Bayo Lawal

Applied Statistical Methods in Agriculture, Health and Life Sciences

EXTRAS ONLINE

Springer

Applied Statistical Methods in Agriculture, Health and Life Sciences

Bayo Lawal

Applied Statistical Methods in Agriculture, Health and Life Sciences

Springer

Bayo Lawal
Department of Statistics and Mathematical Sciences
Kwara State University
Malete
Nigeria

Additional material to this book can be downloaded from http://extra.springer.com

ISBN 978-3-319-05554-1 ISBN 978-3-319-05555-8 (eBook)
DOI 10.1007/978-3-319-05555-8
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014939028

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To
To my wife, Nike, all our children, and Damilola

Preface

This book is aimed at exposing statistical techniques which are very essential to understand research work in the Biological, Agricultural, and Health Sciences. Other disciplines may also find the book useful. This book is born out of my teaching Experimental, field experimentation, and Biostatistics courses at various universities in the United States and Nigeria. The book has also benefited from lecture notes during my graduate program at the University of Reading, Berkshire, UK.

The book covers the basic aspects of statistics, such as data description, probability, sampling distributions, estimation, and hypotheses testing. These topics are covered in Chaps. 1 to 5. Regression analysis and analysis of categorical data are covered in Chaps. 7 and 8 respectively. Chapter 6 covers an introduction to analysis of variance. Here, students are first introduced to treatment comparison methods as well as multiple comparison procedures. This chapter also introduces students to the concepts of contrasts and orthogonality. Chapter 9 introduces students to the principles of experimental design, while Chap. 10 covers the completely randomized design including more coverage on contrast and multiple comparisons as well as the analysis of experiments designed with quantitative levels. Chapter 11 covers the randomized complete block design, including discussion on group balanced block design while Chap. 12 covers Latin square designs as well as cross-over designs. Several examples are introduced in this chapter. This chapter also covers materials relating to multiple Latin squares.

Chapter 13 covers the analysis of covariance in both the completely randomized design (CRD), and the randomized complete block design (RCBD). Chapter 14 introduces students to simple factorial designs in both 2^n and 3^2 designs. The concept of confounding and partial confounding is similarly introduced in this chapter. Resolutions III and IV designs are also introduced in this chapter. The split plot design is introduced in Chap. 15. Also introduced here are the strip-plot and the split-split plot designs with examples. Incomplete block and lattice designs are introduced in Chap. 16. Quantal-bioassay and the logistic regression are introduced in Chap. 17 including the probit

model. The repeated measures design for single and two-factor models is introduced in Chap. 18. Chapter 19 introduces students to survival analysis. Here, the concept of censoring and estimating survival functions is discussed. Hazard and proportional hazard models are similarly discussed. Chapter 20 discusses combined analysis of experiments over time, season, and sites.

Several different examples are presented in the text to illustrate the diversity of the various models. All examples in this text have been analyzed using MINITAB version 16. These examples have therefore been accompanied with their corresponding MINITAB codes embedded in the text. The examples have also been analyzed with R programs, and these are made available at the Springer site which is dedicated to this text. We have presented partial outputs arising from the use of MINITAB 16. To facilitate data entry, many of the data sets for examples and exercises are provided on the book's website (http://extra.springer.com). The example data files are contained in the folder DATAFILES and are presented chapter by chapter. All R program codes for analyzing the examples in the text are contained as ASCII files in RCODES folder. Partial outputs generated from the R programs are contained in the Routput.pdf. This also contains the necessary information on all the examples.

The book is intended for use in undergraduate courses in Agricultural Sciences, Nursing and Health Sciences as well as in Biological Sciences.

Acknowledgments

I gratefully acknowledge the support and guidance of Ms. Hannah Bracken, Springer acquisition editor and sincere thanks to Deepshikha Chauhan, and Neelu Sahu, Production Editor and Manager respectively at Springer. Thank you all.

I have benefited from lecture notes taken while I was a graduate student at the University of Reading, UK, to which I am grateful. I am also grateful to an anonymous scholar from Cornell University, Ithaca, NY, whose lecture notes are copiously used in my introductory write ups on hypothesis testing, principles of experimental design in Chap. 9 and description of data in Chap. 2.

I am grateful to Messers Olorede and Dauda, of the Department of Statistics and Mathematical Sciences at Kwara State University, Malete, Nigeria, for their assistance with the R program codes for some of the examples in this text. I am also indebted to John Wiley and Sons, Brooks/Cole, Oxford University Press, Sage Publisher, Elsevier Limited, McGraw-Hill Companies, Duxbury Press, CRC Press, Cambridge University Press, Dr. M. Friendly, and Dr. S. Long for permission to use copyrighted materials.

I thank the MINITAB Inc., for licensing me the MINITAB software free under their authorship program. This software is used in the analyses of all the data examples in this text, the results of which are presented in the text.

Kwara State University
Malete, Nigeria
January 24, 2014

Bayo Lawal

Contents

List of Figures

List of Tables

About the Author

Bayo Lawal is Professor of Statistics at Kwara State University, Malete, Kwara State Nigeria. He received his B.Sc. (Hons) degree in Mathematics from the Ahmadu Bello University, Nigeria and his Master's degree in Biometry from the University of Reading, UK. His Ph.D. in Statistics is from the University of Essex, UK. Professor Lawal, taught for several years at the University of Ilorin, Nigeria, St Cloud State University, St Cloud, MN, Temple University in Philadelphia, USA. He was a visiting professor in the Department of Biometry and Epidemiology (now Department of Biostatistics) at the Medical University of South Carolina, Charleston, SC between 2009 and 2000. He also served as chair of the Departments of Statistics at St. Cloud State University, MN, USA and at the University of Ilorin in Nigeria. He has also served as Dean of the School of Sciences, Auburn University at Montgomery between 2004 and 2008 as well as the Dean of the School of Arts and Sciences, and Interim Academic Vice-President at the America University of Nigeria, Yola, Nigeria between 2008 and 2011. He is currently serving as head of the Department of Statistics and Mathematical Sciences at Kwara State University, Malete, Nigeria

Chapter 1
Introduction

1.1 Concepts

In a certain university, suppose we were to obtain the height, sex, weight, and age of each student in a large class in Biostatistics, then such a collection of numbers or survey represents characteristics of the group of individuals contained in this class. If the aggregate of individuals in this class represents the only individuals of interest in the survey, they constitute the Universe or population of interest. If, however, the Universe is composed of a wider aggregate or group of individuals, say the full time Undergraduate students at this University, then the members of this class represent only a part or sample of the Universe or population. Complete enumeration or survey of a characteristic in the population is defined as a Census whereas enumeration on only a part of the population is known as a Sample or as a Sample Survey. Population or Universe may consist of characteristics of people, or acres in farms, Sex, etc. etc. Examples are—all progeny of a particular rat, the birthweights of pigs in one litter, all possible values of millet yield per acre in Kano State. Populations are classified and described by numbers. Students in this University for example, are described by their registration or matriculation numbers. It is a fact, however, that the more developed a society is, the more that society will be characterized by numbers. Knowledge of characteristics of a society allows intelligent action to be taken in order to further develop the society. A sample on the other hand is a part of the population (in some cases, a sample may include the whole of the population). Often, we are interested as researchers in the behavior of a variate throughout a population, but observations on every member of the population may be impossible. For instance, we cannot contemplate catching and weighing every fish in a particular study pond or counting the number of every deficient seed in a kilogram bag of seeds. Sometimes too, the restriction is stronger than consideration for economy or speed. The observations or measurements may involve classification of the individual (weight of a rat's heart or amount of a certain trace element in a Tilapia fish) so that full records for a population would prevent any continuing study of that population. Thus, the intention

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_1,

is to use sample information to have an inference about a given population. It is therefore very important to concisely define the population of interest and to obtain a representative sample from such a population so that valid inferences could be made. Numbers used to describe a characteristic of a population and which are derived from all members of the Universe are called Parameters. Parameters represent the facts about the population. Numbers derived from a sample and which may be used to "guesstimate," to estimate, or to approximate the value of the parameters are called Statistics. Thus, a weighted average of a student in any given semester is an estimate of the students grade average for his 4 years at the University, in this example, the one Semester specified weighted average is a statistic, and it estimates the weighted average for a total of eight Semesters. A statistic is subject to variation (i.e. it is a variable). As opposed to the above uses of Statistics as represented by columns of numbers, averages, percentages, ratios, and its like, there is a subject of Statistics which is a field and a science unto itself. Statistics are concerned mainly with the following items:

(i) "to design or to plan experimental investigations (experiments) and Sample Surveys.
(ii) to summarize the numbers collected from experiments and Sample Surveys', and
(iii) to relate or to infer facts about the population utilizing facts from the sample."

Statistics as a science and subject unto itself is a branch of applied mathematics and probability. As such, it is rigorous and well-defined within a framework of definitions and assumptions. Whenever a statistical procedure is applied to a real-life situation, the assumption may or may not be justified. This means then that the application of statistical procedures always involves a degree of subjectiveness. The degree of subjectiveness should be constantly questioned and evaluated in order to make proper use of the statistical procedure under consideration. We may also note here the following:

1. The population is a set of data that characterizes some phenomena.
2. The sample is a set of data selected from a population.
3. A statistical inference is a decision, estimate, prediction, or generalization about the population based on information contained in the sample.

1.1.1 Why Study Statistics?

The growth in data collection associated with scientific phenomena as well as the operations of business and government (quality control, Statistical auditing, forecasting, etc) has been truly outstanding over the past several decades. Published results of political, economic, and social surveys as well as increasing government emphasis on drug and product testing provide vivid evidence

of the need to be able to evaluate data sets intelligently. Consequently, you will want to develop a discerning sense of rational thought that will enable you to evaluate numerical data. You may be called upon to use this ability to make intelligent decisions, inferences, and generalizations. For this reason, the study of statistics is an essential prerequisite for a role in modern Society. Indeed, it is the key Technology. Because the use of Statistics has manifested itself in several aspects of human endeavor (education, research, economic data, etc.), it is necessary therefore that an understanding of the subject of Statistics should form part of our educational training and experience. Such an understanding will acquaint us with the language of the discipline as well as the basic concepts of statistics at least. We will be exposed to the applicable properties of statistical concepts in Biology, Medicine, Agriculture, and social Sciences. However, we must keep in mind that Statistics is intended to be a tool for research.

1.2 Methods of Describing Data

At the beginning of this chapter, we refer to some characteristics of members of this class. Some of this data are measured (e.g., weight) whereas others are classified (e.g., sex which must be male or female).We call these classified records "attributes." Each of the quantities or attributes recorded on each student is called a variate.

Definition A variate is any quantity or attribute whose values varies from one unit of investigation to another.

Definition An observation is the value taken by a variate for a particular unit of investigation. With large data sets, it will be clear for reasons that will be given later that we would need some method for summarizing the information in a data set. Methods for describing data sets are also essential for Statistical inference. Most populations are large data sets. Consequently, if we are going to make descriptive statements (inferences) about a population based on information described in a sample, we will once again need methods for describing a data set.Two methods for describing data are presented in the next chapter—one graphical and the other numerical. As we shall see later, both play an important role in Statistics.

1.2.1 Types of Data

Although the number of phenomena that can be measured is almost limitless, data can generally be classified as one of two types: Quantitative or qualitative.

Definition A quantitative data are observations that are measured on a numerical scale. The most common type of data is quantitative data, since many descriptive variables in nature are measured on numerical scales. Examples of quantitative data are: Number of leaves per plant, yield of cowpea, the heights (or weights) of students in a class, the number of Lecturers in the faculty of Science, University of Ilorin, Nigeria. The measurements in these examples are all numerical. All data that are not quantitative are qualitative. Quantitative variates can also be divided into two types. They may be continuous, if they can take any value we care to specify within some range or discrete if their values change by steps or jumps. Thus the 1000 seed weight of a crop for instance is continuous, because there is no reason why 1000 seeds should not have a weight of 6.94326254 Kg even if no scales could measure it accurately. However, a variate like the number of plants per plot must be whole number $0, 1, 2, \cdots$ going up in steps; decimal values are certainly not allowed here. Heights and weights are obvious examples of continuous variates. On the other hand, discrete observations are integers because they arise from counts.

Definition A qualitative variate or attribute is a variate whose values cannot be put in any numerical order. That is, they are observations that are categorical rather than numerical and are not capable of being measured. Examples of this are "The political affiliations of a group of people." Each person would have one and only one political affiliation. Sex of a person is also another example as it can be either male or female. This type of variate can either be ordinal (if there are intrinsic ordering about its categories, e.g., severity of a disease or a variable with three categories: good, adequate. and poor) or nominal (if its categories are unordered and mutually exclusive). Gender, marital status, flower color are examples of nominal qualitative or categorical variable.

1.3 Measurements

Most biological, agricultural, and medical experiments involve measurements which are numbers that characterize certain variables of a population. It is therefore necessary to have a device for producing meaningful and consistent numbers, or a measuring device. To have repeatable or reproducible measurements or numbers, it is necessary to have a measuring device with a prescribed or measurable margin of error. Note that we do not say that the measuring instrument or device must be error-free, but only that the error of measurement must fall within prescribed limits. Knowing the limits of error of measurement, we are then in a position to determine whether we can or cannot measure a characteristic on the individuals of the sample or of the population with the desired accuracy.

1.3.1 Measuring Devices

The measuring devices utilized by experimenters, lecturers, merchandizers, consumers, etc. are many and diverse. We shall list some of the more common, at least to certain fields, measuring devices with some comments on their use. Perhaps the first measuring device that comes to mind is a ruler or similar device used for linear measurements. This instrument is usually calibrated in feet, inches, 1/6 inches, centimeters, millimeters or other Units of measurements. It is implied from our primary school days that these units are fixed units and never vary. This implication is of course, never made explicit until Secondary school Physics courses are encountered. Even here, they receive only limited attention. Do we ever stop to question how much variation there is between the same calibration marks on the rulers manufactured as brand X? Our experience has told us that rulers commercially available are calibrated closely enough so that we need not worry about errors of calibration in every day life. Unfortunately, this "Safe-feeling" may be carried over into scientific research requiring very precise measurements with sometimes not so happy results. Would any of us recognize the fact that brand Y rulers were only 11.99 inches even though calibrated as a 12 inch ruler? Do any of us know how tall we are to the nearest centimeter when you we arise in the morning? Or how tall we are to the nearest 1/2 inch when we retire at night? Is our height measured with or without shoes and stockings? Such questions lead us to the idea that the height of a person must be defined in precise terms or we shall be unable to determine what is meant by the height of a person except in very general terms. Another measuring device is the scale which is calibrated in pounds and ounces. For scientific investigations, the scale is calibrated in kilograms, grams, centigrams and milligrams. We have spring and balance scales with all degrees of accuracy for both types. Do we ever bother to ascertain the accuracy of the scales used? A few years ago, a research organisation checked the scales used for weighing heavy objects; the scale was found to weigh low for relatively light objects. This would mean that the differences in weight between heavy and medium, heavy and light, and medium and light objects were smaller than they should have been. The error in measurement of weights could have led to erroneous conclusions. A simple check would have revealed this error which had gone undetected for an unknown length of time. If a scale is utilized for precise weights which have important consequences, e.g., in certain research investigations, it should be calibrated against a known standard throughout the total range of weights employed on the scale.Another very common measuring device is the human judge. Humans serve as measuring devices for sports events, beauty contests, taste panels, reading other measuring devices, scoring plant strains for disease, infection, etc. One of the key criteria for a useable judge is the ability to discriminate and to differentiate between levels of the characteristic under consideration. For example, if all "moimo"

within a specified range taste the same to a person, he is useless for discriminating between the small differences that a researcher in home economists is studying. Also, in a beauty contest, if all the girls involved were equally beautiful to a person, this person would be useless as a judge, since he would be unable to pick a winner. If all plant strains appeared to be equally infected with a disease to the judge, when in fact they were not, that person's scores would be useless in differentiating between the strains.The ability to discriminate can be sharpened in many cases with adequate training. However, some individuals may never be able to attain a high level of discrimination with regard to a particular characteristic despite considerable training. One of the key characteristics of outstanding research is their ability to observe and to discriminate among the various types of evidence encountered and then to organize and sort out the pertinent facts. Successful researchers are keen observers. A fourth type of measuring instrument with which we have wide acquaintance is the questionnaire. The questionnaire has many and diverse forms, but they all have one common goal and that is illicit information from or about people and their activities and attitudes. The most widely known form is the ordinary test given in courses. As you all know, there are as many forms of tests as there are lecturers or persons giving the test. There are true–false, multiple-choice, completion, matching, discussion, etc., types of tests and various combinations of these types. Another form of the questionnaire which is associated with surveys and censuses seeks to determine information on such items as type of dwelling, occupancy, and ownership of dwelling, income and expenditure of occupants, attitudes of people toward various items ranging from prejudice to choice of political opponents. These questionnaires are constructed by people who often forget one simple fact and that is—if the person being interviewed does not understand the question and an answer is given, the answer might as well has been generated by a random or chance process. Application forms represent another form which attempts to obtain information about individuals for University admission, job application, etc. Often these forms are very brief, but occasionally, the inventor of forms becomes a little too enthusiastic. The forms of questionnaires are varied and we constantly have to complete one form or another almost daily. Many of you will be involved with developing questionnaires in your life time. Please be precise, exact, and unambiguous. Another type of measuring device is the chemical determination. Large laboratories are constructed for the sole purpose of performing chemical determinations in plants, animals, humans and mineral samples. The results are utilized in several ways. For example, the Drug Department checks on the contents and quality of foods and drugs. Limits of variation in individual items are set and manufacturers must conform to these standards. Other chemical laboratories check soil samples for fertility content, milk samples for butterfat contents, food samples for pesticide residues, concentration and content of drugs, concentration and content of alcoholic beverages, concentration of tars, resins, nicotine etc. in cigars and cigarettes, contents of cosmetics, concentration and identity of weed seeds in

crop or lawn seed, etc. A very large statistical problem in connection with all these items is the design of the sampling procedure and establishment of limits of variation that will be tolerated in the samples. For many items, the statistical standards have not yet been established; there are too many items and too few statisticians. In other cases, standards have been arbitrary, but many of these may be shown to be relatively impossible to attain when studied statistically. For example, in certified seed, the presence of one specified noxious weed makes the entire lot of seeds unsuitable for sale. Now in order to find one noxious in a lot of seeds, it would be necessary to inspect the entire lot seed by seed. This is too expensive and time consuming for commercial seed production, some other means, e.g., field inspection, must be used to eliminate the specified noxious weed seeds from the sample.

1.3.2 Standardization of Measuring Devices

Around 4 m sticks were purchased and arrived in a box which originally contained 12 m sticks. Two of the meter sticks differed from the other by two or more than one millimeter in the calibration marks. The meter sticks carried the same brand name and lot number. This points up the fact that whenever a new measuring device is utilized, it should be checked against a standard—the standard should have known accuracy. A measuring device with unknown accuracy may be useless for the purpose at hand. If we have a meter stick of known accuracy reading the calibration marks, we could check the newly purchased meter sticks against the standard. Scales should be checked for accuracy throughout their usable range of weights prior to using the scale for precise and accurate work. Human judges should be checked for discriminatory power and for level of discrimination. Questionnaires should be pretested prior to use in order to eliminate ambiguities and lack of clarity. Chemical and physical procedures should be checked when first initiated and occasionally thereafter in order to ascertain that the process remains accurate within the prescribed levels. Procedures that are usefully accurate for one type of material may be inaccurate for a second kind of material. The above, as well as all other measuring devices should be calibrated against known calibrated standards; they should be recalibrated at intervals in order to ascertain that the measuring device remains accurate. Duplicate samples and samples of known content are often included along with the unknown samples as a method of checking on the measuring device.

1.3.3 Variability in Measurements

Variability is always present in measurements and it is universal in characteristics of all populations. We live in a variable world. Since, it is universal, we must learn to live with it and to design experimental investigations and

surveys in such a way as to overcome the effects of accomplishing this and they will all be treated under the broad name "Experimental Design Techniques." Variability is very important when biological materials are involved rather than inanimate materials.

1.3.4 Bias in Measurement

An unstated tenet in the collection of numbers utilizing a measuring device is that the plus errors are about equal to the negative errors. Over a large number of trials, one would expect the errors to sum near to zero. Suppose that this is not the case and that the magnitude of inaccuracies in one direction, say, plus, exceeds those in the negative direction. The nature of this type of discrepancy is termed a systematic error, or more commonly, a bias. To illustrate, let us suppose that experimenter A always reads the measuring device one unit higher than does experimenter B. The bias of A compared to B is $+1$, and the bias of B compared to A is -1. Note that we did not state which, if either of the two experimenters, takes correct measurements in the sense that if they measured all individuals in the population, they would obtain the population parameter for the characteristic measured.

1.3.5 Error in Measurement

The causes of variation in measurements are many and varied. These are, as we have pointed out previously, systematic errors and biases, personal errors, mistakes, and errors due to assignable causes. In addition, variation in measurements may be caused by unassignable causes due to the combination of a number of uncontrolled and often unknown variables each with individually small effects. If the magnitude and sequence of these variations are completely unpredictable, i.e., they form a random sequence, we denote them as random variation or random error. The sum of the random errors over all individuals in the population should be zero. The total variation in measurements may be written as:

$$\text{Total variation} = \text{Assignable causes} + \text{bias} + \text{random error}$$

The error of measurement is often defined as

$$\text{Error of measurement} = \text{bias} + \text{random error}.$$

Quite often, the bias is ignored when in fact it may be the larger factor in the error of measurement. If differences between individuals in the population are utilized rather than the individual measurements, the bias term

adds to zero. Regardless of the procedure in order for the measuring device to be useful, some measure of the reliability and accuracy should be known. Thus, the method of utilizing numbers may affect the effect of the bias term; this method must have meaning to the experimenter. We shall discuss in Chaps. 9–11, methods (experimental designs) and randomization techniques that are often employed to minimize random error terms and biases respectively.

1.4 Exercises

1. Classify each of the following into either qualitative (nominal or ordinal) or quantitative (continuous or discrete)
   - Birth weight, date of birth and father's race of a new baby
   - Level of cholesterol in a cubic milliliter of blood
   - smoking status (never, former, or current)
   - injury (severe, moderate, mild, none)
   - sex of a new born
   - Species of a tree (redwood, cedar, pine, oak)
   - Blood group type (A, B, AB).
2. What is meant by descriptive statistics?
3. Define the following: Quantitative variable; Discrete variable.

Chapter 2
Frequency Distributions

2.1 Introduction

One principal aim of any statistical enquiry is to be able to understand and describe the population of interest. For example, a farm survey is aimed at estimating current crop output and evaluating the impact of various government policies; a consumer survey will be interested in assessing how much of its product is being consumed and what is the chance of increasing production if some action is taken. Thus, the first task of a statistical staff is that of organizing the data in the form that salient characteristics can be easily seen.

Suppose in your enumeration area, 35 farming households were sampled, and the weights of heads of households in kilograms (to nearest whole number) as obtained from the field are shown in Table 2.1:

Table 2.1 Weights of heads of households in kilograms

70	66	60	55	61	63	72
68	60	60	63	60	75	68
59	71	53	76	64	64	52
64	64	68	64	66	67	63
64	70	69	68	63	59	57

These data are what we call *raw data*, that is, data as obtained from the field. With the data in this form, very little information can be obtained about the population. The first possible thing that we can do is to put the data in what we call an *array*. An array is the arrangement of the values in ascending or descending order of magnitude. For example, if we put the data in an ascending array we have the following results:

52	53	55	57	59	59	60
60	60	60	61	63	63	63
63	64	64	64	64	64	64
66	66	67	68	68	68	68
69	70	70	71	72	75	76

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_2,

```
NAME C1 'WEIGHTS'
SET C1
DATA>70 66 60 55 61.........57
DATA>END
SORT C1 C2
PRINT C1 C2
```

2.2 Frequency Distributions

2.2.1 Ungrouped Distribution

The above initial analysis can be improved by finding out how many farmers have specific weights.

Sample No.	Weights	No. of farmers having such weights
1	52	1
2	53	1
3	55	1
4	57	1
5	59	2
6	60	4
7	61	1
8	63	4
9	64	6
10	66	2
11	67	1
12	68	4
13	69	1
14	70	2
15	71	1
16	72	1
17	75	1
18	76	1

Note that the total should be equal to the number of households. This classification tells us more about the sample; for example, we could see that:

(i) most farmers have different weights
(ii) the most popular (or common) weight of household head is 64 kg.

This is an example of ungrouped frequency distribution. The display is called a *frequency table*.

Definition

The number of farmers having a certain weight is called its *frequency*. In general, the number of times a particular variable/individual occurs is called its frequency. This is represented by "f." For example, the frequency of 67 is 1, that of 68 is 4, etc.

2.2.2 Grouped Distribution

One serious disadvantage of the classification above is that the table may be too long. Take an example when we consider the weights of a sample of 200 households. The analysis in the form of the preceding section becomes too cumbersome and uninformative.

A more convenient way of summarizing a large mass of raw data is to group the observations/variables (in this case) weights into categories and find out how many household heads belong to each category, for example, how many household heads have weights?

- 52 kg to a weight less than 54 kg
- 54 kg to a weight less than 56 kg
- 56 kg to a weight less than 58 kg, etc.

We write the above in a more shortened form:

- 53 kg - under 54 kg
- 54 kg - under 56 kg
- 56 kg - under 58 kg

Each of these categories is called a class interval. A simple procedure we use is what we call *Tally Score Method*. This method consists of making a stroke in the proper class for each observation and summing these for each class to obtain the frequency. It is customary for convenience in counting to place each fifth stroke through the preceding four as shown below.

Weights in kg	Tally	No. of farmers (f)
52 - under 56	111	3
56 - under 60	111	3
60 - under 64	11111 1111	9
64 - under 68	11111 1111	9
68 - under 72	11111 111	8
72 - under 76	11	2
76 - under 80	1	1
	Total	35

Descriptive Analysis

(i) No household head has weight that is less than 52 kg and more than 80 kg.
(ii) The most common weight is somewhere between 60 and 68 kg.
(iii) Most of the farmers have weights from 56 to 72 kg, that is, $3+9+9+8=29$ or 83 % of the farmers.

This is an example of a grouped frequency distribution.

Definitions

- Class Interval: Each category is called a class interval or simply a class.
- Class Limits: These are the end numbers of each class, e.g., 52, 56, 58, etc.
- Upper Class Limit: This is the larger number of the class intervals, e.g., 56.
- Lower Class Limit: This is the smaller number of the class intervals, e.g., 52.
- Size or Width of a class interval: This is the difference between the upper and lower class limits, e.g., $56 - 52 = 4$.
- Class Mark: This is the midpoint of the class interval and is defined as

$$\frac{\text{Upper Class Limit} + \text{Lower Class Limit}}{2}, \quad \text{e.g.,} \quad \frac{56 + 52}{2} = 54, \quad \text{etc.}$$

- Class Boundary: When the upper limit of each class is the same as the lower limit of the next class, the class limits are called class boundaries (above example).

2.2.3 Constructing a Frequency Distribution

There is no hard and fast rule for the construction of frequency distribution, but the following procedures may be followed:

(i) Try to use equal class interval width. This is useful for comparative purposes and for easier calculations.
(ii) The number of classes should not be too many or too few. A rough guideline for constructing k classes for a sample data is the smallest integer value of k such that $2^k \geq n$, where n is the sample size. In the example above, the sample size is 35 and since $2^5 \leq 35 \leq 2^6$, we would employ $k = 6$ classes. Note that in our example above, we have used seven classes.
(iii) It is advisable to use class interval width of multiples of 2, 5, or 10.

In our example above, if we choose $k = 6$ classes, then, the class width is computed as

$$\frac{\text{Largest value} - \text{Smallest value}}{\text{class size}} = \frac{\text{Range}}{\text{Class size}} = \frac{\text{Range}}{k} = \frac{76 - 52}{6} = 4$$

We would usually increase this class width by a little notch say, to 4.2 or 4.5. Suppose we choose 4.5. We can now start the construction of our classes by starting from a value that is slightly less than the minimum. Our minimum in this case is 52. Suppose we start with 51.5. We then give below the construction of the six classes with a class width of 4.5.

Weights	Midpoints	Tally	Frequency (f)
51.5 - <56.0	53.75	111	3
56.0 - <60.5	58.25	11111 11	7
60.5 - <65.0	62.75	11111 11111 1	11
65.0 - <69.5	67.25	11111 111	8
69.5 - <74.0	71.75	1111	4
74.0 - <78.5	76.25	11	2
Total			35

Note The idea of having equal class interval may be waived in a lot of cases. For example, when we have a lot of classes with very few values, it might be advisable to lump them together.

Another example is the case when some classes are unbounded, that is, when we have the case of *open class intervals*. The table below gives the ages of pupils in a primary school in years.

Age (years)	Frequency (f)
Under 6	3
6 - 7	39
8 - 9	42
10 - 11	40
12 - 13	36
Above 13	7

Note that the classes Under 6 and Above 13 have no lower limit and upper limit, respectively.

2.2.4 Other Forms of Frequency Distribution

Relative Frequency

We may be interested in the proportion of our sample or population that falls in a certain class. In this case, we make use of relative frequency. The result of dividing each class frequency by the total frequency of all classes and multiplying the result by 100 is the relative frequency.

Weights	Frequency	Relative frequency (%)
52 - under 56	3	$\frac{3}{35} \times 100 = 8.6$
56 - under 60	3	$\frac{3}{35} \times 100 = 8.6$
60 - under 64	9	$\frac{9}{35} \times 100 = 25.7$
64 - under 68	9	$\frac{9}{35} \times 100 = 25.7$
68 - under 72	8	$\frac{8}{35} \times 100 = 22.9$
72 - under 76	2	$\frac{2}{35} \times 100 = 5.7$
76 - under 80	1	$\frac{1}{35} \times 100 = 2.9$
Total	35	100.10

The relative frequency is mostly useful for easy comparison of two or more frequency distributions. A biological example for instance is the situation where we wish to compare the number of seeds germinating in two varieties of a plant.

The following data in Table 2.2 are used to illustrate the comparative use of the relative frequency approach.

Table 2.2 Age distribution of grade and pupils in Gabon, 1962.

	Frequency (f)		
Age (years)	Boys	Girls	Total
10 - 11	6	5	11
12 - 13	119	49	168
14 - 15	210	102	312
16 - 17	169	75	244
18 - 19	34	4	38
20 - 21	12	-	12
22 - 23	2	-	2
Total	552	235	787

Source: Fundamentals in Educational Planning, (UNESCO)

One cannot compare these values straightaway because the population of the boys in the school is greater than that those of girls, so expectedly, the figures for boys will be greater than those for girls. However, to compare both results, we would need to convert both frequencies into relative frequencies. The relative frequency is very useful for an easy comparison of two or more frequency distributions. We give an example of such a use with the data below which relate to the age distribution of pupils in Gabon in 1962.

	Relative frequencies		Total relative
Age (years)	Boys	Girls	frequency
10 - 11	1.1	2.1	1.4
12 - 13	21.5	20.9	21.3
14 - 15	38.0	43.4	39.6
16 - 17	30.6	31.9	31.0
18 - 19	6.2	1.7	4.8
20 - 21	2.2	0	1.5
22 - 23	0.4	0	0.3
Total	100	100	100

The results from the above analysis suggest the following:

(i) Gabonese government should encourage more girls to school.
(ii) The proportional distribution of ages by sex is close enough except for age group 14 - 15 (difference = 5.4 %) and 18 - 19 (difference = 4.5 %).
(iii) More boys of older age stay at school.

2.2.5 Cumulative Frequency Distributions

Suppose for the data in Table 2.1, we are interested in answering questions such as:

- How many household heads weigh less than 53 kg?
- How many household heads weigh more than 52 kg?

The answers to these and other similar questions are best answered through *cumulative frequency* distributions.

Weights in kg	Frequency	Cumulative frequency from below	Cumulative frequency from above
52 - under 56	3	3	35
56 - under 60	3	6	32
60 - under 64	9	15	29
64 - under 68	9	24	20
68 - under 72	8	32	11
72 - under 76	2	34	3
76 - under 80	1	35	1
Total	35		

No. of farmers whose weights are less than 52 kg = 0
No. of farmers whose weights are less than 56 kg = 3
No. of farmers whose weights are less than 60 kg = 6
No. of farmers whose weights are less than 64 kg = 15
No. of farmers whose weights are less than 68 kg = 24
No. of farmers whose weights are less than 72 kg = 32
No. of farmers whose weights are less than 76 kg = 4
No. of farmers whose weights are less than 80 kg = 35

The above are obtained from the cumulative frequency distribution from below. Similarly, we have,

No. of farmers whose weights are greater than 52 kg = 35
No. of farmers whose weights are greater than 56 kg = 32
No. of farmers whose weights are greater than 60 kg = 29
No. of farmers whose weights are greater than 64 kg = 20
No. of farmers whose weights are greater than 68 kg = 11
No. of farmers whose weights are greater than 72 kg = 3
No. of farmers whose weights are greater than 76 kg = 1
No. of farmers whose weights are greater than 80 kg = 0

The above are similarly obtained from the cumulative frequency distribution from above.

2.3 Graphical Representation of Data

Many people have a strong aversion for anything having numbers and tables, so it might be useful to represent frequency distribution in a more appealing form. One method is to represent the frequency distribution in graphic form which is more informative to the layman. We consider some cases:

2.3.1 The Dotplot

One very useful and simple graphical way to display data is by the use of the graphical method called the *dotplot*. The plot employs a horizontal line with the appropriate axis mark to reflect the range of the data. Each sample observation is then represented in the graph by a single dot above the horizontal line at the specified value. For instance, a data value of 55 is represented by a single dot in the figure below, while a value of 64 is represented by six dots that are stacked above one another. The figure below is a MINITAB output of the dotplot for the data in Table 2.1. We could see that the interval 59–70 contains most of our data values. Further, the plot provides visual information which otherwise could not be discerned from mere looking at the original data in Table 2.1.

```
MTB > DotPlot 'WGT'.

Dotplot: WGT
                                        :
                                  :     : :       :
                 . .   .   .    : : .   : :    : . : . : . .      . .
            +---------+---------+---------+---------+---------+-------WGT
          50.0      55.0      60.0      65.0      70.0      75.0
```

A very useful advantage of the dotplot is in comparative analysis of two distributions.

2.3.2 The Stem and Leaf Display

The stem and leaf plot offers a quick way to graphically display the shape of continuous type data while including the actual numerical values in the graph. That is, the plot retains the original values of the data. The stem and leaf works best for small numbers of observations as each item of data must be listed. Below is a MINITAB command to construct a stem and leaf display of the data in Table 2.1 which was stored in column 1, C1.

```
MTB > STEM AND LEAF C1;
SUBC> INCREMENT=5.

Stem-and-leaf of weights   N  = 35
Leaf Unit = 1.0

    2    5 23
    6    5 5799
  (15)   6 000013333444444
   14    6 66788889
    6    7 0012
    2    7 56
```

The first column from the MINITAB output for stem and leaf display gives the cumulative frequencies, both from above and below to the interval in which the *median* is located. Thus the parentheses around 15 indicate that the median is in that class interval. The column also tells us that 6 household heads have weights below 60 and 14 who have weights of at least 70.

To construct the stem and leaf display, we note that the minimum datum here is 52 and the maximum is 76. Thus, we could make this a *one-stemmer* by having as stems the tens digits 5, 6, and 7, while the ones digit would then constitute the leaves. This would only result in only three classes, which would not give a fair pictorial representation of the data. This approach is displayed in the following:

```
MTB > STEM AND LEAF C1;
SUBC> INCREMENT=10.

Stem-and-leaf of weights   N  = 35
Leaf Unit = 1.0

    6    5 235799
  (23)   6 00001333344444466788889
    6    7 001256
```

The stem and leaf display we have in the figure above is an example of a *two-stemmer* display. Here the 5's for instance are broken into two groups; 50 - 54 and 55 - 59. That is, the leaves in both groups are respectively the digits {1, 2, 3, 4} and {5, 6, 7, 8, 9}. The two stemmers can be invoked in MINITAB by using the subcommand *increment = 5* while the one-stemmer can similarly be invoked by using the subcommand *increment = 10*. Other forms of the stem and display are the *five-stemmer* and the *ten-stemmer*. For a five stemmer, we would have for the 5's the following stems.

Stems	Leaves
2*	With unit digits 0 or 1
2t	With unit digits 2 or 3
2f	With unit digits 4 or 5
2s	With unit digits 6 or 7
5•	With unit digits 8 or 9

In this splitting, the symbol t is used for the digits 2 and 3; f for four and five; and s for six and seven. We again give this display for our data in Table 2.1. The display is generated by the MINITAB subcommand *increment = 2*.

```
MTB > stem and leaf c1;
SUBC> increment=2.

Stem-and-leaf of weights   N  = 35
Leaf Unit = 1.0

    2    5 23
    3    5 5
    4    5 7
    6    5 99
   11    6 00001
   15    6 3333
   (6)   6 444444
   14    6 667
   11    6 88889
    6    7 001
    3    7 2
    2    7 5
    1    7 6
```

We note that in all the above MINITAB displays of the stem-and-leaf plots, the MINITAB orders the leaf units. However, one needs to very careful with stem-and-leaf displays because the display itself does not tell you the actual value of the data. The actual value is provided by the leaf *unit* = statement which is given just above the display. For example, if the leaf unit = 1.0 had been leaf unit = 10, then the smallest data element would have been 520. Similarly, if the leaf unit had been leaf unit = 0.001 instead of 1.0, then the smallest data element would have been 0.052. We give an example below where the data in Table 2.1 were multiplied each by 10, and the resulting stem and leaf display below (a five-stemmer) gives the leaf unit = 10, indicating that the minimum data element is 520 and the maximum being 760. Notice that this display is very similar in every respect to the five-stemmer display above, except for the leaf unit value.

```
MTB > LET C3 = C1*10
MTB > STEM AND LEAF C3

Stem-and-leaf of C3        N  = 35
Leaf Unit = 10

    2    5 23
    3    5 5
    4    5 7
    6    5 99
   11    6 00001
   15    6 3333
   (6)   6 444444
   14    6 667
   11    6 88889
    6    7 001
    3    7 2
    2    7 5
    1    7 6
```

2.3.3 Histograms

A histogram consists of a set of rectangles whose:

(i) bases are on the horizontal axis (X-axis) with lengths equal to the size of the class intervals
(ii) areas are proportional to the class frequencies.

Remark

We will consider the case when the size of all class intervals are equal. The frequencies in this case represented on the vertical Y-axis are taken numerically to be equal to the height of the rectangle. As an example, consider the data in Table 2.1 with the corresponding frequency distribution displayed earlier using the tally method. There we have six classes with the class width of 4.5. We can implement this in MINITAB by doing the following.

```
MTB > SET C1
DATA> 70 66 60 55 61 63 72 68 60 60 63 60 75 68
DATA> 59 71 53 76 64 64 52 64 64 68 64 66 67 63
DATA> 64 70 69 68 63 59 57
DATA> END
MTB > GStd.
* NOTE  * Character graphs are obsolete.

MTB > Histogram 'WGT';
SUBC>   Start 53.75 76.25;
SUBC>   Increment 4.5.

Histogram of WGT   N = 35

Midpoint       Count
   53.75           3  ***
   58.25           7  *******
   62.75          11  ***********
   67.25           8  ********
   71.75           4  ****
   76.25           2  **
```

A graphical version of the histogram can be accomplished with the following commands with the resulting histogram (Fig. 2.1).

```
MTB > GPro.
MTB > Histogram 'WGT';
SUBC>   MidPoint 53.75:76.25/ 4.5;
SUBC>   Bar;
SUBC>   ScFrame;
SUBC>   ScAnnotation.
```

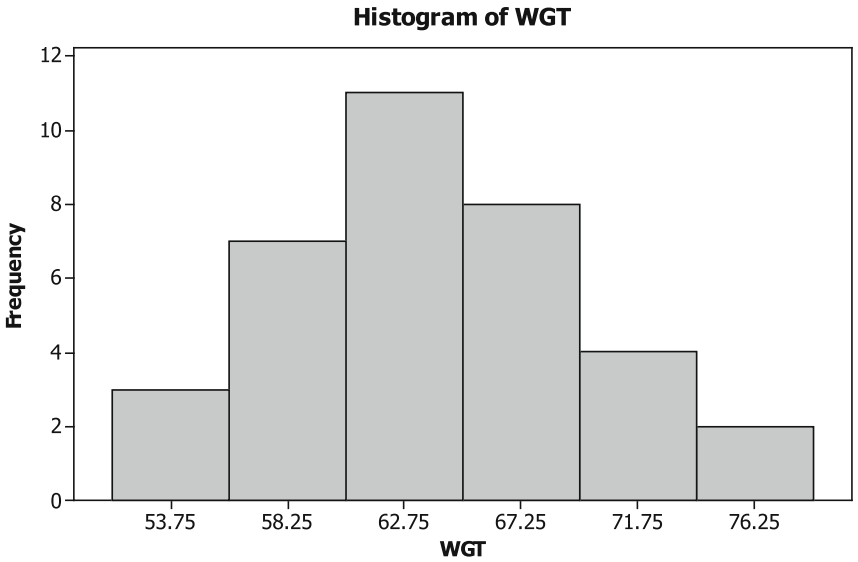


Fig. 2.1 Histogram plot for the data

Example 2.3.1

We shall again illustrate the construction of an histogram with the following example:
In a plot of 130 tillers, the following information was obtained:

No. of plants	0	1	2	3	4	5	6	Total
No. of tillers	25	19	22	15	33	13	7	130

Which of them is the variable?
Which of them is the frequency?

The procedure is as follows:

1. The variable, number of plants will be on the X-axis.
2. The frequency, number of tillers will be on the Y-axis.
3. Scale your Y-axis from zero and in such a way as to take the highest frequency (33). An example is to scale from 0 to 35.
4. On the X axis, mark 0, 1, 2, 3, 4, 5, 6 spacing with equal intervals.
5. Draw each rectangle in such a way that the variable values are at the center and the heights equal the number of families.

The resulting histogram with midpoints corresponding to $0, 1, \cdots 6$ is displayed in Fig. 2.2.

Fig. 2.2 Histogram for this example

2.3.4 Polygons

Frequency Polygon

This assumes that observations in a class interval are clustered around the central value, that is, the class mark.

The frequency polygon is a line graph constructed by plotting the class frequencies of the various classes at their respective class marks and connecting these points by means of straight lines. Again, we can use the data in Table 2.1 as an example.

Remark

It is customary to complete the picture by adding one class at each end of the distribution with zero frequencies.

That is, the class 48 - 52 with class mark of 50.0 and $f = 0$.
Also, the class 80 - 84 with class mark $= 82.0$ and $f = 0$. We then join the points.

2.3.5 Cumulative Frequency Polygon - The Ogive

This is a graph of cumulative frequencies plotted against the class boundaries. Consider the previous examples.

Less than 52 kg = 0; Less than 56 kg = 3

Less than 60 kg = 6; Less than 64 kg = 15

Less than 68 kg = 24; Less than 72 kg = 32

Less than 76 kg = 34; Less than 80 kg = 35

We will concern ourselves with less than ogive.

```
MTB > NOTE OGIVE FOR HOUSEHOLD HEAD WEIGHT DATA
MTB > SORT C1 C2
MTB > SET C3
DATA> 1:35
DATA> END
MTB > NAME C3 'CUMFREQ'
MTB > Plot 'CUMFREQ'*'WGTS';
SUBC>   Connect;
SUBC>   ScFrame;
SUBC>   ScAnnotation.
```

All the above graphs can be drawn using relative frequencies. The advantages of using graphical representations are:

(i) The pattern of a distribution is easily seen from a graph.
(ii) It is more informative to the layman.

The ogive for the data in Table 2.1 is constructed in MINITAB with the following statements and the graph is presented in Fig. 2.3.

```
MTB > ECDF 'WGT';
SUBC>   Connect;
SUBC>   Distribution.
 Empirical CDF of WGT
```

We may note here that a normal distribution with computed mean and standard deviation is superimposed on the the ogive in Fig. 2.3. If we do not want this overlay, we can simply remove the Distribution statement in the MINITAB statement and simply put a period after the "connect" statement.

2.4 Presentation of Data: Charts and Diagrams

So far we have considered diagrams used to illustrate variables. Attributes (that is, qualitative variables) can also be illustrated pictorially. We consider some cases below.

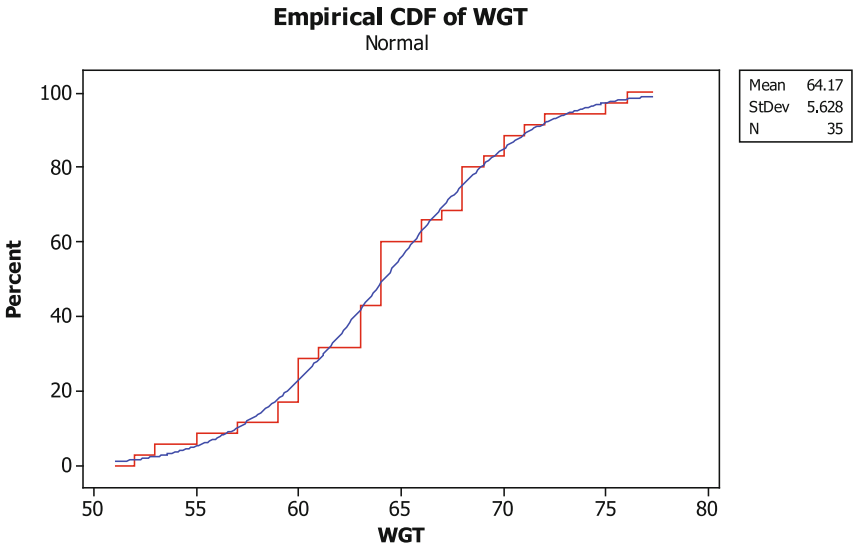


Fig. 2.3 Plot of the ogive for the data in Table 2.1

2.4.1 The Bar Chart

The Simple Bar Chart

We shall use a simple *Bar Chart* to illustrate the volume of cocoa exported from Nigeria between 1960 and 1965 (Fig. 2.4). The table below gives the volume of cocoa in metric tons (thousands) exported by Nigeria between 1960 and 1965.

Year	Metric tons
1960	73.6
1961	67.4
1962	66.8
1963	64.8
1964	80.2
1965	85.4

The procedure for drawing a bar graph is the following:

(a) Each value is represented with a bar (rectangle) and its height to its value.
(b) The width of all rectangles is the same (that is, equal).
(c) The bars are separated by intervals (or gaps) of equal size.

```
Data Display

 Row    YEAR   Tonnage

   1    1960   73.6
   2    1961   67.4
   3    1962   66.8
   4    1963   64.8
   5    1964   80.2
   6    1965   85.4

MTB > Chart Mean( 'Tonnage' ) * 'Year';
SUBC>   Bar.
```

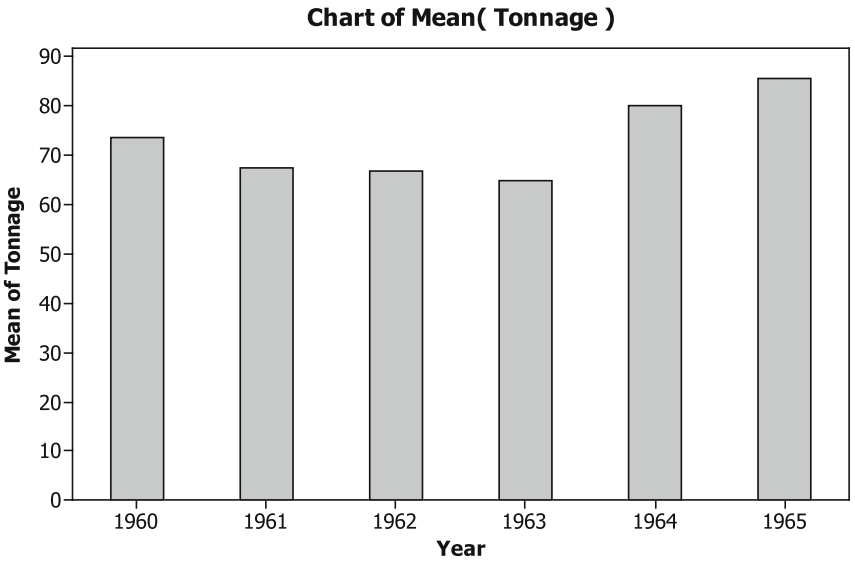


Fig. 2.4 Simple bar chart

2.4.2 Multiple Bar Chart

We can also construct *Multiple Bar Chart* which is mostly used for comparative purposes. We shall use this technique to compare the purchase of palm kernels in Kwara State from Okene/Okehi and Oyun local government areas between 1971/1972 and 1973/1974.

	Palm kernels (in tons)		
	1971/1972	1972/1973	1973/1974
Okene/Okehi	33	19	6
Oyun	84	26	44
Total	117	45	50

Source: Kwara State Statistical Year Book 1977/1978.

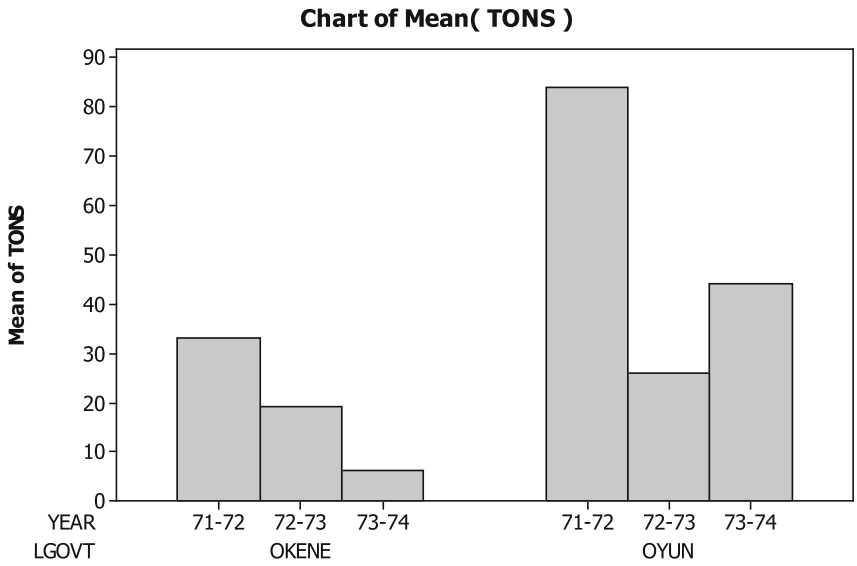


Fig. 2.5 Multiple bar chart

We see from the multiple bar graphs in Fig. 2.5 that Oyun local government of Kwara State produced far more tons of palm kernel than Okene local government in the years 1971 and 1973/1974. The exception perhaps being 1972/1973. This graph in 2.5 can easily be implemented in MINITAB with the following statements.

```
MTB > print c1-c3

Data Display

Row  LGOVT  TONS  YEAR
  1  OKENE    33  71-72
  2  OKENE    19  72-73
  3  OKENE     6  73-74
  4  OYUN     84  71-72
  5  OYUN     26  72-73
  6  OYUN     44  73-74
MTB > Chart Mean( 'TONS' ) * 'LGOVT';
SUBC>   Group 'YEAR';
SUBC>   Overlay;
SUBC>   Bar.

Chart of Mean( TONS )
```

2.4.3 Component Bar Chart

This is a simple bar chart divided into sections such that each division (height) corresponds in magnitude to the value it represents. For example, a component bar chart for the data employed for the multiple bar chart in the previous section can be constructed as follows.

(a) Draw simple bars of the totals.
(b) Divide each simple bar into components by just marking off respective values.

We give in Fig. 2.6 an implementation of the component bar graph for the last example.

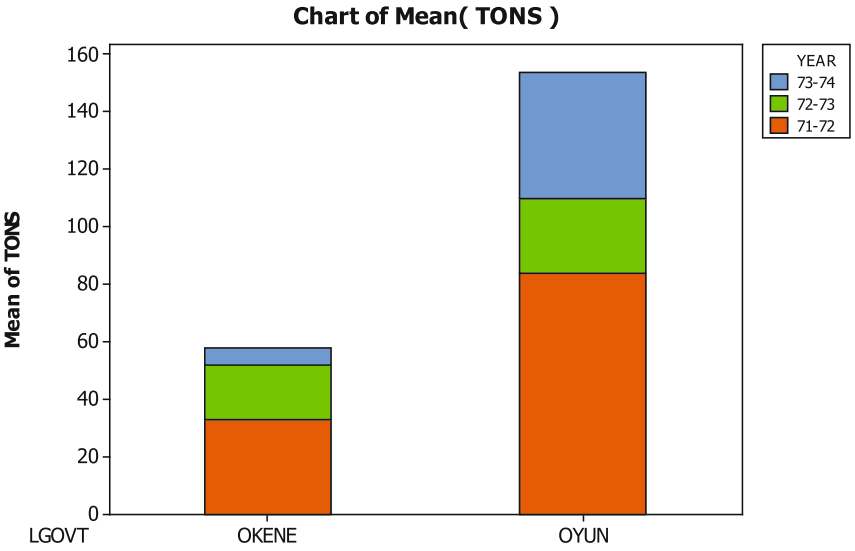


Fig. 2.6 Component bar chart

2.4.4 Pie Charts

The pie chart is mostly suitable for categorical variables and represents our variables or attributes in the form of circles. As an example, we will construct a pie chart from the table below which gives the amount of money realized from the export of the principal crops of Nigeria in 1965 in millions of Naira. Here, we consider crop as a categorical nominal variable with three categories.

Crop	Value (in million N)	Relative frequency
Cocoa	85.4	0.3142
Palm produce	80.2	0.2951
Groundnut	106.2	0.3907
Total	271.8	1.0000

To draw a pie chart, we first note that since a circle spans 360°, the circle can thus be divided into sections such that the size of each section is obtained as:

N 271.8 million is represented by 360° (whole circle). Therefore, N 1 million will be represented by $\frac{360}{271.8} = 1.325°$. We therefore have the distribution for each of the produce as follows (that is, the slices of the pie corresponding to each category):

$$\text{Cocoa} = \frac{360}{271.8} \times 85.4 = 113.1°$$

$$\text{Palm} = \frac{360}{271.8} \times 80.2 = 106.2°$$

$$\text{Groundnut} = \frac{360}{271.8} \times 106.2 = 140.7°$$

The above slices in degrees can also be obtained by multiplying the relative frequencies with 360°. Using a protractor and a compass, we can easily draw a pie chart. Fig. 2.7 gives the pie chart for the principal export crops in Nigeria in 1965.

```
Data Display

 Row   CROP   VALUE

   1  COCOA    85.4
   2  PALM     80.2
   3  GNUT    106.2

MTB > PieChart ( 'VALUE' ) * 'CROP';
SUBC>   Combine 0.02;
SUBC>   Panel.

Pie Chart of CROP
```

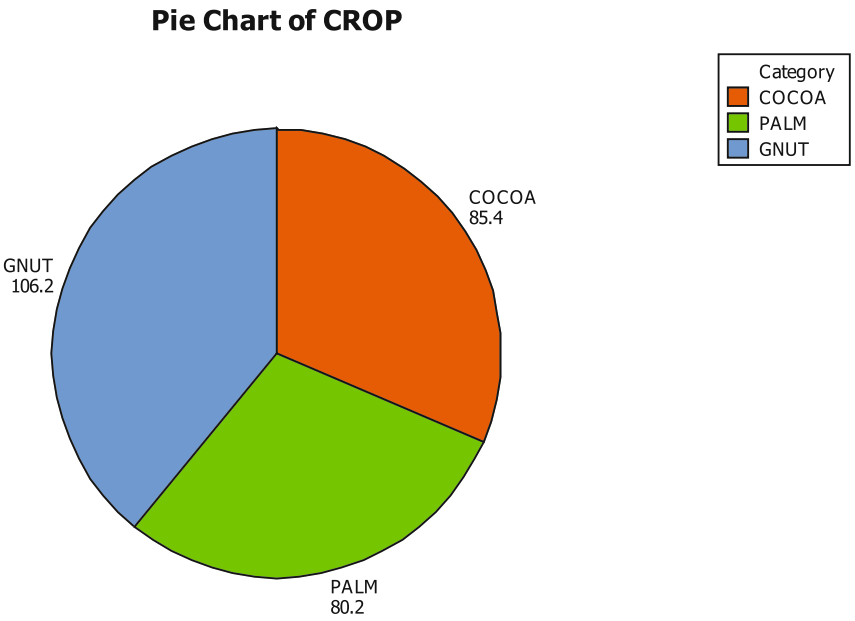


Fig. 2.7 A pie chart example

2.5 Exercises

1. The data below give the weight in kilograms of 100 college students taken at random in fall 1996.

Weight (kg)	Frequency (f)
60 - 62	5
63 - 65	18
66 - 68	42
69 - 71	27
72 - 74	8

   Find the mean and median of the grouped data. Also, calculate Shannon's index of diversity and interpret your result.
2. The data below relate to ozone levels measured as high as 220 parts per billion (ppb) in a forested area of Edo State. Concentrations this high can cause eyes to burn and are a hazard to both plants and animal life.

```
160   176   160   180   167
164   165   163   162   168
173   179   170   196   185
163   162   163   172   162
167   161   169   178   161
```

Construct a two-stemmer stem-and-leaf display for the data. What can you say about the shape of the data?

3. The angle between two adjacent toes was measured from radiographs of the affected feet of 50 young adults undergoing treatment for a foot abnormality:

ANGLE BETWEEN TOES (DEGREE)

42	32	33	33	29	31	33	29	40	31
27	30	29	43	34	29	34	29	28	30
36	46	30	41	45	31	30	33	29	29
33	35	37	27	29	43	32	27	32	32
39	41	44	32	35	29	31	28	28	29

Choose a suitable class interval, arrange the results in a frequency table. Construct a histogram of the data.

4. The cholesterol levels for a sample of 100 subjects are classified as follows:

Cholesterol level	Number of students
Recommended	25
Borderline	10
Moderate risk	50
High risk	15

a. Construct a bar chart to display the distribution.
b. Use a pie chart to present the distribution.

Chapter 3
Numerical Description of Data

3.1 Introduction

The graphic procedures described in the last chapter help us to visualize the pattern of a data set. To obtain a more objective summary description and a comparison of data sets, we must go one step further and formulate quantitative measures for important aspects such as, the location of center of the data and the amount of variability present in the data. To effectively present the ideas and associated formulas, it is convenient to represent a data set by symbols to prevent the discussion from becoming anchored to a specific set of numbers. A data set consists of a number of measurements symbolically represented by $x_1, x_2, \cdots, x_n$. The last subscript n denotes the number of measurements in the data and $x_1, x_2, \cdots,$ represents the first observation, the second observation and so on.

The notation $\sum_{i=1}^{n} x_i$ represents the sum of n numbers $x_1, x_2, \cdots, x_n$ and is read as the sum of all x_i, with i ranging from 1 to n or

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

Examples

1.

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4$$

2. If $x_1 = 3$, $x_2 = 5$, $x_3 = 4$ and $x_4 = 3$, then,
   (i)

$$\sum_{i=1}^{4}(x_i - 2) = (x_1 - 2) + (x_2 - 2) + (x_3 - 2) + (x_4 - 2)$$

$$= \sum x_i - 4(2)$$

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_3,

$$= 15 - 8$$
$$= 7$$

(ii)

$$\sum_{i=1}^{4} 3x_i = 3x_1 + 3x_2 + 3x_3 + 3x_4 = 3\left(\sum_{i=1}^{4} x_i\right) = 3 \times 15 = 45$$

(iii)

$$\sum_{i=1}^{4} x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 = 3^2 + 5^2 + 4^2 + 3^2 = 59$$

(iv)

$$\sum_{i=1}^{4}(x_i - 2)^2 = (x_1 - 2)^2 + (x_2 - 2)^2 + (x_3 - 2)^2 + (x_4 - 2)^2$$
$$= (3 - 2)^2 + (5 - 2)^2 + (4 - 2)^2 + (3 - 2)^2$$
$$= 1 + 9 + 4 + 1$$
$$= 15$$

3.1.1 Properties OF $\sum$

If a and b are constants, then,

$$(i)\ \sum_{i=1}^{n} bx_i = b\sum x_i$$

$$(ii)\ \sum_{i=1}^{n}(bx_i + a) = b\sum x_i + na$$

$$(iii)\ \sum_{i=1}^{n}(x_i - a)^2 = \sum x_i^2 - 2a\sum x_i + na^2$$

$$(iv)\ \sum_{i=1}^{n} a = na, \quad \text{and hence,} \quad \sum_{i=1}^{n} 1 = n$$

3.2 Measures of Center or Central Tendency

Perhaps the most important aspect of studying the distribution of a sample of measurements is the position of a central value, i.e., a representative value about which the measurements are distributed. Any numerical measure intended to represent the center of a data set is called a measure of location

or central tendency. The two most commonly used measures of center are the *mean* and the *median*.

3.2.1 The Mean, $\bar{x}$

The sample mean or average of a set of n measurements $x_1, x_2, \cdots, x_n$ is the sum of these measurements divided by n. The mean is denoted by $\bar{x}$ and is expressed as:

$$\bar{x} = \frac{\sum x_i}{n}$$

Examples

(1) Given the heights in inches of five men as 66, 73, 68, 69, and 74. Then the mean equals

$$\bar{x} = \frac{\sum x_i}{n} = \frac{66 + 73 + 68 + 69 + 74}{5} = \frac{350}{5} = 70$$

i.e., $\bar{x} = 70$ inches.

(2) The birth weights in pounds of five new born babies at a hospital on a certain day are 9.2, 6.4, 10.5, 8.7, and 7.8. Hence, the mean birth weight for this data is

$$\bar{x} = \frac{9.2 + 6.4 + 10.5 + 8.1 + 7.8}{5} = \frac{42.0}{5} = 8.4 \text{ lbs}$$

The pattern in the data can be seen more easily if the readings are arranged in order of magnitude as shown below:

$$6.4, 7.8 \vee 8.7, 9.2, 10.5$$

$$\text{mean } 8.4$$

We can see that the mean is a good summary figure. About half of the readings are smaller than the mean and half larger. Even when the mean is not a good summary figure, it generally provides a useful mental or visual focus when looking at the data. So, it is a good idea to calculate the mean of a data set in the early stages of analysis. At the very least, this can make it easier to see whether the data are symmetrical or skewed.

Many times, our data appear in frequency tables where we no longer know the actual values of the observations, but only to which class interval they belong. In these instances, the best we can do is to approximate the sample mean. The mean is given by the expression:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{n}. \tag{3.1}$$

3.3 Weighted Means

If for instance, we are interested in finding the mean of several means which are themselves obtained on different numbers of observations, then it is appropriate to weight the means or observations by using weights to depend on the number of observations in each mean. A *weighted mean* is therefore defined by,

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

where, w_i are the weights. Consider for example the data in Table 2.1. The means of the first 14 observations for this data is 64.36, while the mean of the remaining 21 observations is 64.06 respectively. If we think, therefore that the mean of the entire 35 observations would be $\frac{64.36 + 64.06}{2} = 64.21$, which clearly does not agree with the actual mean of 64.18. The reason here is that the means are not weighted. The true mean will be computed as:

$$\bar{x}_w = \frac{14(64.36) + 21(64.06)}{14 + 21} = \frac{2246.3}{35} = 64.18$$

Here, the respective weights are $w_1 = \frac{14}{35} = 0.4$ and $w_2 = \frac{21}{35} = 0.6$ respectively.

3.3.1 Geometric Mean

If $x_i, x_2, \cdots, x_n$ are all positive numbers, then the *geometric* or *harmonic* mean is given by

$$\begin{aligned} G &= \sqrt[n]{(x_1 x_2 \cdots x_n)} \\ &= (x_1 x_2 x_3 \cdots x_n)^{1/n} \end{aligned}$$

$$\text{and,} \quad \frac{1}{H} = \frac{1}{n} \sum \left[\frac{1}{x_i} \right]$$

The geometric mean is mainly useful in calculating relative values such as index numbers and in averaging ratios and rates.

3.3.2 Mean of Grouped Data

If we refer to our data in Table 2.6, we have the following:

Wt. in kg	x_c	f	$f\,x_c$
52–56	54	3	162
56–60	58	3	174
60–64	62	9	558
64–68	66	9	594
68–72	70	8	560
72–76	74	2	148
76–80	78	1	78
Total		35	2274

Hence, using Eq. (3.1),

$$\bar{x} = \frac{2274}{35} = 64.9714$$

Consider, the data below in which the midpoints of the intervals as well as the frequency values are given.

x_c	f	$f\,x_c$
57	1	57
52	1	52
47	3	141
42	4	168
37	6	222
32	7	224
27	12	324
22	6	132
17	8	136
12	2	24
Total	50	1480

Here again, we have

$$\bar{x} = \frac{\sum f x_c}{n} = \frac{1480}{50} = 29.60.$$

3.3.3 The Median

The sample *median* of a set of n measurements $x_1, x_2, \cdots, x_n$ is the middle value when the measurements are arranged in order of magnitude, e.g., from smallest to largest. If n is an odd number, there is a unique middle value and it is the median. If n is an even number, there are two middle values and the median is defined as their average.

Roughly speaking, the median is the value that divides the data into two equal halves. In other words, 50 % of the data lie below the median and 50 % lie above it. Simple formulae for finding the median are given below.

Case I: n Odd

If the number of observations n is odd, then the position of the median in an ordered array and given by,

$$\left(\frac{n+1}{2}\right).$$

For example, if $n = 15$, then the median is in the 8th position. Similarly, if $n = 25$, then the median is in the 13th position.

Case II: n Even

If the number of observations n is even, then the median is the average of the two observations whose positions in the ordered array are given by,

$$\left(\frac{n}{2}\right)^{\text{th}}, \quad \text{and} \quad \left(\frac{n}{2}+1\right)^{\text{th}}$$

i.e., the median equals,

$$\frac{1}{2}\left[\left(\frac{n}{2}\right)^{\text{th}} + \left(\frac{n}{2}+1\right)^{\text{th}}\right].$$

For example, if $n = 10$, then the median is the average of the fifth and the sixth observations. Similarly, if $n = 20$, the median is also given by the average of the 10th and the 11th observations in the ordered array.

Examples

To find the median of the birth weight data given in the example above, we first order the measurements from smallest to largest as:

$$6.4, 7.8, 8.1, 9.2, 10.5$$

The middle value is 8.1 and the median is therefore 8.1 pounds.

Consider another example, the monthly incomes in Naira of eight members of an engineering firm in Lagos are 500, 750, 600, 550, 550, 700, 2000, and 550. To calculate the mean and median income, we note that $\sum x = 6200$ and hence, $\bar{x} = 775$. Thus, the mean monthly income for the group is N 775.00. To find the median, first we order the data. The ordered values are:

$$500, \quad 550, \quad 550, \quad 550, \quad 600, \quad 700, \quad 750, \quad 2000.$$

Here, $n = 8$, an even number. Thus the median is the average of the fourth and fifth observations, i.e., the median income is:

$$M = \frac{550 + 600}{2} = N\,575.$$

Here the median of N 575 appears to be a more sensible measure of the center than the mean.

Other measures of location are *quartiles*, *deciles*, and *percentiles*. These are points which divide distributions of ranked values (e.g., smallest to largest) into quarters, tenths, and hundredths respectively. Thus, the median is the second quartile, fifth decile, and fiftieth percentile. These are discussed in the next section.

3.4 Percentiles

The pth percentile of a data array (arranged in order of magnitude) $x_1, x_2, \cdots, x_n$ is number x such that *at least* $p\,\%$ of the data fall below it and $(100-p)\,\%$ of the data fall above it. To calculate the pth percentile of a data set, we do the following:

- Arrange the data in ascending order
- Compute and index i using the expression:

$$i = \left(\frac{p}{100}\right) n \tag{3.2}$$

where p is the percentile of interest and n is the number of observations in the data set.

(a) if i is not an integer, then round up to the next highest integer and this will denote the position of the *pth* percentile.
(b) if i is an integer, then the pth percentile is the arithmetic mean of the ith and $(i+1)$th observations in the ordered array.

Example: Finding the Specific Percentile from a Data Set

Consider the data in Chap. 2 relating to the weights of heads of households in kilograms. The data has been arranged in order of magnitude from smallest to largest.

```
WGT
   52   53   55   57   59   59   60   60   60   60   61   63   63   63   63
   64   64   64   64   64   64   66   66   67   68   68   68   68   69   70
   70   71   72   75   76
```

The 80th percentile is computed as follows:

$$i = \left(\frac{80}{100}\right) 35 = 28$$

Since this is an integer, the 80th percentile therefore is the average of the 28th and 29th observations in the array, i.e., the 80th percentile is $\dfrac{68+69}{2} = 68.5$.

Alternatively, suppose we wish to know the percentile ranking of a head of household, whose weight is 71 kg. Here, we compute this as:

$$\frac{\#\text{ of data values less than } 71}{35} \times 100 = \left(\frac{31}{35}\right) 100 = 88.6 \approx 89$$

i.e., this will correspond to the 89th percentile. In general, we obtain the ranking for a specific data value x as:

$$\frac{\#\text{ of data values less than } x}{n} \times 100, \quad \text{and round up to the nearest integer}$$

3.4.1 Quartiles

The most common percentiles of interest are *Quartiles*, which divide the data set into four equal parts and are defined as follows:

$$Q_1 = \text{first quartile, or 25th percentile}$$
$$Q_2 = \text{second quartile, or 50th percentile}$$
$$Q_3 = \text{third quartile, or 75th percentile}$$

To obtain Q_1, Q_2, and Q_3, we use the expression in (3.2) noting that each corresponds to $i = 25, 50$, and 75 respectively. Thus for Q_1, we have,

$$i = \left(\frac{25}{100}\right) 35 = 8.75$$

Rounding this up, Q_1 is therefore the 9th observation, i.e., $Q_1 = 60$.
Similarly for Q_2, we have,

$$i = \left(\frac{50}{100}\right) 35 = 17.5$$

Rounding up again, gives Q_2 (median) as the 18th observation, i.e., $Q_2 = 64$.
The index for Q_3 is also computed as:

$$i = \left(\frac{75}{100}\right) 35 = 26.25.$$

Rounding up again gives Q_3 as the 27th observation, i.e., $Q_3 = 68$.

3.4.2 Checking for Outliers with Quartiles

Having computed Q_1 and Q_3, we can then compute the *interquartile range* or IQR which is defined as, IQR $= Q_3 - Q_1 = 68 - 60 = 8$. The inner fences are located at 1.5(IQR) distances below Q_1 and above Q_3, thus, $Q_1 - 1.5(IQR) = 60 - 1.5(8) = 48$ and $Q_3 + 1.5(IQR) = 68 + 1.5(8) = 80$ respectively. If our data value falls either below the lower inner fence or above the upper inner fence, then we consider such data as a possible outlier. To be truly certain the

data value is an outlier, it must fall outside the outer fences which are $3(IQR)$ distances below Q_1 and above Q_3 respectively, i.e., our data value must be less than $Q_1 - 3(IQR) = 60 - 24 = 36$ and above $Q_3 + 3(IQR) = 68 + 24 = 92$; a data value x is an outright outlier if $x \leq 36$ or $x \geq 92$ in this case.

The Box plot displayed below, gives a graph of the *five number summary* for our data, viz., the minimum, Q_1, Q_2, Q_3, and the maximum. For our data, these values are respectively, $\{52, 60, 64, 68, 76\}$ and these can also be obtained in MINITAB with the DESCRIBE statement as displayed below. MINITAB gives these values as well as the mean and standard deviation.

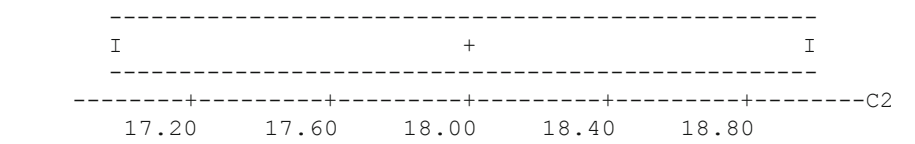
```
MTB > describe c1

Descriptive Statistics: WGT

Variable  N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3  Maximum
WGT      35   0  64.171    0.951  5.628   52.000  60.000  64.000  68.000   76.000


MTB > boxplot c1


Boxplot

                          -----------------
              ----------------I       +       I----------------
                          -----------------
         +---------+---------+---------+---------+---------+------WGT
       50.0      55.0      60.0      65.0      70.0      75.0
```

3.5 The Boxplot

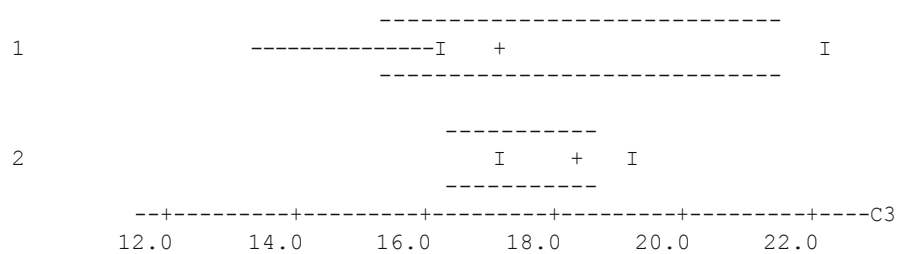
The box plot, apart from displaying the five number summary data can also be employed to compare distributions. Below we have constructed box plots for each of two data sets separately and jointly for comparative purposes.

```
MTB > DESCRIBE C1-C2

                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
C1             10    18.00    17.00    18.13     3.46     1.10
C2             10   18.000   18.000   18.000    0.816    0.258

               MIN      MAX       Q1       Q3
C1           13.00    22.00    15.50    22.00
C2         17.000   19.000   17.000   19.000

MTB > BOXPLOT C1


                             ------------------------------
              ---------------I    +                         I
                             ------------------------------
     --+---------+---------+---------+---------+---------+----C1
    12.0      14.0      16.0      18.0      20.0      22.0
```

```
MTB > BOXPLOT C2
```




```
MTB > STACK C1-C2 C3;
SUBC> SUBSCRIPT C4.

MTB > BOXPLOT C3;
SUBC> BY C4.
```



3.5.1 Mean of Grouped Data

The mean of grouped data is often easily obtained by coding the data, generally, this usually involves a variate x_i which can be transformed into variate U_i with the following transformation:

$$U_i = \frac{x_i - x_0}{c}$$

or $x_i = cU_i + x_0$, where, x_0 is any value of x taken as an arbitrary average and c is the class interval width. With this transformation, the mean is computed as:

$$\bar{x} = x_0 + \frac{c\sum f_i U_i}{\sum f_i}. \tag{3.3}$$

For the data in the preceding section, suppose we chose $x_0 = 37$, then we have the table:

x_i	f_i	U_i	f_iU_i	$f_iU_i^2$
57	1	4	4	16
52	1	3	3	9
47	3	2	6	12
42	4	1	4	4
37	6	0	0	0
32	7	−1	−7	7
27	12	−2	−24	48
22	6	−3	−18	54
17	8	−4	−32	128
12	2	−5	−10	50
Totals	50		−74	328

From the table, we have,

$$\frac{\sum f_iU_i}{\sum f_i} = \frac{-74}{50} = -1.48$$

$$\begin{aligned}\bar{x} &= 37.0 + c(-1.48) = 37 + 5(-1.48), \quad \text{since c equals 5} \\ &= 29.60.\end{aligned}$$

We observe that this value of $\bar{x}$ agrees with the value that we obtained earlier for the ungrouped data. However, the arithmetic of the coded procedure is much simpler and this latter procedure is often recommended when hand calculators are not available or simply to simplify the calculations.

3.5.2 The Median of Grouped Data

We shall illustrate again with the grouped data above. Here, $\sum f_i = n = 50$, hence, the median is somewhere half way of this, i.e., 25. Thus, counting frequencies from the bottom upward (i.e., from below), we find 2+8+6+12 = 28 cases, three more than what we want—this is at an x value of 27. To make 25 cases exactly, we need 9 of the 12 cases in this class. The median lies somewhere within the interval 25–29 whose exact limits are 24.5 and 29.5. Thus, we interpolate that we must go $9/12 = 3/4$ of the way. The total distance is 5. Hence, $\frac{3}{4} \times 5 = 3.75$. Thus adding this to the lower limit, we have $24.5 + 3.75 = 28.25$ as the median. Generally, the median M is given as (if interpolation is from below),

$$M = l_0 + \left[\frac{n/2 - F_b}{f_0}\right] c \tag{3.4}$$

where, l_0 equals the exact lower limit of the class interval containing the median, F_b is the sum of all frequencies below l_0 and f_0 is the frequency of

the interval containing the median. c and n are defined as usual. Employing this formula, for the example above, we have,

$$M = 24.5 + \frac{(25 - 16)5}{12} = 28.25$$

A slightly similar formula is available for computing the median of grouped data by interpolation from above, but the above will suffice for our purpose.

3.5.3 The Mode

The mode is the item that occurs most often in a distribution, i.e., the item that has the highest frequency. The procedure for obtaining the mode is by simply putting the observations in form of a frequency distribution and picking the one that has the highest frequency.

As an example, the age in (years) of ten students are

$$14, \quad 15, \quad 16, \quad 16, \quad 17, \quad 17, \quad 22, \quad 22, \quad 22, \quad 22.$$

Then a frequency display of the data is as follows:

x	f
14	1
15	1
16	2
17	2
22	4

The highest frequency here is 4, and hence, the mode is 22 years. The above can be implemented in MINITAB by specifying

TALLY C1

where the data is assumed to be stored in column one (C1).

As another example, a newspaper wants to predict an election result of a certain constituency. Five political parties (A, B, C, D, and E) are in the race. The newspaper then interviews a selected sample of 1000 potential voters, the results of this interview are summarized below.

Party	f
A	13
B	294
C	298
D	344
E	9
Undecided	42
Total	1000

Based on the above summary table, Party D is the favorite on the basis of the result from the survey. Therefore, the mode or the most popular party is party D.

3.5.4 Comparisons Between Mean, Median, and Mode

The mean, median and mode are all measures of central tendency (or location) in a specific way. The question that one is often facing in practical application is: "Which one of the measures of central tendency is most appropriate?" The question is not easily answered. As will be illustrated later, the one we use depends on the objectives for conducting an enquiry and the type of data gathered. Let us now compare them.

(i) The mean and mode (when it exists) are easy to calculate. However when the number of observations are large, it is tedious putting a set of data in an array, thus the median may be tedious to calculate.
(ii) Mean and median always exist in a distribution, whereas, mode may not exist and if it exists may not be unique. For example, in the summary data below,

x	f	fx
3	3	9
5	2	10
7	3	21
9	1	9
11	3	33
Total	12	82

The above frequency table can be re-written in the ungrouped form as:

$$3, \quad 3, \quad 3, \quad 5, \quad 5, \quad 7, \quad 7, \quad 7, \quad 9, \quad 11, \quad 11, \quad 11.$$

Hence,

$$\bar{x} = \frac{82}{12} = 6.83$$

and

$$M = \frac{(6\text{th} + 7\text{th})}{2} = \frac{7+7}{2} = 7.$$

The mode in this case equals 3, 7, 11, i.e., the mode takes three different values at the central value. Thus the mode is not unique in this case, and we would describe such a data as having a *tri-modal* distribution.

(iii) Generally, the median provides a better measure of the center when there are extremely large or small observations in a set of data. For example, in the above example of monthly income of eight people, we had $\bar{x} = 775$, the median M = 575. The extreme value item is 2000. The median is more central than the mean, i.e., the median is not affected by extreme or abnormal values (or outliers).

(iv) When faced with qualitative data, mean and median are meaningless. Thus, the mode is the only appropriate means of measure of central location in this case.

Example A manufacturing company is carrying out a market research of the use of its brands of University vests in three different colors. He noted 30 students in a certain class wearing vest of the following colors. Green, Yellow, Brown, Blue, Green, Yellow, Red; Blue, Brown, Yellow, Blue, Black, Blue, Brown, Red; Blue, Green, Blue, Yellow, Red; Blue, Red, Brown, Blue, Yellow, Brown, Blue, Black, Yellow, Blue. Find the best average.

Solution Surely mean and median are meaningless in this problem. So our alternative is the modal choice of color.

Color	Frequency
Green	3
Yellow	6
Brown	5
Blue	10
Red	4
Black	2

The modal color is blue.

3.6 Relationship Between Mean, Median, and Mode

1. For symmetrical distributions (Fig. 3.1), the mean, mode, and median coincide, i.e., Mean = Mode = Median or $\bar{X} = \text{M} = \text{m}$ theoretically. However, for real life data, they will seem to be close.
2. For skewed distributions, the following empirical relationship exists between the three measures, viz.,

$$\text{Mode} = \text{Mean} - 3\,(\text{Mean} - \text{Median}) \tag{3.5}$$

Of course we can write the above as,

$$\text{Mean-Mode} = 3\,(\text{Mean-Median}) \quad \text{or,}$$

$$\text{Mean} - \text{Median} = \frac{1}{3}\,(\text{Mean} - \text{Mode}). \tag{3.6}$$

In general, the above is more succinctly written as,

$$\text{Mode} = 3\ \text{Median} - 2\ \text{Mean} \tag{3.7}$$

i.e.,

$$\text{m} = 3\ \text{M} - 2\ \bar{x}$$

3. For positively skewed or right-skewed distributions (see Fig. 3.3) $\text{m} < \text{M} < \bar{x}$. Similarly, for negatively skewed or left-skewed distributions (see Fig. 3.2), we have $\bar{x} < \text{M} < \text{m}$.

In the next figure, we present the histogram and corresponding probability plot of a right-skewed distribution.

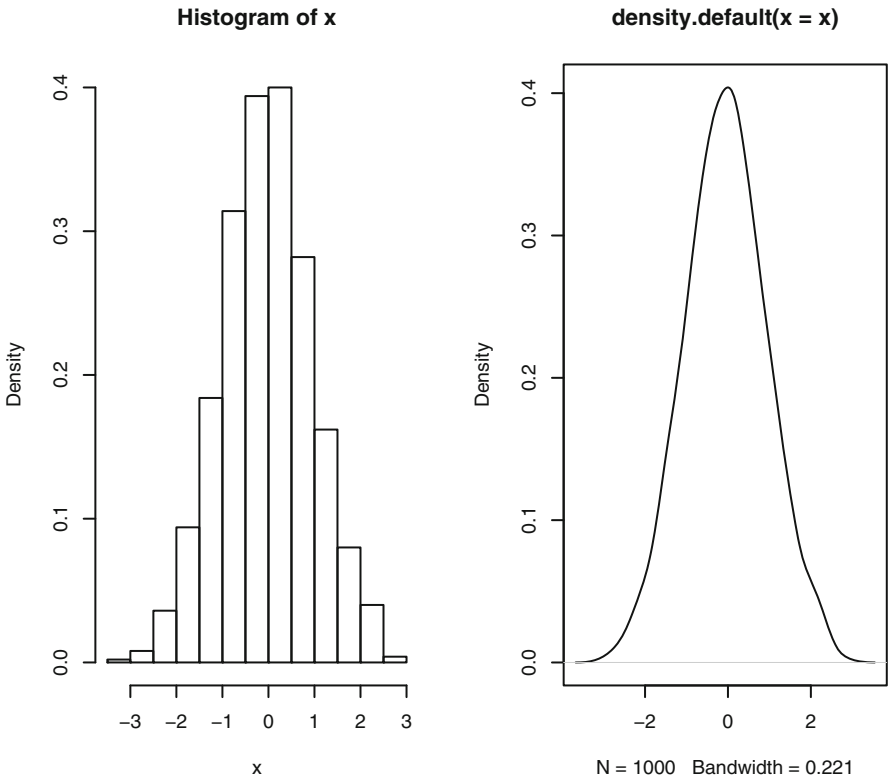


Fig. 3.1 A symmetric distribution. Here, $\bar{X} = M$

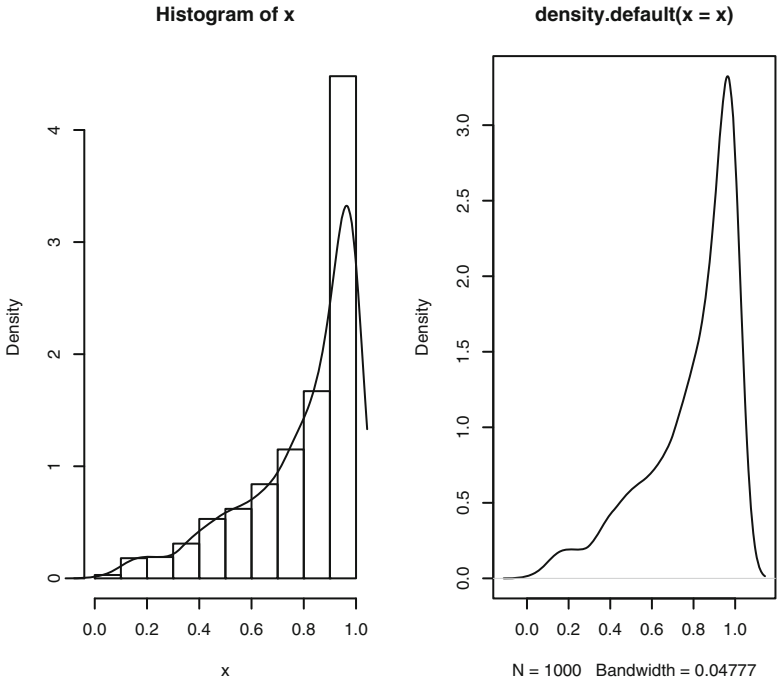


Fig. 3.2 A left-skewed distribution. Here, $\bar{X} << M$

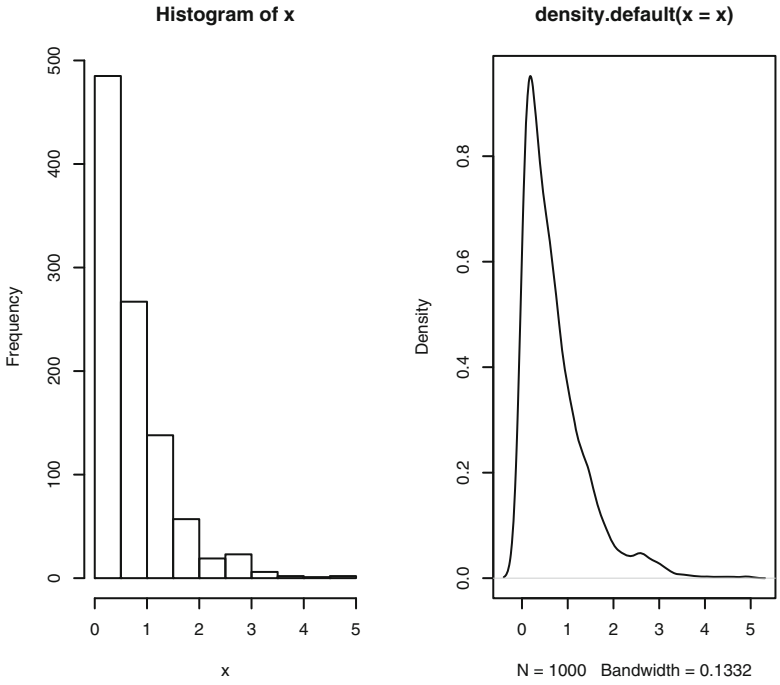


Fig. 3.3 A right-skewed distribution. Here, $\bar{X} >> M$

3.7 Measures of Variation or Dispersion

No two objects are exactly alike. Even "identical twins" differ. The universe is filled with objects and individuals which vary from one another by some characteristics more so in biological and medical data. The averages (mean, median, and mode) or measures of location which we have treated measure the center of data. The assumption is that all data in the observation takes a single value. This in most cases does not hold.

Therefore, there is a need to measure the degree of spread or variation of our data from one another and (or around the average). This degree of spread is called *variation* or *dispersion*.

When the value of our observations are the same, then there is no variation and our degree of variation equals zero.

Example

The age (in years) of ten boys in form V from two different secondary schools are:

School I 13, 14, 16, 16, 17, 17, 21, 22, 22, 22. Hence, $\bar{x}_1 = \dfrac{180}{10} = 18$ years

School II 18, 18, 19, 17, 19, 19, 17, 18, 18, 17. Hence again, $\bar{x}_2 = \dfrac{180}{10} = 18$ years

We can conclude that the average age of boys in the two schools are the same. However, a careful look at the data shows that the ten boys in School II are far more uniform than those in School I. Thus School II boys are likely to behave more like 18-year-olds than those in School I. A measure of variation is out to measure this degree of variability.

We discuss below the various measures of variation that have been suggested from various literature.

3.7.1 The Range

The range which is defined as *Highest data value − Lowest data value*, is a basic measure of variability or spread. For example, in the two schools data above, we have for both the schools.

School I: Range $= 22 - 13 = 9$

School II: Range $= 19 - 17 = 2$.

Thus we see straight from the values of the ranges for both data sets that the data for School I is more widely spread than those from School II. Hence, we would expect the boys in School II to behave more like 18 years old than those from School I. In other words, the data in School II is said to be more *homogeneous* than those from School I.

3.7.2 Variance and Standard Deviation

The variance of a sample is always represented by S^2 and the standard deviation will be represented by the square root of S^2, i.e., $S = \sqrt{S^2}$. While the standard deviation is an absolute measure of dispersion, it is however, measured in units—does it depend on the units of measurement? The *coefficient of variation* on the other hand is a relative measure of dispersion based on the standard deviation and is defined as,

$$\text{CV} = \frac{s}{\bar{x}} \times 100\ \% \tag{3.8}$$

The coefficient of variation being a ratio, it is a dimensionless quantity. Thus for comparing the variability of say, two distributions, we compute their CVs. The distribution with the smaller CV would be more homogeneous than the other with a higher CV.

We consider in the next section two cases of obtaining sample variances for both ungrouped and grouped data.

3.7.3 Case I: Variance of Ungrouped Data

Given a set of values $x_1, x_2, \cdots, x_n$, the variance is defined as

$$\begin{aligned} s^2 &= \frac{\sum (x_i - \bar{x})^2}{n-1} \\ &= \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right] \end{aligned}$$

Example

To find the variance of the age of boys in School I above, we note that $\bar{x} = 18$, hence,

X (age)	$x - \bar{x}$	$(x - \bar{x})^2$
13	−5	25
14	−4	16
16	−2	4
16	−2	4
17	−1	1
17	−1	1
21	3	9
22	4	16
22	4	16
22	4	16
Total	0	108

Variance $= s^2 = \dfrac{108}{10-1} = \dfrac{108}{9} = 12$ years^2. Hence, $s = \sqrt{S^2} = \sqrt{12} =$ 3.4641 years. Alternatively, we could use the second formula for finding the variance. Here, we have,

$$\sum x_i^2 = 13^2 + 14^2 + 16^2 + 16^2 + 17^2 + \cdots + 22^2 = 3348$$

$n = 10$, and $\sum x_i = 180$, thus,

$$\sum x_i^2 - \frac{(\sum x)^2}{10} = 3348 - \frac{180^2}{10} = 3348 - 3240 = 108.$$

Hence, $s^2 = \frac{108}{10-1} = 12$ years^2 and the corresponding coefficient of variation CV is $(3.4641/18) \times 100 = 19.2\,\%$. This second approach is most useful when hand calculators that perform statistical functions are available.
For the second school, i.e., School II, we also have $\bar{x} = 18$ and $s^2 = \frac{2}{3}$years^2, i.e., $s = 0.8165$ years, and hence, the CV $= (0.8165/18) \times 100 = 4.5\,\%$

The results above support our initial observation that the age of boys in School I vary more widely than the age of boys in School II since VAR(I) > VAR(II) or CV(I) > CV(II).

Steps to Follow when Calculating S^2 for Case I

(i) Obtain the mean.
(ii) From each observation, deduct the mean to obtain the deviations $x - \bar{x}$.
(iii) Square each deviation to obtain $(x - \bar{x})^2$.
(iv) Obtain the sum $\sum (x - \bar{x})^2$.
(v) Divide this sum by $n - 1$.

As mentioned earlier, most hand calculators these days have facilities for the calculation of means and variances for ungrouped data. The above can be implemented in MINITAB as follows:

```
MEAN C1 K1
LET C2=C1-K1
NOTE NOW FIND THE SUM OF DEVIATIONS
SUM C2
LET C3=C2*C2
NOTE FIND SUM OF SQUARED DEVIATIONS AND PUT IN K2
SUM C3 K2
LET K3=K3/(N(C1)-1)
PRINT K3
```

of course we could simply ask MINITAB to give us the following:

```
MEAN C1
STDEV C1 k1
LET K2=K1*K1 (VARIANCE)
MEDIAN C1
```

3.7.4 Calculating the Variance of Grouped Data

The data below gives the weight in kilograms of 100 students at a given University.

Weight (kg)	Frequency
60–62	5
63–65	18
66–68	42
69–71	27
72–74	8

To find the variance of the above grouped data, we form the table below.

Wt.(kg) (kg)	Class mark x_i	Freq. f_i	$f_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
60–62	61	5	305	−5.45	41.6025	208.0125
63–65	64	18	1152	−3.45	11.9025	214.2450
66–68	67	42	2814	−0.45	0.2025	8.5030
69–71	70	27	1890	2.55	6.5025	175.5675
72–74	73	8	584	5.55	30.8025	246.4200
Totals		100	6745			852.7500

From the above table, we have

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{6745}{100} = 67.45$$

$$S^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i - 1} = \frac{852.7500}{99} = 8.6136$$

Hence, $S = \sqrt{8.6136} = 2.9349$ kg.

Another formula for computing the variance of grouped frequency is:

$$\frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i - 1} = \left[\frac{\sum f_i x_i^2}{\sum f_i} - \bar{x}^2\right]\left[\frac{\sum f_i}{\sum f_i - 1}\right]$$

Implementing this for the example above, we have,

x_i	f_i	$f_i x_i$	$f_i x_i^2$
61	5	305	18,605
64	18	1152	73,728
67	42	2814	188,538
70	27	1890	132,300
73	8	584	42,632
Total	100	6745	455,803

Hence, $\bar{x} = \frac{6745}{100} = 67.45$, and,

$$\begin{aligned}\frac{\sum f_i x_i^2}{\sum f_i} - \bar{x}^2 &= \frac{455803}{100} - 67.45^2 \\ &= 4558.03 - 4549.5025 \\ &= 8.5275.\end{aligned}$$

Hence, $S^2 = \frac{100}{99} \times 8.5275 = 8.6136$ and $S = \sqrt{8.6136} = 2.9349$ kg.

3.7.5 Use of Coding to Simplify Calculations

The calculations above could have been simplified if we had coded the data. We give an example of the procedure involved for using this method. First, suppose we define $u = (x - 67)/3$, where 67 is an arbitrary value and 3 is the class width in the example above. Then, $x = 67 + 3u$. Thus, $E(X) = \bar{x} = 67 + 3\bar{u}$. To obtain the variance, we note that the variance of a constant is zero. Further, if $\text{Var}(y) = \sigma^2$, then $\text{Var}(ay) = a^2\sigma^2$. Similarly, $\text{Var}\left[\frac{x}{b}\right] = \frac{\sigma^2}{b^2}$, where a and b are constants. Hence Var(x) above equals $3^2\text{Var}(u) = 9\text{Var}(u)$. We illustrate this in the following table.

x_i	u_i	f_i	$f_i u_i$	$f_i u_i^2$
61	−2	5	−10	20
64	−1	18	−18	18
67	0	42	0	0
70	1	27	27	27
73	2	8	16	32
Total	100	15	97	

$$\begin{aligned}\bar{u} &= \frac{15}{100} = 0.15, \text{ i.e.,} \\ \bar{x} &= 67 + 3\bar{u} = 67 + 3(0.15) \\ &= 67.45\end{aligned}$$

and

$$\text{Var}(u) = \frac{100}{99}\left\{\frac{97}{100} - 0.15^2\right\} = 0.9571.$$

Hence, $\text{Var(x)} = c^2\ \text{Var(u)} = 9(0.9571) = 8.6136$. This agrees with the earlier result obtained and we see that the sums are not too large in the computational table.

3.8 Empirical Rule

The mean and the standard deviation of a data set can be used to find the proportion of the total observations that fall within a given interval about the mean. We mostly consider the intervals (i) $\bar{x} \pm s$, (ii) $\bar{x} \pm 2s$, (iii) $\bar{x} \pm 3s$. The empirical rule relates to only *mound-shaped* or *bell-shaped* distribution. The rule states that for mound shaped distribution, approximately:

(a) 68 % of the observations fall within the interval $\bar{x} \pm s$, i.e., within one standard deviation of the mean.
(b) 95 % of the observations fall within the interval $\bar{x} \pm 2s$, i.e., within two standard deviations of the mean.
(c) 99.7 % of the observations fall within the interval $\bar{x} \pm 3s$, i.e., within three standard deviations of the mean.

As an example, consider the data in Table 2.1 (see p. 11). There are 35 observations. $\bar{x} = 64.171$ and $s = 5.628$. Hence, $\bar{x} \pm s = 64.171 \pm 5.628 = [58.54, 69.80]$. With our data arranged in order of magnitude and counting how many observations are between 59 and 69, we have 25, (from the MINITAB display below) that is, $\frac{25}{35} = 0.714$ or 71.4 % of the total observations lie within this interval. Corresponding intervals for $\bar{x} \pm 2s$ and $\bar{x} \pm 3s$ are computed as follows:

$$\bar{x} \pm 2s = 64.171 \pm 2(5.628) = 64.171 \pm 11.256 = [52.92, 75.43]$$

$$\bar{x} \pm 3s = 64.171 \pm 3(5.628) = 64.171 \pm 16.884 = [47.29, 81.06]$$

From the above, we see that approximately, 33 and 35 observations fall respectively in the intervals $\bar{x} \pm 2s$ and $\bar{x} \pm 3s$. Consequently, we say that for this data set, approximately 94.2 and 100 % of the data fall within these intervals respectively. These results are not consistent with the empirical rule, hence, we can rightly conclude that this data set is not mound or bell shaped, i.e., it is skewed.

```
MTB > sort c1 c2
MTB > print c2

Data Display

C2
   52   53   55   57   59   59   60   60   60   60   61   63   63   63   63
   64   64   64   64   64   64   66   66   67   68   68   68   68   69   70
   70   71   72   75   76
```

3.9 Exercises

1. Five measurements in a data set are $x_1 = 7,\ x_2 = 5,\ x_3 = 6,\ x_4 = 8$ and $x_5 = 6$. Compute the numerical values of

$$(i)\sum_{i=1}^{5} x_i \quad (ii)\sum_{i=1}^{4} x_i, \quad (iii)\sum_{i=1}^{5} 2x_i, \quad (iv)\sum_{i=1}^{5}(x_i - 6), \quad (v)\sum_{i=1}^{5}(x_i - 6)^2$$

2. Demonstrate your familiarity with the summation notation by evaluating the following expressions when $x_1 = 1,\ x_2 = 2,\ x_3 = 4$ and $x_4 = 5$

$$(i)\sum_{i=1}^{4} x_i \quad (ii)\sum_{i=1}^{2}(x_i - 4) \quad (iii)\sum(x_i - 2), \quad (iv)\ \sum_{i=1}^{4}(x_i - 2)^2$$

3. The residues of fungicide measured in parts per million in a random sample of 50 fresh oranges harvested 40 days after receiving the last of last of six sprays of the fungicide were as follows:

1.63	1.40	1.64	1.30	1.49	1.58	1.03	1.06	1.33
1.52	1.87	1.83	1.97	1.62	1.21	1.01	1.14	1.58
1.43	1.41	1.51	1.15	1.61	1.10	1.03	1.84	1.61
1.71	1.32	1.29	1.82	1.99	1.43	1.53	1.56	1.48
1.82	1.81	1.21	1.73	1.59	1.99	1.34	1.23	1.65
1.20	1.76	1.54	1.58	1.99				

   Arrange the results in a grouped frequency table and calculate the sample mean, variance, and median from this table. Comment on your results.

4. The angle between two adjacent toes were measured from radiographs of the affected feet of 50 young adults undergoing treatment for a foot abnormality.

ANGLE BETWEEN TOES (DEGREE)

42	32	33	33	29	31	33	29	40	31
27	30	29	43	34	29	34	29	28	30
36	46	30	41	45	31	30	33	29	29
33	35	37	27	29	43	32	27	32	32
39	41	44	32	35	29	31	28	28	29

   Similar measurements were made on the feet of 40 normal young adults.

12	18	13	15	16	12	15	18	15	15
17	15	16	17	17	16	18	13	12	15
14	15	12	14	14	18	17	18	12	14
13	12	12	14	17	16	12	16	15	13

Obtain the mean, variance, and median for these sets of data. Now choose a suitable class interval, arrange the results in a frequency table and compute the same statistics. Compare your results.

5. Calculate the means and median of the following sets of data for the situations.

   (i) Ungrouped
   (ii) Grouped using a suitable class interval width

   (a) The time interval in minutes between the arrival of successive customers at a cash desk of a self service store was measured over 56 customers and the results are given below:

1.05	1.68	0.78	1.10	0.32	1.61	0.10	3.12	0.21
3.30	0.15	0.54	2.16	1.14	0.16	0.31	0.91	0.18
0.57	0.65	4.60	1.72	0.52	2.32	0.08	2.68	1.16
1.19	0.11	0.05	3.70	1.48	3.80	2.08	0.09	1.76
2.71	2.12	2.81	0.04	1.16	0.62	0.58	0.57	0.04
0.63	1.21	0.01						

   (b) The intelligence quotients of 100 children are given below:

72	112	100	116	99	111	85	82	08	85	94	91
118	103	102	133	98	106	92	102	115	109	100	57
108	77	94	121	100	107	104	67	11	88	87	97
102	98	101	88	90	93	85	107	80	106	120	91
101	103	109	100	127	107	112	98	83	98	89	106
79	117	85	94	119	93	100	90	102	87	95	117
142	94	93	72	98	105	122	104	104	79	102	104
107	97	100	109	103	107	106	96	83	107	102	110
102	76	98	88								

6. The following sample of serum cholinesterase indices in normal individuals is taken from Kaufman (1954). The data has been sorted. Use this to find the followings:

INDICES

1.03	1.03	1.04	1.04	1.04	1.06	1.08	1.09	1.13	1.15
1.15	1.15	1.16	1.16	1.18	1.21	1.22	1.23	1.24	1.24
1.25	1.26	1.27	1.30	1.32	1.32	1.35	1.35	1.37	1.39
1.40	1.40	1.42	1.43	1.44	1.44	1.46	1.48	1.51	1.52
1.52	1.54	1.55	1.57	1.59	1.59	1.61	1.65	1.65	1.65
1.67	1.68	1.69	1.70	1.70	1.70	1.71	1.72	1.75	1.75
1.75	1.78	1.82	1.83	1.84	1.86	1.86	1.88	1.89	1.91
1.92	1.92	1.92	1.92	1.92	1.93	1.94	1.95	2.02	2.10
2.12	2.13	2.14	2.14	2.15	2.17	2.23	2.26	2.27	2.29
2.52	2.54	2.55	2.59	2.60	2.65	2.67	2.76	3.09	3.27

(i) The mean, the median, and mode for the data if $\sum x = 170.40$.
(ii) The upper and lower quartiles (Q_1, Q_3) and hence obtain IQR, the interquartile range.
(iii) Obtain the *five-number summary* for the data.
(iv) Obtain variance s^2 for this data if $\sum x^2 = 313.60$. Find the percentage of the data values that fall between $\bar{x} \pm s$.

7. The data below relate to serum CK concentrations (creatine phosphokinase) of 36 male volunteers. The data has been sorted from smallest to largest.

```
CSK
   25   42   48   57   58   60   62   64   67   68   70   78   82
   83   84   92   93   94   95   95  100  101  104  110  110  113
  118  119  121  123  139  145  151  163  201  203
```

For these data, $\sum x = 3538.0$ and $\sum x^2 = 404,778.00$.

(a) Compute $\bar{x}$ and s^2, the sample variance and hence s the sample standard deviation.
(b) What percentage of the data fall in the interval $\bar{y} \pm s$?
(c) What percentage of the data fall in the interval $\bar{y} \pm 2s$?
(d) What percentage of the data fall in the interval $\bar{y} \pm 3s$?
(e) Based on your results above, what can you say about the shape of the distribution of these data.

8. The mean and the median of a set of test scores are 75 and 60 respectively. Circle the letter of the statement which is most defensible.

(a) The distribution of test scores is skewed to the left.
(b) The distribution of test scores is skewed to the right.
(c) Half of the test scores are greater than 60.
(d) A few test scores are very small, pulling the median down.
(e) Both (a) and (d) are true.

9. The data below give the weight in kilograms of 100 college students taken at random in Fall 2006.

Weight (kg)	Frequency (f)
60–62	5
63–65	18
66–68	42
69–71	27
72–74	8

Find the mean and median for the grouped data. Also, compute Shannon's index of diversity and interpret your result.

Chapter 4
Probability and Probability Distributions

4.1 Introduction

The concept of probability is relevant to experiments that have some uncertain outcomes. These are the situations in which, despite every effort to maintain fixed conditions, some variation in the result during repeated trials of the experiment is unavoidable. As used in here, the term "experiment" is not restricted to laboratory experiments but includes any activity that results in the collection of data pertaining to phenomena that exhibit variation. The domain of probability encompasses all phenomena for which outcomes cannot be exactly predicted in advance.

Examples of experiments with uncertain outcomes are

(i) Tossing a coin
(ii) Rolling a die
(iii) Gender of the first two newborns in town tomorrow
(iv) Tossing two coins
(v) Rolling two dice
(vi) Planting a seed, etc.

Though, in the above examples, each experimental outcome is unpredictable, we can describe the collection of all possible outcomes as:

Definition

The collection of all possible outcomes of an experiment is called the *Sample Space* of outcomes and each distinct outcome is called a *simple event*, an elementary outcome or an element of the sample space. They are usually denoted by S and E respectively.

Before we discuss the concept of probability and probability distributions completely, let us familiarize ourselves with some counting methods, that will be most useful to us in this chapter.

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_4,

4.2 Counting Methods

In calculating probabilities, it is very essential that we be able to count sample points corresponding to S and E in the event. However, this sometimes becomes a tedious job, and thus compact counting methods are necessary. A branch of Algebra, called "Permutations" and "Combinations" is very useful here.

Suppose two operations A and B are carried out, and if there are "m" different ways of carrying out A and "k" different ways of carrying out B, then the combined operation of A and B may be carried out in $m \times k = mk$ different ways.

4.2.1 Permutation

The number of permutations (or arrangements) of n distinct objects, taken all together is

$$n! = n(n-1)(n-2)\cdots \times 2 \times 1$$

$$0! = 1$$

$$1! = 1 \times 1 = 1$$

$$2! = 2 \times 1 = 2$$

$$3! = 3 \times 2 \times 1 = 6$$

$$4! = 4 \times 3 \times 2 \times 1 = 24$$

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$$

$$6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$$

and so on. Note that $n!$ is read n *factorial*.

Example

Consider the three letters A, B, C, the number of possible arrangements of these three letters will be $3! = 3 \times 2 \times 1 = 6$. These arrangements are given by

$$ABC, ACB, BAC, BCA, CAB, CBA.$$

The number of permutations of n distinct objects taken r at a time, written nP_r is given by:

$$n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

Thus in the above example, $3P_2 = \dfrac{3!}{(3-2)!} = \dfrac{3!}{1!} = 6$ and these are.

$$AB, AC, BC, BA, CA, CB$$

Example

What is the number of permutations of ten distinct digits taken two at a time?

The first digit can be chosen in ten ways and having filled the first place, the second digit can be chosen in nine ways. Hence there are $10 \times 9 = 90$ permutations or

$$10P_2 = \frac{10!}{8!} = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} = 10 \times 9 = 90 \text{ ways}$$

If two or more letters (numerals, items, objects, etc.) are identical (or of the same form), then, the number of permutations is appropriately reduced. For example, Consider the letter ABCDAF. The number of arrangements of the six letters is

$$\frac{6!}{2!} = 360$$

The denominator or divisor is because there are two As in the letter. In general, if we have n objects which are composed of p objects of one kind, q of another, r of another and so on, then the number of different arrangements is given by

$$\frac{n!}{p!q!r!\cdots}$$

Example 1

Out of 12 tulip bulbs to be planted in a row along a border, four are yellow flowers, six are red flowers, and two are orange flowers. How many color patterns could be created? The number of color patterns that can be obtained by varying the planting order will be

$$\frac{12!}{4!6!2!} = 13,860$$

Example 2

In a clinic there are two specialists, one for ear patients and one for nose patients. If during a day, six patients are to arrive, four for ear treatment and two for nose treatment, in how many ways can the duty roster for the specialists be arranged?

The number of ways of arranging the duty roster equals

$$\frac{6!}{4!2!} = 15.$$

4.2.2 Combinations

Combinations deal with the number of distinct arrangements of n objects taken r at a time. It is written as nC_r or $\binom{n}{r}$ and is given by

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

Combinations unlike permutation disregards order.

Example

In the previous example with three letters A, B, and C, we saw that there were $3P_2 = 6$ ways of arranging two letters at a time, namely, AB, AC, BC, BA, CA, and CB. However, the number of distinct ways of arranging two of these at a time is $\binom{3}{2} = 3$, that is,

$$\frac{3!}{2!1!} = 3 \quad \text{ways}$$

and these arrangements are:

$$AB, AC, \textit{ and } BC.$$

Since, these arrangements disregard the ordering of the letters, AB is not distinct from BA.

At the beginning of this chapter, we defined the sample space and simple events. We can now list the simple events and their corresponding sample spaces for each of the examples of experiments with uncertain outcomes enumerated at the beginning of this chapter. These are:

(i) $S = \{\text{H, T}\}$
(ii) $S = \{1, 2, 3, 4, 5, 6\}$
(iii) $S = \{\text{BB, BG, GB, GG}\}$, where BG stands for Boy first, Girl second.
(iv) $S = \{\text{HH, HT, TH, TT}\}$
(v) $S = \{x_i, x_j; 1 \leq i \leq 6, 1 \leq j \leq 6\}$
(vi) $S = \{\text{G, NG}\}$, where G stands for germination and NG for no germination.

The sample space for (v) is as shown in Table 4.1.

Each of the outcomes in the Table 4.1 is equally likely. Note that the events $(1, 2) \neq (2, 1)$, etc.

Table 4.1 Sample space for rolling two dice

	j					
i	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

4.2.3 Probability of an Event

This is the proportion of times an event occurs say, A is expected to occur when the experiment is repeated under identical conditions and is denoted as *P(A)*.

Thus in our above examples for case (i), *P(H)*=1/2 while for case (ii), *P(4)*=$\frac{1}{6}$. The probability of one boy or one girl is given by P(1 boy or 1 girl) $= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.
For all events:

(a) $0 \leq P(A) \leq 1$
(b) $P(S) = 1$

That is, computed probabilities are never greater than 1 or less than 0.

Examples

(i) Twenty discs are marked with the numbers 1–20 inclusive. They are placed in a box and one disc is drawn from it. What is the probability that the number on the disc will be a multiple of 5?

Solution Here $S = 20$ and E(the events) $= \{5, 10, 15, 20\}$. Hence *P(E)* $= 4/20 = 1/5$

(ii) A bag contains five blue balls, three red balls, and two black balls. A ball is drawn at random from the bag then, the probabilities

(a) Prob (red ball) $= \frac{3}{10}$
(b) Prob (black ball) $= \frac{2}{10}$
(c) Prob (not a black ball) $= \frac{8}{10}$ or $1 - \frac{2}{10} = \frac{8}{10}$

(iii) A die is rolled, calculate the probability that

(a) it will give a five?
(b) a number less than three
(c) an even number.

Solution $S = \{1, 2, 3, 4, 5, 6\}$

(a) $E = \{5\}$, hence, $P(E) = \frac{1}{6}$
(b) $E = \{1, 2\}$, hence, $P(E) = \frac{2}{6} = \frac{1}{3}$
(c) $E = \{2, 4, 6\}$, hence, $P(E) = \frac{3}{6} = \frac{1}{2}$

(iv) Refer to Table 4.1. Find the probability that the total sum of the two dice will be (a) 5 (b) less than 5 (c) more than 5 (d) 7 (e) 11.

Solutions $S = 36$ sample points

(a) If total is 5, then $E = \{(1, 4), (4, 1), (2, 3), (3, 2)\}$, hence, $P(E) = \frac{4}{36} = \frac{1}{9}$
(b) If total is less than 5, then $E = \{2, 3, \text{or } 4\}$ and they are given by the following sample points, $E = \{(1, 1), (1, 2), (2, 1), (2, 2), (1, 3), (3, 1)\}$. Hence $P(E) = \frac{6}{36} = \frac{1}{6}$
(c) 1-P(less than or equal to 5) $= 1 - (\frac{1}{6} + \frac{1}{9}) = \frac{13}{18}$. Alternatively, (c) can be obtained as follows: More than 5 implies total sums equal to 6, 7, 8, 9, 10, 11, or 12. Thus,

```
E={(3,3), (1,5), (5,1), (2,4), (4,2), (1,6), (6,1), (2,5)
   (5,2), (3,4), (4,3), (2,6), (6,2), (3,5), (5,3), (4,4)
   (3,6), (6,3), (4,5), (5,4), (4,6), (6,4), (5,5), (5,6)
   (6,5), (6,6)}
```

Hence $P(E) = \frac{26}{36} = \frac{13}{18}$ which gives the same result as in (c) above.

(d) If total $= 7$, then, $E = \{(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)\}$. Therefore, $P(E) = \frac{6}{36} = \frac{1}{6}$
(e) If total $= 11$, then, $E = \{(5, 6), (6, 5)\}$ and $P(E) = \frac{2}{36} = \frac{1}{18}$

Example 4.1.1

When an experimental stimulus is given to an animal, it will either respond or fail to respond. In other words, there are only two possible outcomes when a stimulus is applied to an animal. Either the animal responds (R) or it does not (N). The experiment consists of administering the stimulus to three animals in succession and recording R or N for each animal. Find the probability of the following events

(i) Only one animal responds
(ii) There is a response in the first trial
(iii) Both the first and second animals fail to respond.

Solution

There are two possible outcomes R (response) and N (no response) for each of the animals. Since there are three animals, there are $2 \times 2 \times 2 = 8$ elementary

outcomes for this experiment. These are listed below and are identified by the symbols $E_1, E_2, \cdots, E_8$.

RRR (E_1)	RRN (E_2)	RNR (E_3)	NRR (E_4)
RNN (E_5)	NRN (E_6)	NNR (E_7)	NNN (E_8)

(i) For this, the events consists of points A= $\{E_5, E_6, E_7\}$, hence, $P(A) = \frac{3}{8}$.
(ii) The events here are B = $\{E_1, E_2, E_3, E_5\}$ and hence, $P(B) = \frac{4}{8} = \frac{1}{2}$.
(iii) Here the events consist of C = $\{E_7, E_8\}$, hence $P(C) = \frac{2}{8} = \frac{1}{4}$.

Table 4.2 Distribution of probabilities

	Presence of Rh+		
Gender	Yes (Rh+)	No (Rh−)	Total
Boys	0.4335	0.0765	0.51
Girls	0.4165	0.0735	0.49
Total	0.85	0.15	1.00

4.3 Marginal & Conditional Probabilities

In a certain city hospital 85 % of newly born babies are Rh+ (that is, they all have the Rh+ antigen on the surface of their red blood cells, otherwise, it is Rh-). It is also known that about 51 % of all babies born at this particular hospital are boys. Let the distribution of the probabilities be as given in Table 4.2.

Suppose an individual newly born baby is randomly selected from this hospital, then

(i) $P(\text{Boy}) = 0.51$ and $P(\text{Girl}) = 1 - 0.51 = 0.49$.
(ii) $P(\text{Rh+}) = 0.85$ and $P(\text{Rh−}) = 1 - 0.85 = 0.15$.
(iii) Probability that the child is a boy and is Rh+ = $P(\text{Boy and Rh+}) = 0.4335$

The probabilities in (i) and (ii) mentioned above are called *marginal* probabilities. We can also construct marginal probability tables with frequency data. Consider the following data which relate to a group of 1000 randomly selected adults who were asked if they are in favor of abortion or are against it. The results of this survey is presented in Table 4.3.

Table 4.3 Survey frequency distribution

	Response		
Gender	In favor	Against	Total
Male	248	203	451
Female	310	239	549
Total	558	442	1000

Thus,

$$P(\text{Male}) = \frac{451}{1000} = 0.451$$
$$P(\text{Female}) = \frac{549}{1000} = 0.549$$
$$P(\text{In favor}) = \frac{558}{1000} = 0.558$$
$$P(\text{Against}) = \frac{442}{1000} = 0.442$$

The above are the four marginal probabilities or *simple probabilities*. Notice that these probabilities involve single events.

4.4 Laws of Probability

4.4.1 The Addition Law of Probability

For any two events A and B, the probability that either A or B or both will occur, denoted by *P(A or B)* is defined as:

$$P(\text{A or B}) = P(A) + P(B) - P(A \text{ and } B) \tag{4.1}$$

The Venn diagram below displays this concept for two events A and B.

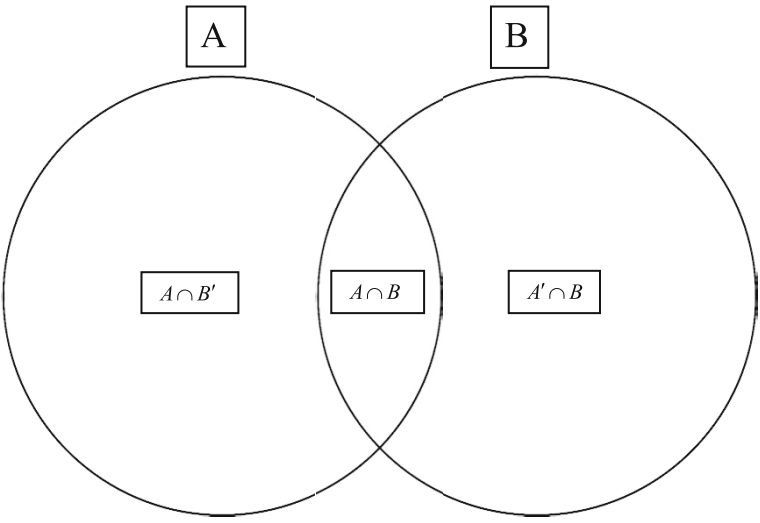


Fig. 4.1 Venn diagram to illustrate $A \cap B$

However, if A and B are mutually exclusive events, then the definition in (4.1) becomes,

$$P(A \text{ or } B) = P(A) + P(B) \tag{4.2}$$

Since in this case, $P(A \text{ and } B) = 0$.

Example 4.3.1

If 30 % of Nigerians are obese (A) and that 4 % of Nigerians suffer from diabetes (B). 2 % are both obese and suffer from diabetes. What is the probability that a randomly selected person is obese or suffers from diabetes?

Here, $P(A) = 0.3$, $P(B) = 0.04$ and $P(A \text{ and } B) = 0.02$. Then,

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\ &= 0.3 + 0.04 - 0.02 = 0.32 \end{aligned}$$

Example 4.3.2

Refer to Table 4.3. What is the probability that the individual selected is male or against abortion. Let $A = \{\text{Male}\}$ and B the event $B = \{\text{against}\}$. Here, $P(A) = \frac{451}{1000}$, $P(B) = \frac{442}{1000}$ and $P(A \textit{ and } B) = \frac{203}{1000}$, hence,

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\ &= \frac{451}{1000} + \frac{442}{1000} - \frac{203}{1000} \\ &= \frac{690}{1000} = 0.690 \end{aligned}$$

4.4.2 Multiplication Law

For any two events A and B, the **conditional probability** of A given B is defined as:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \quad \text{provided } P(B) \neq 0 \tag{4.3}$$

Of course it also follows that

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} \quad \text{provided } P(A) \neq 0 \tag{4.4}$$

Example 4.3.3

It is estimated that 15 % of the adult population has hyper-tension, but that 75 % of all adults feel that personally, they do not have this problem. It is also estimated that 6 % of the population has hyper-tension but does not think that the disease is present. If an adult patient reports that he or she does not have hyper-tension, what is the probability that the disease is, in fact, present?

If we let A denote the event that the patient does not feel that the disease is present and B, the event that the disease is present, we are given that $P(A) = 0.75$, $P(B) = 0.15$ and $P(A \text{ and } B) = 0.06$. We want to find $P(B \mid A)$. From the definition in (4.4), we have

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{0.06}{0.75} = 0.08$$

Thus, there is an 8 % chance that a patient who expresses an opinion that she or he has no problem with hypertension does, in fact, have the disease.

Example 4.3.4

Refer to the previous example (Example 4.3.3). If the disease is present, what is the probability that the patient will suspect its presence?

Here, we wish to find $P(\bar{A} \mid B)$, where $\bar{A}$ denotes the *compliment* of event A, that is, the event that A does not occur. Hence $P(\bar{A}) = 1 - P(A) = 1 - 0.75 = 0.25$. Thus,

$$\begin{aligned} P(\bar{A} \mid B) &= \frac{P(\bar{A} \text{ and } B)}{P(B)} = \frac{P(\bar{A}) \times P(B \mid \bar{A})}{P(B)} \\ &= \frac{0.25 \times 0.09/0.25}{0.15} = \frac{0.09}{0.15} \\ &= 0.60 \end{aligned}$$

That is, if the patient expresses the opinion that he or she has hypertension, there is a 60 % chance of the patient being right.

We have used the multiplication rule in the last example. This rule states for two events A and B that,

$$P(A \text{ and } B) = P(B) \times P(A \mid B) \quad \text{or} \tag{4.5a}$$

$$P(A \text{ and } B) = P(A) \times P(B \mid A) \tag{4.5b}$$

Refer to Table 4.2, what is the probability that the baby selected is a boy who is Rh+? that is,

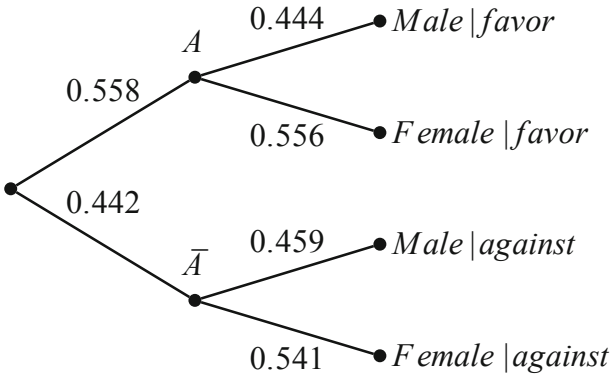
$$3P(Boy|Rh+) = \frac{P(Boy \text{ and } Rh+)}{P(Rh+)} = \frac{0.04335}{0.85} = 0.51$$

Example

Refer to Table 4.3, the probability of a male who is in favor is

$$\frac{P(\text{Male and favor})}{P(\text{favor})} = \frac{248/1000}{558/1000} = \frac{248}{558} = 0.444$$

We can represent this in a tree diagram as follows: First let A represents the event 'in favor'. Then $P(A) = 0.558$ and $P(\bar{A}) = 1 - 0.558 = 0.442$. Here, $P(\bar{A})$ is often described as the compliment of event A which is $1 - P(A)$. The tree diagram below represents these conditional probabilities. For instance, the probability that the person selected is a female who is opposed to abortion is 0.541.



4.5 Relationship Between Probability and Odds

If the odds in favor of an event A are a to b, then the probability that the event occurs is $p = \dfrac{a}{a+b}$ and the probability that event does not occur is $1 - \dfrac{a}{a+b} = \dfrac{b}{a+b}$. As an illustration, the odds in favor of obtaining a sum of seven in throwing two dice is 1:5. Thus, probability of observing this event is $1/(1+5) = 1/6$ and the odds against this event would therefore be $b : a = 5 : 1$ with corresponding probability $\dfrac{b}{a+b} = \dfrac{5}{6}$. The odds for success of an event therefore are the probability of success to the probability of failure.

4.6 Specificity, Sensitivity of Tests

Measures for testing the effectiveness of a test procedure (screening test or set of symptoms), such as a medical test to diagnose a disease for *sensitivity*, *specificity*, and *predictive values*.

In considering screening tests, we must be aware of the fact that they are not always infallible. That is, a testing procedure may yield a *false positive* or a *false negative*.

1. A *false positive* results when a test indicates a positive status when the true status is negative. That is, the test indicates that the disease is present, but the person does not really have the disease.
2. A *false negative* results when a test indicates a negative status when the true status is positive. That is, the person has the disease but the test does not detect it, because the person tested negative.

Suppose we have for a sample of n subjects (n always large), the information in the table below:

	Disease		
Test result	Present (D)	Absent ($\bar{D}$)	Total
+ve	a	b	a+b
−ve	c	d	c+d
Total	a+c	b+d	n

- The sensitivity of a test (or symptom) is the probability of a positive test result (or presence of the symptom) given the presence of the disease. That is, it is

$$P(+ve \mid D) = \frac{a}{a+c}$$

- The specificity of a test (or symptom) is the probability of a negative test result (or absence of the symptom) given the absence of the disease. That is,

$$P(-\text{ve} \mid \bar{D}) = \frac{d}{b+d}$$

- The predictive value positive of a screening test or symptom is the probability that the subject has the disease, given that the subject has a positive screening test result (or has the symptom). Or simply defined as the proportion of positive results that are true positives (i.e., have the symptom or disease)

$$P(D \mid +\text{ve}) = \frac{P(D \text{ and } +\text{ve})}{P(+\text{ve})} \tag{4.6}$$

$$= \frac{P(D \text{ and } +\text{ve})}{P(D \text{ and } +\text{ve}) + P(\bar{D} \text{ and } +\text{ve})} \tag{4.7}$$

$$= \frac{P(D) \times P(+\text{ve} \mid D)}{P(D) \times P(+\text{ve} \mid D) + P(\bar{D}) \times P(+\text{ve} \mid \bar{D})} \tag{4.8}$$

$$= \frac{P(D) \times \text{sensitivity}}{P(D) \times \text{sensitivity} + P(\bar{D}) \times (1.0 - \text{specificity})} \tag{4.9}$$

- The predictive value negative of a screening test (or symptom) is the probability that subject does not have the disease, given that the subject has a negative screening test result (or does not have the symptom) or again simply defined as the proportion of negative results that are true negatives (i.e., do not have the symptom or disease).

$$P(\bar{D} \mid -\text{ve}) = \frac{P(\bar{D} \text{ and } -\text{ve})}{P(-\text{ve})} \tag{4.10}$$

$$= \frac{P(\bar{D}) \times \text{specificity}}{P(\bar{D}) \times (1.0 - \text{sensitivity}) + P(\bar{D}) \times \text{specificity}} \tag{4.11}$$

Example 4.6.1

If a woman takes an early pregnancy test, she will either test positive, meaning that the test says she is pregnant, or test negative, meaning that the test says she is not pregnant. Suppose that if a woman is really pregnant, there is 98 % chance that she will test positive. Also, suppose that if a woman is *not* pregnant, there is a 99 % chance that she will test negative.

(1a) Suppose that 1000 women take early pregnancy tests and that 100 of them are really pregnant. What is the probability that a randomly chosen woman from this group will test positive?
(1b) Suppose that a woman tests positive, what is the probability that she is really pregnant?
(2a) Suppose that 1000 women take early pregnancy tests and 50 of them are really pregnant. What is the probability that a randomly chosen woman from this group will test positive?
(2b) Suppose that a woman test positive, what is the probability that she is really pregnant?

	Disease	
Test result	Pregnant (D)	Not pregnant ($\bar{D}$)
Positive +ve	0.98	0.01
Negative (−ve)	0.02	0.99
Total	1.00	1.00

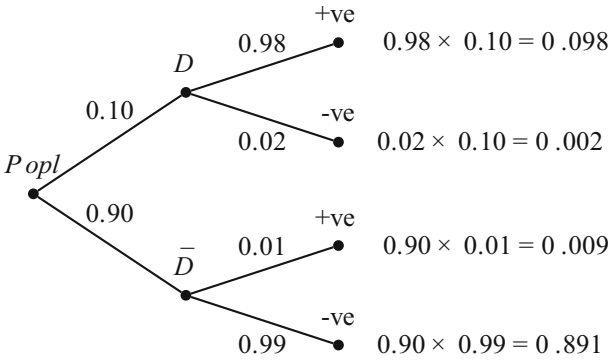
a The sensitivity of the test is given by:

$$P(+\text{ve} \mid \text{pregnant}) = 0.98$$

b The specificity of the test is given by:

$$P(\text{-ve} \mid \text{not pregnant}) = 0.99$$

In problem (1a), in a population of 1000 women, 100 of them were actually pregnant. Hence $P(D) = 0.10$ and $P(\bar{D}) = 0.90$ in the population. We can now use a probability tree diagram to solve our problem.



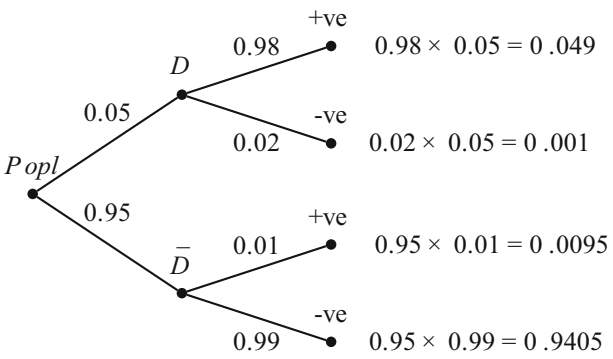
The probabilities are computed as follows using the multiplication rule:

$$P(D \cap +ve) = P(D) \times P(+ve|D) = 0.10 \times 0.98 = 0.098$$
$$P(\bar{D} \cap +ve) = P(\bar{D}) \times P(+ve|\bar{D}) = 0.90 \times 0.01 = 0.009$$

Similar calculations lead to those for the $(-ve)$s. Hence, the probability that a randomly chosen woman from this group will test positive equals $0.098 + 0.009 = 0.107$. In problem (1b), the probability that she is really pregnant given that she tested positive is the predictive value positive of the test and is computed as:

$$\frac{P(D \cap +ve)}{P(+\text{ve})} = \frac{0.098}{0.107} = 0.9159$$

For (2a), we now have $P(D) = 0.05$ and $P(\bar{D}) = 0.95$. The following probability tree diagram displays the various probabilities relating to this problem.



Hence, the probability that a randomly chosen woman from this group will test positive equals $0.049 + 0.0095 = 0.0585$. In problem (2b), the probability that she is really pregnant given that she tested positive is the predictive value positive of the test and is computed as:

$$\frac{P(D \cap +ve)}{P(+\text{ve})} = \frac{0.049}{0.0585} = 0.8376$$

Example 4.3.3

The blood type distribution in a certain country at the time of war was thought to be type A, 41 %; type B, 9 %; type AB, 4 %; and type O, 46 %. It is estimated that during this war, 4 % if inductees with type O blood were typed as having type A, 88 % of those with type A blood were correctly typed, 4 % with type B blood were typed as A, and 10 % with type B were typed as A. A soldier was wounded and brought to surgery. He was typed as having type A blood. What is the probability that this was his true blood type?

Here we wish to find $P(A_1 \mid B)$, where, let,

A_1 :He has type A blood

A_2 :He has type B blood

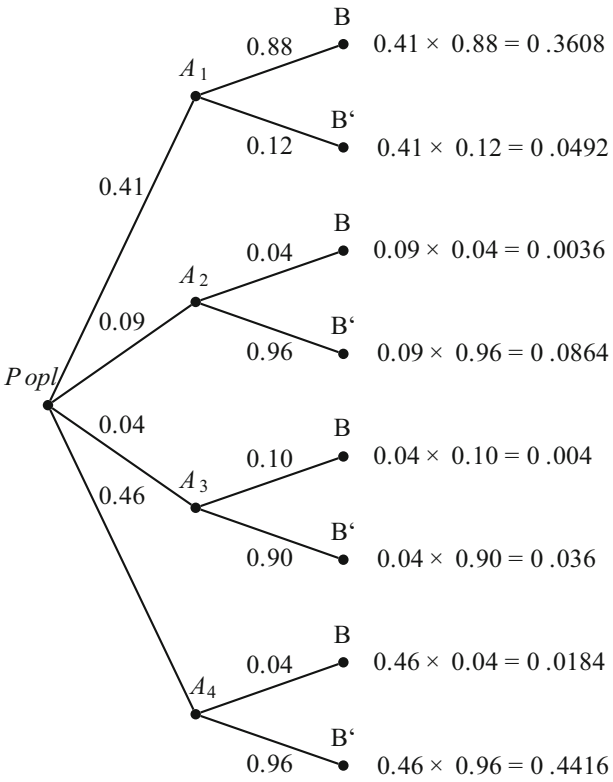
A_3 :He has type AB blood

A_4 :He has type O blood

Similarly we are given,

$$\begin{aligned} P(A_1) &= 0.41 \quad & P(B \mid A_1) &= 0.88 \\ P(A_2) &= 0.09 \quad & P(B \mid A_2) &= 0.04 \\ P(A_3) &= 0.04 \quad & P(B \mid A_3) &= 0.10 \\ P(A_4) &= 0.46 \quad & P(B \mid A_4) &= 0.04 \end{aligned}$$

The following tree diagram together with the computed probabilities are presented.



From the above,

$$P(B) = P(A_1 \text{ and } B) + P(A_2 \text{ and } B) + P(A_3 \text{ and } B) + P(A_4 \text{ and } B)$$
$$= 0.3608 + 0.0036 + 0.0040 + 0.0184 = 0.3868$$

But $P(A_1 \text{ and } B) = 0.3608$. Hence,

$$P[A_1 \mid B] = \frac{0.3608}{0.3868} = 0.9328$$

Thus there is a 93 % chance that the blood type is A if it has been typed as A. There is a 7 % chance that it has been mistyped as A when it is actually some other type. We have used in the example above the use of what is known as *Bayes' theorem.*

4.7 Receiver Operating Characteristics(ROC) Curves

The *receiver operating characteristic (ROC) curve* is a graphical plot of the sensitivity (true positives rate) versus the false positive rate (1−specificity) of the screening test, at different cut-off points used to designate test positive.

4.7.1 Example

Consider the following data for serum ferritin as a test for iron deficiency anemia. The level of serum ferritin (SF) found in blood and measured in milligrams percent is to be used as a diagnostic tool for detecting iron deficiency anemia. Large values of SF is often associated with iron deficiency anemia (Table 4.4).

Table 4.4 Serum ferritin as IDA diagnostic test

Serum ferritin (mmol/l)	# With IDA (% of total)	# Without IDA (% of total)
< 15	474	20
15–34	175	79
35–64	82	171
65–94	30	188
> 94	48	1332

Suppose we adopt cut points ≤ 15, ≤ 34, ≤ 64, and ≤ 94, then the corresponding 2×2 contingency tables for each of the cut points are presented in (i) to (iv) below.

	IDA		
SF	1	0	Total
≤ 15	474	20	494
> 15	335	1750	2085
Total	809	1770	2579

(i)

	IDA		
SF	1	0	Total
≤ 34	649	99	748
> 34	160	1671	1831
Total	809	1770	2579

(ii)

	IDA		
SF	1	0	Total
≤ 64	731	270	1001
> 64	78	1500	1578
Total	809	1770	2579

(iii)

	IDA		
SF	1	0	Total
≤ 94	761	438	1199
> 94	48	1332	1380
Total	809	1770	2579

(iv)

For Table (i), the sensitivity and specificity are computed as:

$$\text{Sensitivity} = \frac{474}{809} = 0.5859$$

$$\text{Specificity} = \frac{1750}{1770} = 0.9887$$

For Table (ii), the computed values are:

$$\text{Sensitivity} = \frac{649}{809} = 0.8022$$

$$\text{Specificity} = \frac{1671}{1770} = 0.9441$$

For Table (iii), these are similarly computed as:

$$\text{Sensitivity} = \frac{731}{809} = 0.9036$$
$$\text{Specificity} = \frac{1500}{1770} = 0.8475$$

and finally for Table (iv), we have:

$$\text{Sensitivity} = \frac{761}{809} = 0.9407$$
$$\text{Specificity} = \frac{1332}{1770} = 0.7525$$

These results are tabulated as follows (Table 4.5):

Table 4.5 Relationship between sensitivity and specificity

SF (cut-points)	Sensitivity	Specificity	1−Specificity
≤ 15	0.5859	0.9887	0.1130
≤ 34	0.8022	0.9441	0.0559
≤ 64	0.9036	0.8475	0.1535
≤ 94	0.9407	0.7525	0.2475
> 94	1.0000	0.0000	1.0000

The above table indicates the relationship between specificity and sensitivity. It clearly shows that as sensitivity increases, the specificity drops and vice versa. Thus there is a trade off and ideally, we would want a test that is highly sensitive and highly specific. A cut point of ≤ 15 for instance to diagnose iron deficiency anemia has sensitivity of 0.5859 and a specificity of 0.9887. We can increase the cutoff point for instance to increase the sensitivity. For instance, if we were to use a cutoff point of ≤ 94, we would have a higher sensitivity indicating that a larger number proportions of the diagnosis will be positive, but we would have decreased the specificity to 0.7525 and thus increasing the probability of false positives. The choice of a cutoff point should be selected carefully as we would not wish to minimize the false negative error (that patients does not have iron deficiency anemia when he/she clearly has it) in this particular case for instance. The relationship between sensitivity and specificity is often graphically illustrated by employing the Receiver operating characteristic (ROC) curve. The ROC curve is a graphical plot of the sensitivity values against 1−specificity curves and compare sensitivity versus specificity across a range of values for the ability to predict a dichotomous outcome. Area under the ROC curve is another measure of test performance.

The area under the ROC curve is a measure of the accuracy of the test. An area of 1 represents a perfect test or complete agreement, while an area of

0.5 is considered worthless. In general, an ROC area of between 0.5 and 0.70 is considered marginally useful. An area of between 0.7 and 0.9 is considered a good test, while an area greater than 0.90 is considered as an excellent test. Here, in our example, the area under the ROC curve is 0.9344 which is therefore considered as an excellent test.

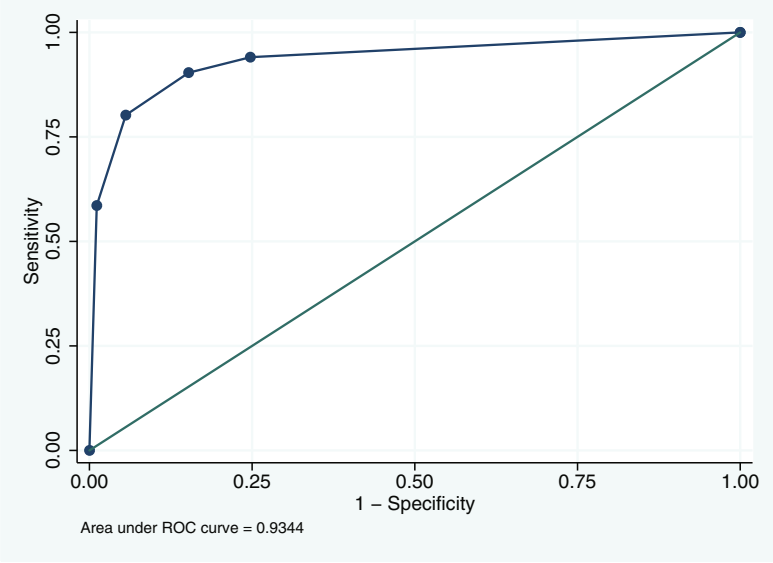


Fig. 4.2 ROC curve for Serum Ferritin

4.8 Probability Distributions

Suppose we denote the number of responses by X in Example 4.1 above. Then, X can take the following values, 0 (corresponding to *no responses*, i.e. NNN), 1, 2 or 3. We present corresponding outcomes for these values of X below.

X	Outcome
0	{NNN}
1	{RNN, NRN, NNR}
2	{RRN, RNR, NRR}
3	{RRR}

We notice that the value of X is not fixed as it could take any of the values 0, 1, 2 or 3. Therefore in statistical terms, X will be described as a discrete random variable. It is discrete because X takes discrete values $0, 1, \cdots$. The

probability density function (pdf) of X, denoted by $p(x)$ can therefore be obtained as follows:

	Response				
X	0	1	2	3	Total
$p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

$p(x)$ as given by the values above is described as the probability density function of the random variable X (No. of responses). Thus *p(x)* satisfies the following:

(i) $p(x) \geq 0$, for all i
(ii) $\sum p(x) = 1$, summed over all values of X

4.8.1 Mean and Variance of X

The mean of X, denoted by μ_x or *E(X)* (expectation of X) is calculated from the expression as mentioned below.

$$\begin{aligned} E(X) &= \sum xp(x), \quad \text{hence,} \\ &= 0\left(\frac{1}{8}\right) + 1\left(\frac{3}{8}\right) + 2\left(\frac{3}{8}\right) + 3\left(\frac{1}{8}\right) \\ &= 0 + \frac{3}{8} + \frac{6}{8} + \frac{3}{8} \\ &= \frac{3}{2} \end{aligned}$$

Similarly, the variance is obtained by using the formula

$$\sigma_x^2 = \sum x^2\ p(x) - \mu_x^2$$

i.e.,

$$\begin{aligned} \text{Var } X &= \sum x^2\ p(x) - E(X)^2 \\ &= 0^2\left(\frac{1}{8}\right) + 1^2\left(\frac{3}{8}\right) + 2^2\left(\frac{3}{8}\right) + 3^2\left(\frac{1}{8}\right) - \left(\frac{3}{2}\right)^2 \\ &= 0 + \frac{3}{8} + \frac{12}{8} + \frac{9}{8} - \frac{9}{4} \\ &= \frac{3}{4} \end{aligned}$$

Hence, the standard deviation $\sigma_x = \sqrt{\frac{3}{4}} = 0.86660$. *E(X)* is read as "Expected value of X" or simply as the "Expectation of X". The expectation gives the mean of the distribution.

We notice that for a discrete probability density function $p(x)$, then $p(x)$ must satisfy

$$\sum p(x) = 1 \tag{4.12}$$

Example 4.8.1

A cell when it multiplies can give birth to a maximum of four daughter cells. The probability of x daughter cells being formed by a cell which has just multiplied is given by the following probability distribution

x	1	2	3	4
$p(x)$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{1}{4}$

It can be seen from the example that X has discrete pdf and

$$\sum_{x=1}^{4} p(x_i) = \frac{1}{4} + \frac{3}{8} + \frac{1}{8} + \frac{1}{4} = 1$$

For the above pdf, we can evaluate the following probabilities

(i) $P(X = 3) = P(3) = \frac{1}{8}$
(ii) $P(X < 2) = P(1) = \frac{1}{4}$
(iii) $P(X > 2) = P(3) + P(4) = \frac{1}{8} + \frac{1}{4} = \frac{3}{8}$
(iv) $P(X \leq 2) = P(1) + P(2) = \frac{1}{4} + \frac{3}{8} = \frac{5}{8}$
(v) $P(2 \leq X \leq 4) = P(2) + P(3) + P(4) = \frac{3}{8} + \frac{1}{8} + \frac{1}{4} = \frac{3}{4}$
(vi) $P(X \geq 3) = P(3) + (p(4) = \frac{1}{8} + \frac{1}{4} = \frac{3}{8}$

4.8.2 Cumulative Probability Distribution Function

The cumulative probability distribution (cdf) for a discrete random variable X is defined as:

$$\mathrm{F}(x) = P(X \leq x)$$

and for the pdf in Example 4.8.1, the corresponding cdf is given by:

x	1	2	3	4
$F\ (x)$	$\frac{1}{4}$	$\frac{5}{8}$	$\frac{6}{8}$	1

Given, the cumulative distribution function $F(x)$, then the probabilities in (i)–(vi) in the previous subsection, are calculated as follows:

(i) $P(X = 3) = F(3) - F(2) = \frac{6}{8} - \frac{5}{8} = \frac{1}{8}$

(ii) $P(X < 2) = P(X \leq 1) = F(1) = \frac{1}{4}$

(iii) $P(X > 2) = 1 - P(X \leq 2) = 1 - F(2) = 1 - \frac{5}{8} = \frac{3}{8}$

(iv) $P(X \leq 2) = F(2) = \frac{5}{8}$

(v) $P(2 \leq X \leq 4) = F(4) - F(1) = 1 - \frac{1}{4} = \frac{3}{4}$

(vi) $P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2) = 1 - F(2) = 1 - \frac{5}{8} = \frac{3}{8}$

The mean and variance of the X in this example are also obtained from the pdf as:

$$E(X) = \sum xp(x) = 1\left(\frac{1}{4}\right) + 2\left(\frac{3}{8}\right) + 3\left(\frac{1}{8}\right) + 4\left(\frac{1}{4}\right) = \frac{19}{8}$$

$$\text{Var}(X) = \sum x^2 p(x) - E(X)^2 = 1^2\left(\frac{1}{4}\right) + 2^2\left(\frac{3}{8}\right) + 3^2\left(\frac{1}{8}\right) + 4^2\left(\frac{1}{4}\right)$$

$$= \frac{1}{4} + \frac{12}{8} + \frac{9}{8} + \frac{16}{4} - \left(\frac{19}{8}\right)^2 = 6.875 - 5.6406 = 1.234$$

4.9 The Binomial Distribution

Suppose, a drug company announces that it has just developed a new drug to cure a certain fictitious disease NNYZ. The company also claims that the cure rate of this new drug is 0.8 or 80 %. Suppose, there are four patients at the local hospital with this particular disease and we are interested in testing the efficacy of this new drug. Consequently, this new drug is administered to these four patients over a specified period of time and the conditions of the patients are re-examined at this time. What is the distribution of the number of patients cured by this new drug.

Let S denote the event that the drug cures and F the event that the drug fails to cure the patient. Then

$$P(S) = 0.8, \qquad \text{and} \qquad P(F) = 1 - 0.8 = 0.20.$$

Further let X be the number of patients cured. Then the possible values of X are {0, 1, 2, 3, 4} corresponding to:

$$\underbrace{FFFF}_{X=0}, \quad \underbrace{SFFF}_{X=1}, \quad \underbrace{SSFF}_{X=2}, \quad \underbrace{SSSF}_{X=3}, \quad \underbrace{SSSS}_{X=4}$$

But since each patient is either cured or not cured, thus we have $2\times2\times2\times2 = 16$ possible outcomes. We have, however, listed only five of these possible

outcomes, hence there is a need for us to locate where the other possible outcomes are lurking (or hiding). For example, the outcome SFFF listed above only indicates that the first patient was cured. It could have been only the second, the third or the fourth. Hence there would be four possible outcomes for the case when $X = 1$, namely {SFFF, FSFF, FFSF, FFFS}. To generate these outcomes we make use our earlier rule. Thus,

$$\text{FFFF} = \frac{4!}{0!4!} = 1$$
$$\text{SFFF} = \frac{4!}{1!3!} = 4$$
$$\text{SSFF} = \frac{4!}{2!2!} = 6$$
$$\text{SSSF} = \frac{4!}{!3!1!} = 4$$
$$\text{SSSS} = \frac{4!}{4!0!} = 1$$

for instance for the SSFF, there are two of one kind (SS) and two of the other kind (FF), hence, there are six possible outcomes corresponding to $X = 2$. In all, the outcomes are represented as:

These are,

FFFF
FFFS FSFF FFSF SFFF
FFSS SSFF SFSF FSFS SFFS FSSF
FSSS SSSF SSFS SFSS
SSSS

The corresponding probabilities are therefore:

$$P(X = 0) = 0.2^4 = 0.0016$$
$$P(X = 1) = 4(0.2^3\ 0.8^1) = 0.0256$$
$$P(X = 2) = 6(0.2^2\ 0.8^2) = 0.1536$$
$$P(X = 3) = 4(0.2^1\ 0.8^3) = 0.4096$$
$$P(X = 4) = (0.8^4) = 0.4096$$

Hence, the distribution of X, the number of patients cured can be summarized in Table 4.6.

Table 4.6 Distribution of the random variable X in this example

X	0	1	2	3	4	Total
$P(x)$	0.0016	0.0256	0.1536	0.4096	0.4096	1

We notice that $\sum P(x) = 1$. $P(x)$ is a pdf and a special pdf for that matter. The number of patients cured X is said to follow a *binomial* distribution with parameters $n = 4$ and $p = 0.8$ and is written as $X \sim \text{b}(4, 0.8)$.

The binomial distribution arises mainly when there are only two possible outcomes in each trial of an experiment, such that the two possible outcomes are mutually exclusive. These outcomes may be success or failure, germination or no-germination, defective or non-defective, yes or no etc. In the example above, a patient is either cured (S) or not cured (F). The probability of success is always denoted by p and that of failure by $1 - p$ or q and a binomial experiment is one that possesses the following properties:

(a) The experiment consists on n repeated trials.
(b) Each trial results in an outcome that may be classified as a success or failure, that is, dichotomous (Greek) or binary (Latin) outcomes.
(c) The probability of success is p and failure q such that $p + q = 1$
(d) The repeated trials are independent.

Then, if X represents the number of successes in n such repeated trials of the experiment, then the possible values of X are $0, 1, 2, \cdots, n$ and the distribution of X is called a *binomial distribution* with parameters n and p and is written as $X \sim b(n, p)$. For the above example, the probabilities can therefore be computed as in the following Table where, $p = 0.8$ and $q = 1 - p = 0.20$.

Values of X	0	1	2	3	4
$p(x)$	q^4	pq^3	p^2q^2	p^3q	p^4
No of sequences	$1 = \binom{4}{0}$	$4 = \binom{4}{1}$	$6 = \binom{4}{2}$	$4 = \binom{4}{3}$	$1 = \binom{4}{4}$

Such that Pr(Cured) $= p$ and Pr(not cured) $= q$. Then, $p(x)$ is given by

$$p(x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \cdots, n.$$

$$\text{or} = \frac{n!}{x!(n-x)!} p^x q^{n-x}, \quad x = 0, 1, 2, \cdots n.$$

For the binomial distribution, it can be shown that

$$E(X) = np \quad \text{and} \quad \text{Var}(X) = npq.$$

That is, the mean of a binomial distribution is np, while its variance is npq.

Example 4.9.1

The probability of a bacterium being infected with a phage is 0.4. If four bacteria are examined under a microscope what is the probability of

(i) No bacteria being infected
(ii) three bacteria being infected
(iii) at least one bacterium being infected?

In this example, each bacterium represents a trial. Since a bacterium is either infected (success) or not infected (failure), we thus have a binomial distribution with $n = 4$ and $p = 0.4$ and if X denotes the number of bacterium infected, then,

$$p(x) = \binom{4}{x}(0.4)^x(0.6)^{4-x}, \quad x = 0, 1, 2, 3, 4$$

Hence,

(i) Prob (No bacteria are infected) $= p(0)$
$= \binom{4}{0}(0.4)^0(0.6)^4 = (0.6)^4 = 0.1296$
(ii) Prob (three bacteria being infected) $= p(3)$
$= \binom{4}{3}(0.4)^3(0.6)^1$
$= 4 \times (0.4)^3(0.6) = 0.1536$
(iii) Prob (at least one bacterium is infected) $= P(X \geq 1)$
$= p(1) + p(2) + p(3) + p(4) = 1 - p(0)$
$= 1 - (0.6)^4 = 0.8704$

Alternatively, I would prefer to use the cumulative distribution function to solve the above problems. We have used MINITAB to generate the cdf for an $X \sim b(4, 0.4)$. These are presented below.

```
MTB > CDF;
SUBC> BINOMIAL 4 0.4.

Cumulative Distribution Function

Binomial with n = 4 and p = 0.400000

          x         F(x)
          0       0.1296
          1       0.4752
          2       0.8208
          3       0.9744
          4       1.0000
```

(i) $P(X = 0) = F(0) = 0.1296$
(ii) $P(X = 3) = F(3) - F(2) = 0.9744 - 0.8208 = 0.1536$
(iii) $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X \leq 0) = 1 - F(0) = 1 - 0.1296 = 0.8704$

Example 4.9.2

The genetic features of two adult mice are such that the probability of an offspring being an albino is 0.2.

If the mice give birth to six offsprings, calculate the probability of

(i) no albino
(ii) one albino only
(iii) two or more albinos

Solution

Here, $p = 0.2$; $q = 1 - 0.2 = 0.8$ and $n = 6$. Hence, $p(x)$ has the form:

$$p(x) = \binom{6}{x}(0.2)^x(0.8)^{6-x}, \quad x = 0, 1, 2, 3, 4, 5, 6.$$

The corresponding cdf is presented below.

```
MTB > CDF;
SUBC> BINO 6 0.2.

Cumulative Distribution Function

Binomial with n = 6 and p = 0.200000

          x        F(x)
          0      0.2621
          1      0.6554
          2      0.9011
          3      0.9830
          4      0.9984
          5      0.9999
          6      1.0000
```

(i) $P(X = 0) = P(X \leq 0) = F(0) = 0.2621$
(ii) $P(X = 1) = F(1) - F(0) = 0.6554 - 0.2621 = 0.3933$
(iii) $P(X \geq 2) = 1 - P(X < 2) = 1 - P(X \leq 1) = 1 - F(1) = 1 - 0.6554 = 0.3446$

Of course we could solve the problems by using the pdf, but these require extensive calculations especially if n is large. We however present these results in what follows:

(i) Prob (No albinos) $= p(0)$
$= \binom{6}{0}(0.2)^0(0.8)^6 = (0.8)^6 = 0.2621$
(ii) Prob (one albino only) $= p(1)$
$= \binom{6}{1}(0.2)^1(0.8)^5$
$= 6(0.2)(0.8)^5 = 0.3933$
(iii) Prob (two or more albinos) $= P(X \geq 2)$
$= p(2) + p(3) + p(4) + p(5) + p(6)$
$= 1 - p(0) - p(1)$
$= 1 - 0.2621 - 0.3932 = 0.3447$

4.10 The Poisson Distribution

Data which come as counts rather than as continuous measurements are often very skewed. Poisson or related theoretical distributions can sometimes be used to describe this type of data. The Poisson distribution, is also widely used in ecology to describe the ways in which shrubs, trees, insects etc are spread over areas. Other examples giving rise to a Poisson distribution are, insect counts in field plots, noxious weed seeds in seed samples, number of egg clusters on a leaf, etc. The Poisson distribution is most often used to model discrete events in time or space and has sometimes been referred to as the *distribution of rare events*.

Thus, if X has a Poisson distribution, then the probability density function (pdf) of X will be given by:

$$p(x) = \frac{e^{-\mu}\,\mu^x}{x!}, \quad x = 0, 1, 2, \cdots, \tag{4.13}$$

where $x!$ stands for factorial.
Since $p(x)$ is a probability density function, it follows that

$$\sum_{x=0}^{\infty} p(x) = 1$$

i.e., the probabilities must sum up to 1.

Example 4.10.1

If insect egg clusters on the leaves of a tree have a Poisson distribution with parameter $\mu = 0.5$. Calculate the probability of a leaf having

(i) No egg clusters
(ii) At least one egg cluster.

Let X denote the number of egg clusters on a leaf. Hence,

$$p(x) = \frac{(0.5)^x e^{-0.5}}{x!}, \quad x = 0, 1, 2, 3, \cdots$$

(i) Prob (no eggs) $= p(0) = \frac{(0.5)^0 e^{-0.5}}{0!} = e^{-0.5} = 0.6066$
(ii) Prob (at least one egg cluster in leaf) is computed as:

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X \leq 0) = 1 - 0.6066 = 0.3934$$

Again if the cdf is available, we would be better off using this for the problem above. We present the MINITAB generation of the cdf of a Poisson with parameter $\mu = 0.5$

```
MTB > CDF;
SUBC> POISSON 0.5.
Poisson with mu = 0.500000

        x         F(x)
        0       0.6065
        1       0.9098
        2       0.9856
        3       0.9982
        4       0.9998
        5       1.0000
```

(i) $P(X = 0) = P(X \leq 0) = F(0) = 0.6065$
(ii) $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X \leq 0) = 1 - F(0) = 1 - 0.6065 = 0.3935$

The Poisson distribution has $E(X) = \mu$ and $Var\ (X) = \mu$. That is, for this distribution the mean and variance are equal to the parameter μ

4.10.1 Fitting of Poisson Distribution to a Sample of Data

The data below shows the number of noxious weed seeds in 98 samples of Phleum praetense (meadow grass). Fit a Poisson distribution to the data.

	No of Noxious weeds											
X	0	1	2	3	4	5	6	7	8	9	10	11+
Frequency	3	17	26	16	18	9	3	5	0	1	0	0

The first step is to estimate the mean μ from the data by computing:

$$\hat{\mu} = \frac{\sum x_i f_i}{\sum f_i} = \frac{(0 \times 3) + (1 \times 17) + (2 \times 26) + \cdots + (11 \times 0)}{3 + 17 + 26 + \cdots + 0}$$
$$= \frac{296}{98} = 3.0204$$

Hence,

$$p(x) = \frac{(3.0204)^x \, e^{-3.0204}}{x!}, \quad x = 0, 1, 2, \cdots, 11 \tag{4.14}$$

and

$$\text{var}(x) = \frac{\sum f_i (x_i - \hat{\mu})^2}{\sum f_i - 1} = 3.2779$$

As a first check, we notice that the variance is very close to the mean. These initial calculations indicate that perhaps the data can best be fitted by a Poisson distribution.

4.10.2 Use of Recursion Formula

We know that

$$p(x) = \frac{\mu^x e^{-\mu}}{x!}, \quad \text{and}$$
$$p(x+1) = \frac{\mu^{(x+1)} e^{-\mu}}{(x+1)!}$$

Hence,

$$\frac{p(x+1)}{p(x)} = \frac{\mu^{x+1}}{(x+1)!} \times \frac{x!}{\mu^x} = \frac{\mu}{x+1}$$

i.e.,

$$p(x+1) = \frac{\mu}{x+1} p(x)$$

Thus, if $x = 0$,

$$p(1) = \frac{\mu}{1} p(0) = \mu p(0)$$

$$p(2) = \frac{\mu}{2}p(1)$$

$$p(3) = \frac{\mu}{3}p(2), \quad \text{etc.}$$

For these data, $p(0) = \dfrac{e^{-3.0204}(3.0204)^0}{0!} = 0.0488$. Hence, the expected values are given by

$$E_i = np(x_i) \quad x = 0, 1, 2, \cdots, 11$$

Table 4.7 gives the result of these computations.

Table 4.7 Table of expected frequencies

No of noxious weeds (x)	Frequency	Poisson multipliers	Expected frequency
0	3	$1 = 1.0$	4.781
1	17	$\hat{\mu} = 3.0204$	14.440
2	26	$\frac{\hat{\mu}}{2} = 1.5102$	21.807
3	16	$\frac{\hat{\mu}}{3} = 1.0068$	21.955
4	18	$\frac{\hat{\mu}}{4} = 0.7551$	16.578
5	9	$\frac{\hat{\mu}}{5} = 0.6041$	10.015
6	3	$\frac{\hat{\mu}}{6} = 0.5034$	5.042
7	5	$\frac{\hat{\mu}}{7} = 0.4315$	2.176
8	0	$\frac{\hat{\mu}}{8} = 0.3756$	0.817
9	1	$\frac{\hat{\mu}}{9} = 0.3356$	0.274
10	0	$\frac{\hat{\mu}}{10} = 0.3020$	0.083
11+	0	$\frac{\hat{\mu}}{11} = 0.2746$	0.030
Total	98		97.998

We shall discuss whether this model fits the data, further in Chap. 8.

4.11 The Normal Distribution

The most important continuous distribution in the entire field of Statistics is the normal (Gaussian) or *bell-shaped* distribution. The graph of the normal curve is bell-shaped and is given in Fig. 4.3 below.

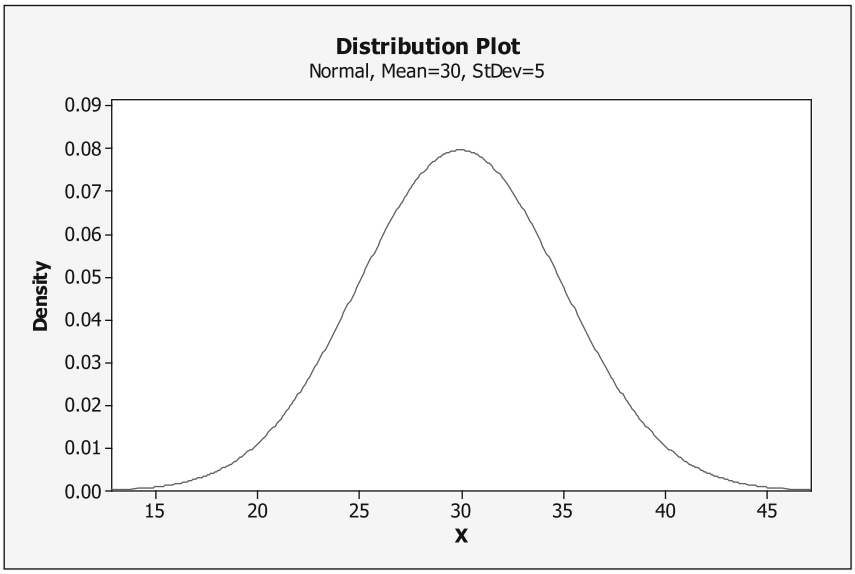


Fig. 4.3 Normal Probability Plot for $\mu = 30$ and $\sigma = 5$

The distribution provides a basis upon which much of the theory of inductive statistics is based.

The variable X is said to have a normal distribution with mean μ and variance σ^2. Its distribution function is of the form:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

and is usually written as $X \sim N(\mu, \sigma^2)$. Once μ and σ are specified, the normal distribution is completely satisfied.

Some Properties

(i) The mode which is the point on the horizontal axis occurs at $x = \mu$.
(ii) The curve is symmetrical about a vertical axis through the mean μ.
(iii) The total area under the curve and above the horizontal axis is equal to 1. Thus the area under the curve to the left of $x = \mu$ equals 1/2, while similarly the area under the curve to the right of $x = \mu$ is also 1/2.

4.11.1 Areas Under the Normal curve

If $X \sim N(\mu, \sigma^2)$, and suppose we are interested in the probability $P(X_1 < X < X_2)$, we can obtain this area by means of a transformation

$$z = \frac{X - \mu}{\sigma}$$

Here, z is called a standardized normal variate and $E(Z) = 0$, i.e., the mean of $z = 0$. Similarly, the variance of $z = 1$. A typical standard normal distribution curve is presented in Fig. 4.4

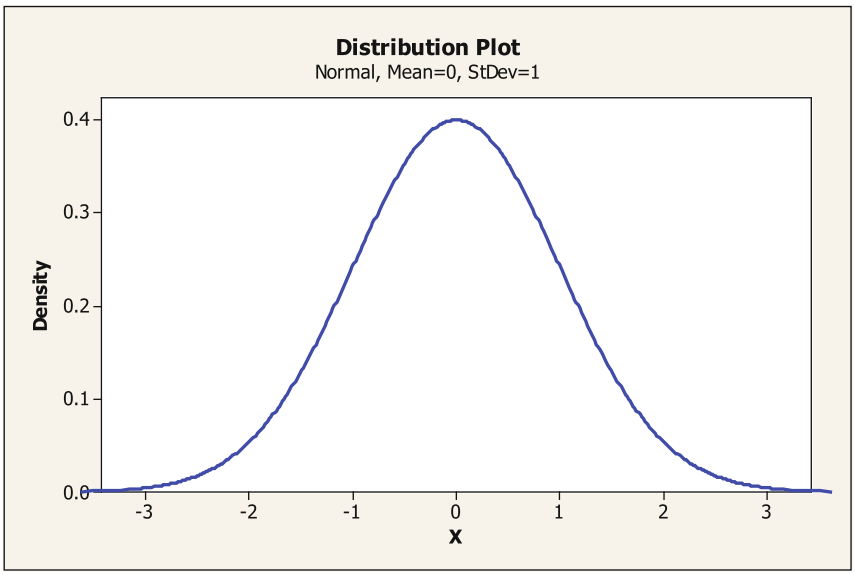


Fig. 4.4 A standard normal distribution

Thus,

$$P(X_1 < X < X_2) = P\left[\frac{x_1 - \mu}{\sigma} < \frac{x - \mu}{\sigma} < \frac{x_2 - \mu}{\sigma}\right]$$
$$= P\left[\frac{x_1 - \mu}{\sigma} < z < \frac{x_2 - \mu}{\sigma}\right]$$
$$= \Phi\left[\frac{x_1 - \mu}{\sigma}\right] - \Phi\left[\frac{x_2 - \mu}{\sigma}\right]$$

Where, because of symmetry, $\Phi(-x) = 1 - \Phi(x)$. For example, $\Phi(-0.88) = 1.0 - \Phi(0.88) = 1 - 0.8106 = 0.1894$. Tables of $\Phi(z)$ are given in Table 1 in the appendix. z is often referred to as the z-score.

Example 4.11.1

Given a $X \sim N(50, 100)$, find the probability that X assumes a value between 45 and 62.

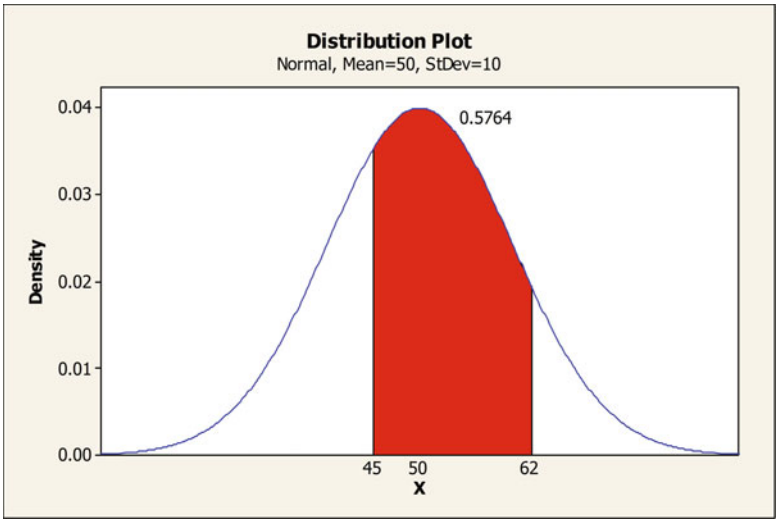


Fig. 4.5 Area required for this example 4.3.1

Here, we are interested in $P(45 < X < 62)$ and since $\mu = 50$, and $\sigma^2 = 100$, i.e. $\sigma = 10$. Hence,

$$\begin{aligned} P(45 < X < 62) &= \Phi\left[\frac{62-50}{10}\right] - \Phi\left[\frac{45-50}{10}\right] = \Phi(1.2) - \Phi(-0.5) \\ &= 0.8849 - \{1 - \Phi(0.5)\} = 0.8849 - 1 + 0.6915 \\ &= 0.5764 \end{aligned}$$

Example 4.11.2

If $X \sim N(\mu, \sigma^2)$, then,

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &= 0.683 \\ P(\mu - 2\sigma < X < \mu + 2\sigma) &= 0.954 \\ P(\mu - 3\sigma < X < \mu + 3\sigma) &= 0.9974 \end{aligned}$$

For example,

$$\begin{aligned} P(\mu - 3\sigma < X < \mu + 3\sigma) &= \Phi\left[\frac{\mu + 3\sigma - \mu}{\sigma}\right] - \Phi\left[\frac{\mu - 3\sigma - \mu}{\sigma}\right] \\ &= \Phi(3) - \Phi(-3) \\ &= \Phi(3) - \{1 - \Phi(3)\} \\ &= 2\Phi(3) - 1 \end{aligned}$$

$$= 2(0.9987) - 1$$
$$= 0.9974$$

Example 4.11.3

Find $Pr(z \leq 1.52)$.

The area required for this problem is plotted in Fig. 4.6 and is computed from the table in the appendix as $P(z \leq 1.52) = 0.9357$.

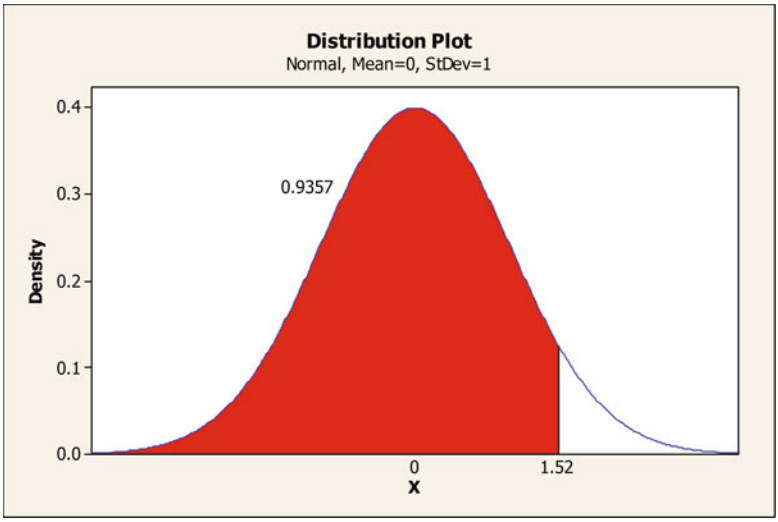


Fig. 4.6 Area required for this example 4.3.3

Example 4.11.4

Find $P(z > 1.52)$.
Similarly, the area required is plotted in Fig. 4.7 and is computed as, $P(z > 1.52) = 1 - P(z \leq 1.52) = 1 - 0.9357 = 0.0643$.

Example 4.11.5

A population of marine gastropods have shell lengths which are normally distributed with mean 7 mm and variance 2 mm^2. What proportion of the population will have a shell length between 5 and 9 mm?

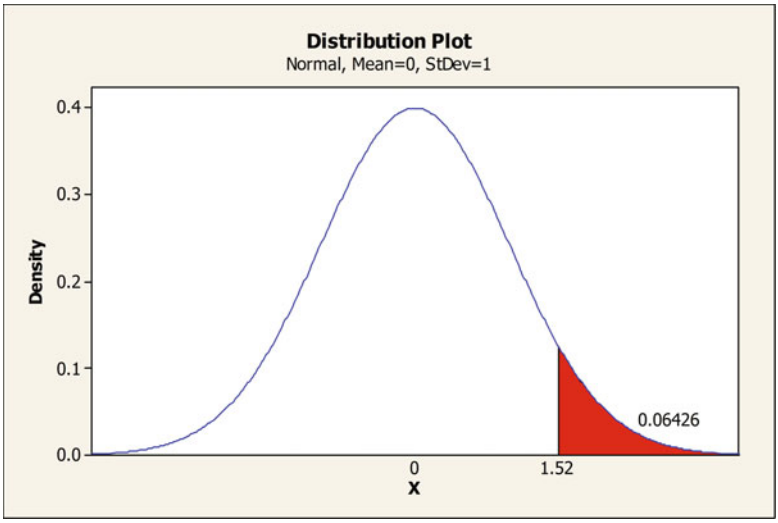


Fig. 4.7 Area required for this example 4.3.4

Solution

Let X be the shell length of a gastropod. Then $X \sim N(7, 2)$, i.e., $\mu = 7$, $\sigma^2 = 2$, which implies that $\sigma = \sqrt{2}$. The area required for this problem is plotted in Fig. 4.8

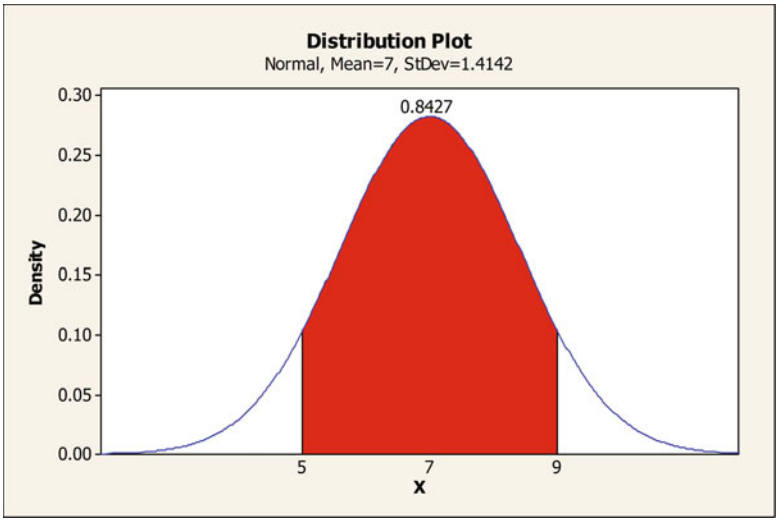


Fig. 4.8 Area required for this example 4.3.5

Hence,

$$\begin{aligned} P(5 \leq X \leq 9) &= \Phi\left[\frac{9-7}{\sqrt{2}}\right] - \Phi\left[\frac{5-7}{\sqrt{2}}\right] \\ &= \Phi\left(\frac{2}{\sqrt{2}}\right) - \Phi\left(\frac{-2}{\sqrt{2}}\right) \\ &= \Phi(1.41) - \{1 - \Phi(1.41)\} \\ &= 2\Phi(1.41) - 1 \\ &= 2(0.9207) - 1 \\ &= 0.8414 \end{aligned}$$

Hence, approximately 84 % of the population will have a shell length in the range 5 to 9 mm.

4.12 Normal Approximations to the Binomial

We recall that the binomial probability density function(pdf) is given by:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, \cdots, n \tag{4.15}$$

However, when n is very large, the expression in (4.15) can be computational burdensome, laborious and time consuming. We may in this case employ the normal approximation to the calculation of binomial probabilities. We recall that for a binomial random variable with parameters n and p, the mean and variance are respectively given by $\mu = np$ and $\sigma^2 = npq$. Thus, $\sigma = \sqrt{npq}$. The figure below, Fig. 4.9 gives a normal approximation to a binomial distribution with parameters 25 and 0.6. That is, b(25, 0.6). Here $\mu = np = 15$, and $\sigma^2 = npq = 6$ and thus $\sigma = \sqrt{6} = 2.4495$.

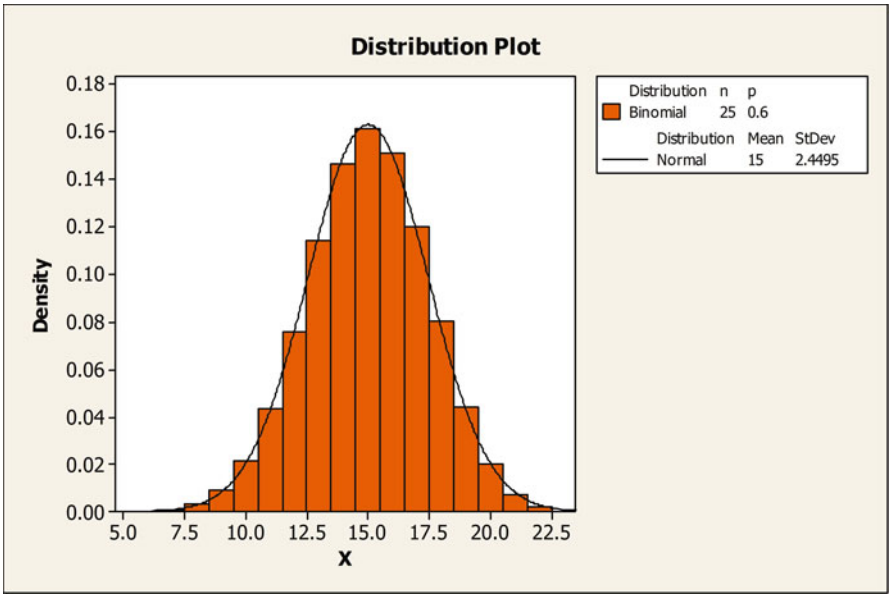


Fig. 4.9 Overlay of normal distribution on binomial

For the approximation to work,

$$\min[np, n(1-p)] \geq 5$$

Further, the continuity correction for the normal approximation must be applied as indicated in (4.16) to (4.16) for a binomial variate with parameters n and p.

$$\begin{aligned} P(x \leq a) &= P\left[z < \frac{(a+0.5) - np}{\sqrt{npq}}\right] \\ P(x \geq a) &= P\left[z > \frac{(a-0.5) - np}{\sqrt{npq}}\right] \\ P(a \leq x \leq b) &= P\left[\frac{(a-0.5) - np}{\sqrt{npq}} < z < \frac{(b+0.5) - np}{\sqrt{npq}}\right] \end{aligned} \tag{4.16}$$

Suppose we wish to find (i) $P(X \leq 12)$ and (ii) $P(10 \leq X \leq 16)$ for the b(25, 0.6) example above. Here the correct solutions are (i) Cumulative Distribution Function

Binomial with $n = 25$ and $p = 0.6$

```
 x  P(X<=x)
12    0.153768
```

and similarly, the second case has the solution (ii) $p = F(16) - F(9) = 0.7265 - 0.0132 = 0.7133$.

The normal approximation probabilities are computed as follows:

$$\begin{aligned} P(X \leq 12) &= P\left[z < \frac{(12 + 0.5) - 15}{2.4495}\right] \\ &= P(z < -1.02) \\ &= 0.1539 \end{aligned}$$

Similarly,

$$\begin{aligned} P(10 \leq X \leq 16) &= P\left[\frac{(10 - 0.5) - 15}{2.4495} < z < \frac{(16 + 0.5) - 15}{2.4495}\right] \\ &= P(-2.25 < z < 0.61) \\ &= 0.7291 - 0.0122 \\ &= 0.7169 \end{aligned}$$

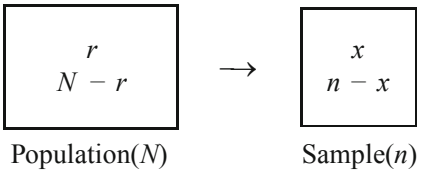
We see the normal approximations are very close to the exact values. The approximation will better be appreciated for say, the case when $x \sim b(150, 0.6)$, where we would need to generate 150! for instance in order to compute the necessary probabilities. In this kind of situation, the binomial approximation will be very useful. Of course such an exact computation will not create any problem in MINITAB.

4.13 The Hypergeometric Distribution

The hypergeometric distribution has the following characteristics:

- The finite population of N subjects, where each subject is classified as either S (success) or F (failure).
- A random sample of size n subjects is drawn without replacement from this population, such that r of which are S's (for Success) and $(N - r)$ are F's (for failure).
- The hypergeometric random variable x is the number of Successes in the sample of size n.

The boxes below demonstrate visually these realizations.

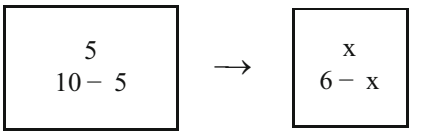


With the above set up, the hypergeometric distribution is therefore has the probability distribution:

$$\Pr(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \quad \text{for } x = 0, 1, 2, \cdots n \tag{4.17}$$

Example

If $N = 10,\ R = 5$ and $n = 6$, then $N - R = 5$. The extreme samples are SSSSSF and FFFFFS. That is, min $= 1$ and max $= 5$.



Hence,

$$\Pr(x) = \frac{\binom{5}{x}\binom{5}{6-x}}{\binom{10}{6}}, \quad \text{for } x = 0, 1, 2, \cdots 5$$

The MINITAB is employed to calculate these probabilities with the following results. Note that the cumulative distribution function $F(x) = P(X \leq x)$ is also displayed and is generated by stating in the MINITAB statement (1): MTB > CDF c1.

```
MTB > Set c1
DATA>   1( 0 : 5 / 1 )1
DATA>   End.
MTB > PDF C1;
SUBC>   Hypergeometric 10 5 6.

Probability Density Function

Hypergeometric with N = 10, M = 5, and n = 6

x  P(X=x)
0    0.000000
1    0.023810
2    0.238095
3    0.476190
4    0.238095
5    0.023810

MTB > CDF C1;
SUBC>   Hypergeometric 10 5 6.
```

```
Cumulative Distribution Function

Hypergeometric with N = 10, M = 5, and n = 6

x  P(X<=x)
0      0.00000
1      0.02381
2      0.26190
3      0.73810
4      0.97619
5      1.00000
```

4.14 Sampling Distributions

4.14.1 Introduction

In sampling theory, we are concerned with the problem of estimating population parameters from the sample statistics. Questions such as "How close to the required parameter can a statistic be expected to lie"? "How can we be sure that our estimate is not more than a specified quantity out"? Can be answered by studying the appropriate sampling distributions of $\bar{x}$ and s^2, which are defined as:

$$\bar{x} = \frac{\sum x}{n} \qquad \text{and} \quad s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Estimation theory, deals with the estimation of the parameters of the population from the sample. Thus $\bar{x}$ estimates the population mean, and s^2 the population variance etc.

Example 4.14.1

The number of heart beats per minute of a patient recorded on ten successive days was a follows:

$$73, 72, 73, 74, 76, 70, 71, 72, 72, 74$$

Here, $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{73 + 72 + \cdots + 74}{10} = 72.7$, and

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left\{\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right\} \\ &= \frac{1}{9}\{73^2 + 72^2 + \cdots + 74^2\} - \frac{(727)^2}{10} \\ &= 2.9 \end{aligned}$$

Hence, $s = \sqrt{2.9} = 1.7$.

If X is normally distributed with mean μ and variance σ^2, that is, $X \sim N(\mu, \sigma^2)$, if we take a random sample of size n from this population, then

(i) $\bar{x} = \dfrac{\sum x_i}{n} \sim N(\mu, \dfrac{\sigma^2}{n})$

(ii) Further, if X has any non-normal probability distribution with mean μ and variance σ^2, then the distribution of $\bar{x}$ approaches the normal distribution with mean μ and variance $\frac{\sigma^2}{n}$ as the sample size n increases. The conventional choice of n is that the above approximation will be true for $n \geq 30$.

The above are called the *Central Limit Theorem*, (CLT).

To illustrate (ii) above, suppose a population consists of five numbers 1, 2, 3, 4, and 5. Then, $\mu = \dfrac{15}{5} = 3$ and $\sigma^2 = \dfrac{\sum(x_i - \mu)^2}{n} = 2$.

Now suppose a random sample of size 2 is drawn from this population. There are 25 samples of size 2 which can be drawn with replacement from the population, they are:

(1,1)	(2,1)	(3,1)	(4,1)	(5,1)
(1,2)	(2,2)	(3,2)	(4,2)	(5,2)
(1,3)	(2,3)	(3,3)	(4,3)	(5,3)
(1,4)	(2,4)	(3,4)	(4,4)	(5,4)
(1,5)	(2,5)	(3,5)	(4,5)	(5,5)

The corresponding sample means are:

1.0	1.5	2.0	2.5	3.0
1.5	2.0	2.5	3.0	3.5
2.0	2.5	3.0	3.5	4.0
2.5	3.0	3.5	4.0	4.5
3.0	3.5	4.0	4.5	5.0

and these can be displayed as distribution as in Table 4.8.

Table 4.8 Sampling distribution of $\bar{x}$

$\bar{x}$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$p(\bar{x})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{1}{25}$

Hence,

$$\begin{aligned} E(\bar{X}) &= \sum \bar{x} p(\bar{x}), \\ &= 1.0\left(\frac{1}{25}\right) + 1.5\left(\frac{2}{25}\right) + \cdots + 5.0\left(\frac{1}{25}\right) \\ &= \frac{75}{25} = 3 \end{aligned}$$

This illustrates the fact that $E(\bar{x}) = \mu$. That is, the mean of all means of the sampling distribution, denoted as $\mu_{\bar{x}} = \dfrac{\sum \bar{x}_i}{25} = \dfrac{75}{25} = 3$.
Similarly, the variance is obtained by using the formula

$$\sigma_{\bar{x}}^2 = \sum \bar{x}^2 \; p(\bar{x}) - \mu_{\bar{x}}^2$$

i.e,,

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \sum \bar{x}^2 \; p(\bar{x}) - \mu_{\bar{x}}^2 \\ &= 1.0^2 \left(\frac{1}{25}\right) + 1.5^2 \left(\frac{2}{25}\right) + \cdots + 5.0^2 \left(\frac{1}{25}\right) - 3^2 \\ &= \frac{250}{25} - 3^2 = 1\end{aligned}$$

The above variance of $\bar{x}$ can also be obtained as:

$$\sigma_{\bar{x}}^2 = \frac{1}{n} \sum_{i=1} 25\bar{x}_i^2 - \mu_{\bar{x}}^2 = \frac{250}{25} - 3^2 = 1$$

But this is equal to $\dfrac{\sigma^2}{2} = \dfrac{2}{2} = 1$. Hence $\text{Var}(\bar{x}) = \dfrac{\sigma^2}{n}$ and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$ is called the standard error of mean. In this case, the standard error is 1.0. To further illustrate this, suppose we use MINITAB to generate 1000 random samples from a discrete distribution [1, 5]. Then, suppose we took random samples of sizes 5, 15, 25 and 30 from this population. We display below their dotplots in Fig. 4.10 as well as the computed sample means and sample variances and standard errors.

```
MTB > Random 1000 c1-c30;
SUBC>   Integer 1 5.
MTB > rmean c1-c5 c31
MTB > rmean c1-c15 c32
MTB > rmean c1-c25 c33
MTB > rmean c1-c30 c34
MTB > Dotplot 'n5' 'n15' 'n25' 'n30';
SUBC>   Overlay.

Dotplot of n5, n15, n25, n30

MTB > describe c31-c34

Descriptive Statistics: n5, n15, n25, n30

Variable     N  N*    Mean  SE Mean   StDev  Minimum      Q1  Median      Q3
n5        1000   0  2.9838   0.0207  0.6542   1.0000  2.6000  3.0000  3.4000
n15       1000   0  2.9959   0.0115  0.3646   1.8667  2.7333  3.0000  3.2667
n25       1000   0  2.9885  0.00883  0.2793   2.1200  2.8000  3.0000  3.2000
n30       1000   0  2.9921  0.00801  0.2532   2.3667  2.8333  3.0000  3.1667

Variable  Maximum
n5         4.6000
n15        4.1333
n25        3.8800
n30        3.8333
```

The dot plots describes the sampling distribution of $\bar{x}$ for sample sizes $n = 5, 15, 25, 30$. The distributions are approximately normal even for small sample sizes. For the population, theoretically, $\mu = 3.0$ and $\sigma^2 = 2$. Hence $\sigma = 1.4142$

In the above simulation, we see that $\mu_{\bar{x}}$ are all about 3.01, all 2.98, thus very close to the theoretical value of 3.0. Similarly, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, $n = 5, 15, 25, 30$ are estimated as $\frac{1.4142}{\sqrt{5}}, \frac{1.4142}{\sqrt{15}}, \frac{1.4142}{\sqrt{25}}, \frac{1.4142}{\sqrt{30}} = (0.6324,\ 0.3651,\ 0.2828,\ 0.2528)$. The column labelled **StDev** in the descriptive statistics above indicate the sample standard errors for the four sample sizes. Again, we see that the simulated values are very close to the theoretical values.

Example 4.14.2

A second example is given below.

Consider a population with the following probability distribution

X	0	1	2	3
$p(x)$	0.2	0.4	0.3	0.1

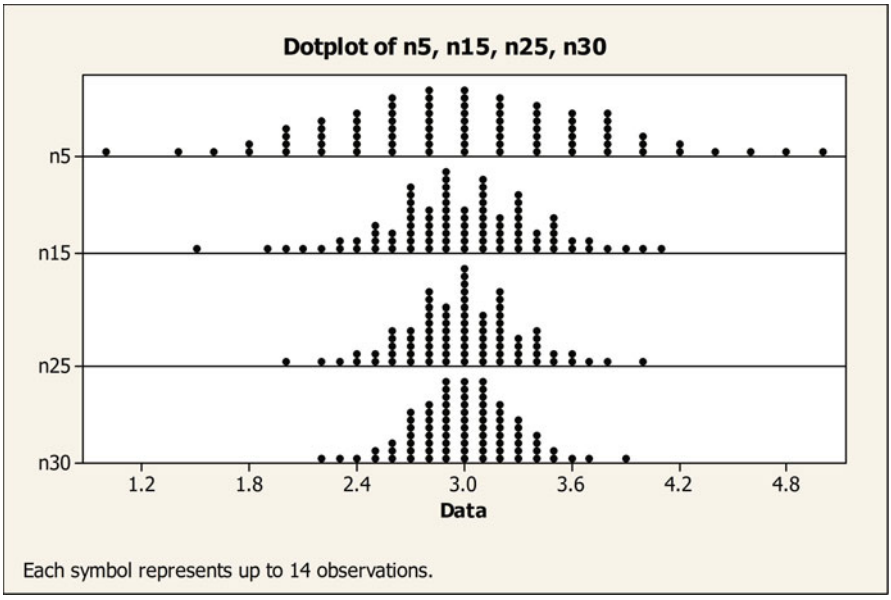


Fig. 4.10 Dot plots for the four sample means

Suppose we take a random sample of size 2 (X_1, X_2) from this population. We could have the following combinations:

$\bar{x} = \frac{x_1 + x_2}{2}$ with the following joint distributions of x_1 and x_2 as displayed in Table 4.9 below.

Table 4.9 Joint distribution of X_1 and X_2

	X_1				
X_2	0	1	2	3	Total
0	0.04	0.08	0.06	0.02	0.20
1	0.08	0.16	0.12	0.04	0.40
2	0.06	0.12	0.09	0.03	0.30
3	0.02	0.04	0.03	0.01	0.10
Total	0.20	0.40	0.30	0.10	1.0

Where for example:

$$P(X_1 = 0, X_2 = 2) = P(X_1 = 0) \times P(X_2 = 2) = 0.2 \times 0.3 = 0.06$$

The possible values of $\bar{x} = \frac{x_1+x_2}{2}$ are given below in Table 4.10, together with their corresponding probabilities.

Table 4.10 Sampling Distribution of $\bar{x}$

Values of $\bar{x}$	0	0.5	1.0	1.5	2.0	2.5	3.0	Total
$p(\bar{x})$	0.04	0.16	0.28	0.28	0.17	0.06	0.01	1

where for example:

For $\bar{x} = 1$, the possible values of $(X_1, X_2) = (1,1), (2,0), (0,2)$ and hence the probability is $0.16 + 0.06 + 0.06 = 0.28$. Similarly, for $\bar{x} = 1.5$, we have the possible values of (X_1, X_2) being (0,3), (1,2), (2,1) and (3,0) with corresponding probability $= 0.02 + 0.12 + 0.12 + 0.02 = 0.28$ etc.
For the parent population:

$$\sum x_i p(x) = 1.3 = \mu$$

$$\sum x_i^2 p(x) = 2.5$$

Hence, its variance equals $2.5 - 1.3^2 = 0.81 = \sigma^2$. That is, the mean and variance of the parent population are 1.3 and 0.81 respectively.

From Table 4.10, for the sample of size 2, we have:

$$\sum \bar{x}_i p(\bar{x}) = 1.3 = \bar{x}$$

$$\sum \bar{x}_i^2 p(\bar{x}) = 2.095$$

Hence, $\text{Var}(\bar{x}) = 2.095 - 1.3^2 = 0.405 = \frac{\sigma^2}{2}$. Thus we note that, generally, $\bar{x}$ is distributed with the same mean and variance $\frac{\sigma^2}{n}$ where n is the sample size.

The standard deviation of the sample variance $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ is called the *the standard error* for the mean.

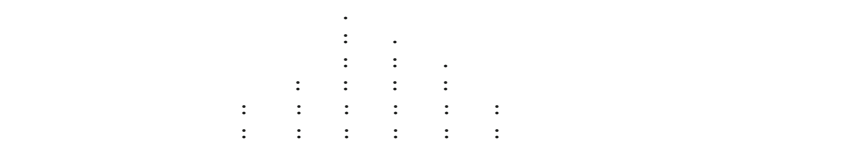
4.14.2 Simulation

We present below a MINITAB simulation program for selecting a random sample of size 2 put in columns C3 and C4. 1000 samples of size 2 were randomly generated from the discrete distribution in example 4.14.2. The distribution of $\bar{x}$ so generated are presented in a dotplot. It is almost symmetrical. The distribution of the 1000 means are presented in the tally and these agree closely with those presented in Table 4.10. From the simulation, $\mu_{\bar{x}} = 1.304$ and $\sigma^2_{\bar{x}} = 0.3725$. These agree very closely with the theoretical values of 1.30 and 0.405 respectively.

```
 Row     X   p(x)

   1     0    0.2
   2     1    0.4
   3     2    0.3
   4     3    0.1

MTB > Random 1000 c3-c4;
SUBC>   Discrete 'X' 'p(x)'.
MTB > RMean C3 C4 C5.
MTB > Describe 'X-bars'.

MTB > DotPlot  'X-bars'.

Dotplot: X-bars

Each dot represents up to 21 points

                        .
                        :         :
                        :         :
                        :         :
               :        :         :        :
               :        :         :        :
               :        :         :        :         .
        :      :        :         :        :         :        .
        +---------+---------+---------+---------+---------+-------X-bars
     0.00       0.60       1.20       1.80       2.40       3.00

MTB > tally c5

Tally for Discrete Variables: X-bars

X-bars  Count
   0.0     28
   0.5    157
   1.0    296
   1.5    292
   2.0    159
   2.5     61
   3.0      7
    N=   1000

 Row   Mean1     StDev1  Variance1

   1   1.304   0.610292   0.372456

MTB >
```

The normality approximation improves when we took a random sample of size 5 from the same discrete distribution. The dot plot of 1000 random samples of size 5 from this distribution together with the estimated $\mu_{\bar{x}} = 1.3212$ and $\sigma^2_{\bar{x}} = 0.4029$ agree more closely with the theoretical values. Further, the shape of the dotplot is more symmetrical, indicating that the approximation improves with increasing sample size.

```
MTB > DotPlot  'means'.

Each dot represents up to 14 points
```




```
Data Display
 Row     Mean1     StDev1  Variance1

   1    1.3212   0.402930   0.162353
```

4.14.3 Sampling Distribution of $\bar{x}$: A Summary

The above results can be summarized succintly in the following.

1. If sampling is from a normally distributed population with mean μ and variance σ^2, then,
   (a) $\mu_{\bar{x}} = \mu$
   (b) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
   (c) The sampling distribution of $\bar{x}$ is normal.
2. If sampling is from a non-normally distributed parent population, then,
   (a) $\mu_{\bar{x}} = \mu$
   (b) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
   (c) The sampling distribution of $\bar{x}$ is approximately normal (Central Limit Theorem).

Example 4.14.3

If the uric acid values in normal adult males are approximately normally distributed with a mean and standard deviation of 5.7 and 1 mg %, respectively,

find the sampling distribution of $\bar{x}$ and the probability that a sample of size 9 will yield a means:

(a) Greater than 6 (b) between 5 and 6
(c) Less than 5.2

The sampling distribution of $\bar{x}$ will be normal with $\mu_{\bar{x}} = \mu = 5.7$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{9}} = 1/3 = 0.3333$.

(a)

$$P(\bar{x} > 6) = P\left(Z > \frac{6 - 5.7}{0.3333}\right) = P(Z > 0.90) = 1 - 0.8159 = 0.1841$$

(b)

$$P(5 < \bar{x} < 6) = P\left(\frac{6 - 5.7}{0.3333} < Z < \frac{5 - 5.7}{0.3333}\right) = P(0.91 < Z < -2.10)$$
$$= 0.8159 - 0.0179 = 0.7980$$

(c)

$$P(\bar{x} < 5.2) = P\left(Z < \frac{5.2 - 5.7}{0.3333}\right) = P(Z < -1.50) = 0.0668$$

Example 4.14.4

The partial pressure of oxygen, PaO_2, is a measure of the amount of oxygen in the blood. Assume that the distribution of PaO_2 levels among newborns has an average of 38 (mm Hg) and a standard deviation of 9. If we take a random sample of size 25, what is the probability that the sample mean will be

(i) greater than 36?
(ii) greater than 41?

In this example, the distribution of all PaO_2 levels in all newborns was not given. Hence by central limit theorem, the distribution of $\bar{x}$ will be approximately normal with $\mu_{\bar{x}} = \mu = 38$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{9}{\sqrt{25}} = \frac{9}{5} = 1.80$. Hence,

(a) $P(\bar{x} > 36) = P\left(Z > \dfrac{36 - 38}{1.80}\right) = P(Z > -1.11) = 1 - 0.1335 = 0.8665$

(b) $P(\bar{x} > 41) = P\left(Z > \dfrac{41 - 38}{1.80}\right) = P(Z > 1.67) = 1 - 0.9525 = 0.0475.$

4.14.4 Sampling Distribution of Population Proportion

Suppose a random sample of size n (usually large) is drawn from a population with an unknown proportion p of a certain attribute. Then the sampling distribution of $\hat{p}$ is approximately normal with

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Example 4.14.5

In a certain population of mussels (*Mytilus edulis*), 80 % of the individuals are infected with an intestinal parasite. A marine biologist plans to examine 100 randomly chosen mussels from the population. Find the probability that 85 % or more of the sampled mussels will be infected.

Here,

$$\mu_{\hat{p}} = p = 0.80$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.80 \times 0.20}{100}} = 0.04$$

Hence,

$$3P(\hat{p} > 0.85) = P\left(Z > \frac{0.85 - 0.80}{0.04}\right) = P(Z > 1.25) = 1 - 0.8944 = 0.1056$$

If in the above example, the biologist takes a random sample of 50, what would be the probability that fewer than 35 of the sampled mussels would be infected? Here again, $\mu_{\hat{p}} = p = 0.80$ and $\sigma_{\hat{p}} = \sqrt{\frac{0.80 \times 0.20}{50}} = 0.0566$
Hence again,

$$3P(\hat{p} < 0.70) = P\left(Z < \frac{0.70 - 0.80}{0.0566}\right) = P(Z < -1.77) = 0.0384$$

Since 35 is 70 % of 50.

4.15 Exercises

1. Evaluate each of the following: $(a)\ \binom{8}{5}$ $(b)\ \binom{7}{0}$ $(c)\ \frac{8!}{5!3!}$ $(d)\ {}_7P_3$
2. Use MINITAB to simulate the rolling of a a fair die a total of 10,000 times. Also obtain the relative frequencies of the number of dots on the die. Are these approximately $\frac{1}{6}$ each?
3. Find the area under the standard normal curve that lies

(a) to the left of $z = 2.85$
(b) To the right of $z = -1.42$
(c) either to the left of $z = -2$ or to the right of $z = 1$.
(d) between $z = -1.66$ and $z = 2.85$.
(e) Find $z_{0.68}$.

4. The number on a branch is a random variable X which takes values x with probability $p(x) = Kx$, $x = 1, 2, \cdots, 10$.
Find the value of K in order that *p(x)* is a probability distribution. Show that the probability that a branch has two shoots is $2/55$ and that the mean number of shoots on a branch will be 7.
5. The time taken after planting for winter wheat crops to reach maturity is normally distributed with a mean of 183 days and variance 100 days (squared). If the crops are planted on 1st October, what percentage will mature
(i) before April (ii) during April (iii) after May.
(Assume the month of February has 28 days)
6. Given a normal distribution with $\mu = 200$ and $\sigma^2 = 100$. Find

(i) the area below 214
(ii) the area above 179
(iii) the point that has 80 % of the areas below it.
(iv) the two points containing the middle 75 % of the area.

7. In a standard normal distribution, find the z-scores that cuts off the top
(a) 28 %, (b) 35 %, (c) 5 %, (d) 18 %
8. Given a normal distribution with $\mu = 40$ and $\sigma = 6$, find:

(i) The area below 32
(ii) The area above 27
(iii) The area between 42 and 51
(iv) The point that has 45 % of the area below it
(v) The point that has 13 % of the area above it.

9. If X is normally distributed with $\mu = 100$ and $\sigma^2 = 64$. Find

(i) $P(X < 107)$ (ii) $P(X < 97)$ (iii) $P(X > 110)$
(iv) $P(X > 90)$ (v) $P(95 < X < 106)$ (vi) $P(103 < X < 114)$
(vii) $P(88 < X < 100)$ (viii) $P(60 < X < 108)$

10. A population of marine gastropods have shell lengths which are normally distributed with mean 7 mm and standard deviation 1.44 mm. What proportion of the population will have a shell length between 5 and 9 mm?
11. The transit times, T, of impulses across a membrane are measured in μ sec, and are distributed according to a $N(1860, 4624)$. Find
(i) $P(T \geq 2000)$ (ii) $P(T \leq 1800)$
If samples of four impulses are observed, find
(iii) $P(\bar{T} \geq 1930)$ (iv) $P(\bar{T} \leq 1870)$
where $\bar{T}$ is the mean transit time of the Sample.

12. If $X \sim N(3, 9)$, find
(i) $P(2 < x < 5)$ (ii) $P(X > 0)$ (iii) $P(\mid X - 3 \mid > 6)$
13. If $X \sim N(100, 64)$, find b such that:
(i) $P(X > b) = 0.8708$
(ii) $P(X < b) = 0.4030$
14. On average 80 % of the tomato seeds in a packet will germinate. Find the probability that
(i) One taken from a packet will germinate
(ii) None of five seeds taken from a packet will germinate
(iii) At least one of five seeds taken from a packet will germinate.
15. In a field trial on crop yield, each of the chemicals phosphorus, nitrogen and potassium can be applied to plots at one of three different concentrations. How many plots will be needed if the crop yields for every combination of phosphorus, nitrogen and potassium concentrations are to be examined?
16. Prior to examination, the packed cell volume (PCV) values in six Zebu and nine N'dama cattle are measured. In how many ways can the examination take place if
(i) all the cattle are to be examined in order to increase PCV value,
(ii) the Zebu cattle only are to be examined in order to increase PCV value,
(iii) The N'Dama cattle are to be examined in order to increase PCV value followed by the Zebu cattle in order to increase PCV value?
17. A medical research team wished to evaluate a proposed screening test for Alzheimer's disease. The test was given to a random sample of 450 patients with Alzheimer's disease and an independent random sample of 500 patients without symptons of the diseases. The two samples were drawn from populations of subjects who were 65 years of age or older where it is assumed that 11.3 % of the US population aged 65 and over have Alzheimer's disease. The data from this study is presented below:

	Alzheimer's diagnosis?		
Test result	Yes (D)	No ($\bar{D}$)	Total
Positive (T)	436	5	441
Negative ($\bar{T}$)	14	495	509
Total	450	500	950

18. Suppose x has a hypergeometric probability distribution with $N = 15, n = 8$, and $r = 5$.
a Compute $p(x)$ for $x = 0, 1, 2, 3, 4, 5$.
b $P(x = 3)$
c $P(x \leq 2)$
d $P(x > 3)$

19. The following table shows the density for a random variable X, the number of persons seeking emergency treatment unnecessarily per day in a small hospital.

X	0	1	2	3	4	5
P(x)	.01	.1	.3	.4	.1	-

Determine:

(i) The missing probability
(ii) Find F, the cumulative density function.
(iii) $P(X = 4)$
(iv) $P(X \leq 2)$
(v) $P(1 < X \leq 4)$
(vi) $P(X > 4)$
(vii) The mean of the random variable X.

20. The probability that a student at a certain college will catch a mild cold in winter is 0.60. If twenty students are randomly selected at random during winter, find using the output below the probability that:

(i) Exactly 14 students will catch the cold?
(ii) Between 10 and 15 (inclusive) will catch the cold.
(iii) At most 14 students will catch the cold?
(iv) At least 11 students will catch the cold?
(v) More than five but less than 16 will catch the cold?

```
 X     CUMPROB

 0      0.0000
 1      0.0000
 2      0.0000
 3      0.0000
 4      0.0003
 5      0.0016
 6      0.0065
 7      0.0210
 8      0.0565
 9      0.1275
10      0.2447
11      0.4044
12      0.5841
13      0.7500
14      0.8744
15      0.9490
16      0.9840
17      0.9964
18      0.9995
19      1.0000
20      1.0000
```

21. Show that $p(x) = \dfrac{9-x}{45}$, for $x = 1, 2, 3, \cdots, 8$ is a probability density function. Hence, compute its mean and standard deviation μ_x and σ_x respectively.
22. In the game of craps, two balanced dice are rolled. Let

    A = event the sum of the dice is 7
    B = event the sum of the dice is 11
    C = event the sum of the dice is 2
    D = event the sum of the dice is 3
    E = event the sum of the dice is 12
    F = event the sum of the dice is 8
    G = event that we observe doubles or sum on the dice is 8

    Compute the probabilities of each of the seven events listed above.
23. When an experimental stimulus is given an animal, it either responds or fails to respond. In other words, there are only two possible outcomes when stimulus is applied to an animal. Either it responds (R) or it does not (N). An experiment consists of administering the stimulus to three animals in succession and recording R or N for each animal. Find the probability of the following events.

    (i) Only one animal responds.
    (ii) There is a response in the first trial.
    (iii) Both the first and third animals fail to respond.

24. Scientists have developed a test for determining when the mercury level in fish is above acceptable level. If the fish actually contain an excessive amount of mercury, then the test is 99 % effective in determining this, and only 1 % will escape detection. On the other hand,if the mercury level is within acceptable limits, then the test will correctly indicate this 96 % of the time. Suppose the test is to be used on fish from a river that that has been polluted by a chemical company, and it is estimated that 30 % of the fish in the river contain excessive amounts of mercury. If a fish is caught and tested by the scientist, and the test indicates that the mercury level is within acceptable limits, what is the probability that the mercury content is actually greater than the acceptable level.
25. It is thought that 30 % of all people in the United States are obese (A_1) and that 3 % suffer from diabetes (A_2). 31 % are obese or suffer from diabetes. What is the probability that a randomly selected person

    (a) Have both obese and diabetes
    (b) Is diabetic given that he/she is obese?
    (c) Is diabetic but is not obese?
    (d) Is diabetic given that he/she is not obese?

26. In a large city, suppose that 10 % of the adult male population has heart disease. A new technique to detect heart disease using a radiopaque dye

is tested. Suppose that positive tests are obtained for 80 % of those with known heart disease and 5 % of those known to be free of heart disease. If an adult male is selected at random from this population:
Find:

(a) The sensitivity of the test.
(b) The specificity of the test.
(c) The false negative rate.
(d) Using a tree diagram, find the probability that he tests positive.
(e) Given that this individual tests positive, what is the probability that he has heart disease?

Among females in the United States between 18 and 74 years of age, diastolic blood pressure is normally distributed with mean $\mu = 77$ mm Hg and standard deviation $\sigma = 11.6$ mm Hg.

(a) What is the probability that a randomly selected woman has a diastolic blood pressure less that 60 mm Hg?
(b) That she has a DBP greater than 90 mm Hg?
(c) That the woman has a DBP which is in the top 8 % of the population?
(d) What is the probability that the mean of a random sample of size 16 from this population, will be greater than 90 mm Hg? That is find $P(\bar{x} > 90)$.

This exercise is drawn from Pagano and Gauvreau (2000). The table below presents results from the study of self-reported smoking status with measured serum cotinine level. Cotinine level was used as a diagnostic tool for predicting smoking status. For a set of cutoff points, the observed sensitivities and specificities are presented below.

Continine level (ng/ml)	Sensitivity	Specificity
5	0.971	0.898
7	0.964	0.931
9	0.960	0.946
11	0.954	0.951
13	0.950	0.954
14	0.949	0.956
15	0.945	0.960
17	0.939	0.963
19	0.932	0.965

(i) How does the probability of a false positive result changes as the cutoff is raised? How does the probability of a false negative result change?
(ii) Use the above data to construct an ROC curve.
(iii) Based on the ROC curve, what value of serum cotinine level would you choose as an optimal cutoff point for predicting smoking status and why?

The following table shows the cumulative distribution function $F(x)$ for a random variable X, the number of wing beats per second of a species of large moth while in flight.

X	6	7	8	9	10
$F(x)$	0.05	0.15	0.75	0.90	1.00

Determine:

(i) Find $P(X \leq 8)$
(ii) $P(X > 7)$
(iii) $P(7 \leq X \leq 9)$
(iv) $P(X = 8)$
(v) $P(X \geq 8)$
(vi) Find $P(x)$, the probability density function of X.
(vii) The mean of the random variable X.

4. Suppose that 60 % of the voting population in a city, about to have a referendum on adding fluoride to the drinking water, favor fluoridation. A sample of 16 persons are interviewed. What is the probability that the number of people who will favor fluoridation is: (USE MINITAB)

(i) Exactly eight.
(ii) Between six and twelve, inclusive.
(iii) At least six.
(iv) No more than 10

An individual is selected at random from a convalescent home in which 30 % have a particular disease and is given a screening test to detect the presence of the disease. Let D denote the event that the person selected has the disease and let S indicate a positive result on the screening test. The probability of a positive screening test result given that the person selected has the disease is 0.92. The corresponding probability for a non-diseased person is 0.15. Find:

(a) The specificity and sensitivity of the test.
(b) The false negative rate.
(c) Using a tree diagram, find the probability that a person has the disease given that the screening test is positive? That is, find $P(D|S)$.

Explain why each of the following distributions is or is not a probability distribution.

x	P(X=x)
0	0.15
1	0.25
2	0.10
3	0.25
4	0.30

(a)

x	P(X=x)
0	0.15
1	-0.20
2	0.30
3	0.20
4	0.15

(b)

x	P(X=x)
-1	0.15
0	0.30
1	0.20
2	0.15
3	0.20

(c)

The probability density function of a random variable X is given by:

X	-2	-1	0	1	2
p(x)	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

Find:

(i) $P(X \leq 2)$
(ii) $P(-1 \leq Xle1)$
(iii) $P(X \leq 1 \text{ or } X = 2)$

Show that the following is a probability density function, and hence, find

$$f(x) = \frac{8}{7}\left(\frac{1}{2}\right)^x, \; x = 1, 2, 3$$

(a) $P(X \leq 1)$
(b) $P(2 < X < 6)$
(c) $P(X \leq 1 \text{ or } X > 2)$.

Chapter 5
Estimation and Hypotheses Testing

5.1 Confidence Intervals

In this chapter, we shall discuss statistical inference where data collected will be viewed as a random sample from some population. The information so gathered from such a sample will then be used to conduct a Statistical estimation which basically comprises determining an estimate of some parameter of the population as well as assessing the precision of such an estimate. We present an example in Table 5.1, relating to contamination counts of a sample of 20 vaccines preserved with phenol.

Table 5.1 Contamination counts from a sample of 20 bacterial vaccines

67	62	52	55	54
61	51	59	54	57
57	60	50	66	68
54	53	52	58	56

From the above data,

$$\bar{x} = \frac{67 + 62 + \cdots + 56}{20} = 57.30$$

Thus if our random sample is assumed to have come from a population with mean μ and standard deviation σ, then, 57.3 is an estimate of μ and $\bar{x}$ will therefore be said to be an estimator for μ or an estimate of μ. Similarly,

$$s = \sqrt{\frac{\sum(x_i - 57.3)^2}{19}} = 5.32$$

Again, 5.32 is an estimate of σ and s will be called an estimator for σ. Both $\bar{x}$ and s are called statistics because they are random variables themselves. They vary with each random sample of size $n = 20$ drawn from this population. The values 57.3 and 5.32 are *point estimates* of μ and σ respectively.

Since $\bar{x}$ itself is a variable, it follows that any single observed value of $\bar{x}$, will not exactly equal the population mean μ, and researchers have found

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_5,

it desirable to have an idea of how close our estimate is to the true population mean. The method often employed for this is the method of interval estimation or *confidence intervals*.

5.1.1 Building a Confidence Interval

From our results in Chap. 4, we recall that if $X_1, X_2, \cdots, X_n$ is a random sample of size n from a population that is normally distributed with mean μ and variance σ^2. Then, $\bar{x}$ will also be normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. Hence,

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{5.1}$$

would be a standard normal variate. Thus to build a 95 % confidence interval for μ, we need L and U such that

$$P(L \leq \mu \leq U) = 0.95 \tag{5.2}$$

To find such L and U, we note that

$$P(-1.96 \leq Z \leq 1.96) = 0.95 \tag{5.3}$$

Hence replacing the Z in (5.3) with the defined Z in (5.1), we have,

$$P(-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95$$

$$P\left(\frac{-1.96\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95$$

Rearranging and isolating μ, we have,

$$P\left(\bar{x} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95 \tag{5.4}$$

From (5.4), we see that

$$L = \bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \quad \text{and} \quad U = \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

The lower and upper bounds therefore are L and U respectively and the 95 % confidence interval can therefore be succinctly written as:

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}.$$

In general, a $100(1-\alpha)\,\%$ confidence interval for the population mean μ if σ^2 were known is:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) = \bar{x} \pm z_{\alpha/2}\,\sigma_{\bar{x}} \tag{5.5}$$

$(1-\alpha)$ is called the confidence coefficient and $z_{\alpha/2}$ is the value of Z such that $\alpha/2$ of the area lies to its right. For instance, to obtain the Z value that corresponds to a 99 % confidence interval, we do the following:

$$\begin{aligned} 100(1-\alpha) &= 99; \quad \text{hence,} \\ 1-\alpha &= 0.99 \\ \alpha &= 0.01 \\ \alpha/2 &= 0.005 \end{aligned}$$

Thus, we need $z_{0.005} = 2.575$. If the population variance σ^2 for the vaccine example had been 25, then, a 95 % confidence interval for μ is:

$$57.30 \pm 1.96\frac{5}{\sqrt{20}} = 57.30 \pm 2.19 = (55.11, 59.49)$$

We can therefore be 95 % confident that the unknown population mean number of contaminations on the vaccines is between 55.11 and 59.49, that is between 56 and 60. We can implement above in MINITAB, by assuming that σ is known.

```
MTB > print c1

Data Display
Counts
     67     62     52     55     54     61     51     59     54     57
     57     60     50     66     68     54     53     52     58     56

MTB > OneZ 'Counts';
SUBC>   Sigma 5;
SUBC>   Confidence 95.

One-Sample Z: Counts

The assumed sigma = 5

Variable        N      Mean     StDev   SE Mean        95.0% CI
Counts         20     57.30      5.32      1.12  (55.11, 59.49)
```

The effect of changing the confidence coefficient is presented in Table 5.2. We observe that increasing the confidence coefficient widens the confidence interval and reducing the coefficient shortens the interval.

Table 5.2 Effect of increasing confidence coefficient on confidence width

$1-\alpha$	Confidence interval	Width
0.90	(55.46, 59.14)	3.68
0.95	(55.11, 59.49)	4.38
0.99	(54.42, 60.18)	5.76

5.1.2 Sample Size Determination

The $100(1-\alpha)\,\%$ confidence interval for μ was found to be equal to $\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$, where $z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$ is often referred to as the margin of error. That is,

$$\bar{x} \pm \underbrace{z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}_{\text{margin of error}}$$

For a given margin of error d, therefore, the sample size required to have a specified $100(1-\alpha)\,\%$ confidence interval in our estimate of the population mean μ is obtained from the following:

$$d = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$d^2 = \frac{z^2\,\sigma^2}{n}, \quad \text{hence}$$

$$n = \frac{z^2\,\sigma^2}{d^2}$$

n should be *ranked* to the next integer.

Example 5.1.1

A research scientist wants to know how many times per hour a certain strand of bacteria reproduces. He believes that the variance is 3.61 and the mean is 9.5. How large a sample would be required in order to estimate the average number of reproductions at 90 % confidence level with an error of at most 0.18 reproductions? Here, $d = 0.18$, $z_{.05} = 1.645$ and $\sigma^2 = 3.61$. Hence,

$$n = \frac{(1.645)^2\ 3.61}{(0.18)^2} = 301.5 \quad \text{That is, } n = 302$$

5.1.3 Case of σ Not Known and n Large ($n \geq 30$)

If σ, the population standard deviation, is unknown but n is large (≥ 30), then we can replace the unknown σ with the corresponding sample standard deviation obtained from $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$ and the $100(1-\alpha)\,\%$ large sample confidence interval for μ is now given by:

$$\bar{x} \pm z_{\alpha/2} \left(\frac{s}{\sqrt{n}}\right) \tag{5.6}$$

Example 5.1.2

5.1.4 Case of σ Not Known and n Small ($n < 30$)

If σ is unknown and $n < 30$, then, a $100(1-\alpha)\,\%$ confidence interval for μ would now be given by

$$\bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}}\right) \tag{5.7}$$

where $t_{\alpha/2}$ is the upper $\alpha/2$ tail of a Students' t distribution with $n-1$ degrees of freedom (d.f.) and s is the sample standard deviation estimated from the data. The above analysis is based on the assumption that the sample was drawn from a normally distributed population with mean μ and unknown variance σ^2. This assumption would would have to be validated with a normality test.

Example 5.1.3

A plant pathologist grew 13 individually potted soybean seedlings as part of a study on plant growth. The plants were raised in a greenhouse under identical environmental conditions (light, temperature, soil, etc.) and she then measured the total length (centimeter) for each plant after 16 days of growth. These data are presented in Table 5.3.

Table 5.3 Stem length of soybean plants

20.2	22.9	23.3	20.0	19.4
22.0	22.1	22.0	21.9	21.5
19.7	21.5	20.9		

The summary statistics for the above data are displayed as:

$$\sum x = 277.4, \quad \sum x^2 = 5937.12, \quad n = 13$$

Hence, $\bar{x} = \dfrac{277.4}{13} = 21.3385$, $s^2 = \dfrac{\sum x^2 - n\bar{x}^2}{n-1} = \dfrac{5937.12 - 13(21.3385)^2}{12} = 1.4859$ and therefore $s = \sqrt{1.4859} = 1.21897$. To obtain a 99 % confidence interval for μ for instance, we note from Table 2 in the appendix that $t_{0.005}$ with 12 d.f. = 3.0545. Hence, the confidence interval is computed as:

$$21.339 \pm 3.0545 \left(\frac{1.2190}{\sqrt{13}} \right) = 21.339 \pm 1.033 = (20.306,\ 22.372)$$

The 99 % confidence interval is implemented in MINITAB with the following, where the data have been read into a column named "Length."

```
MTB > OneT 'Length';
SUBC>   Confidence 99.

One-Sample T: Length

Variable          N      Mean     StDev   SE Mean          99.0% CI
Length           13    21.338     1.219     0.338  (20.306,  22.371)
```

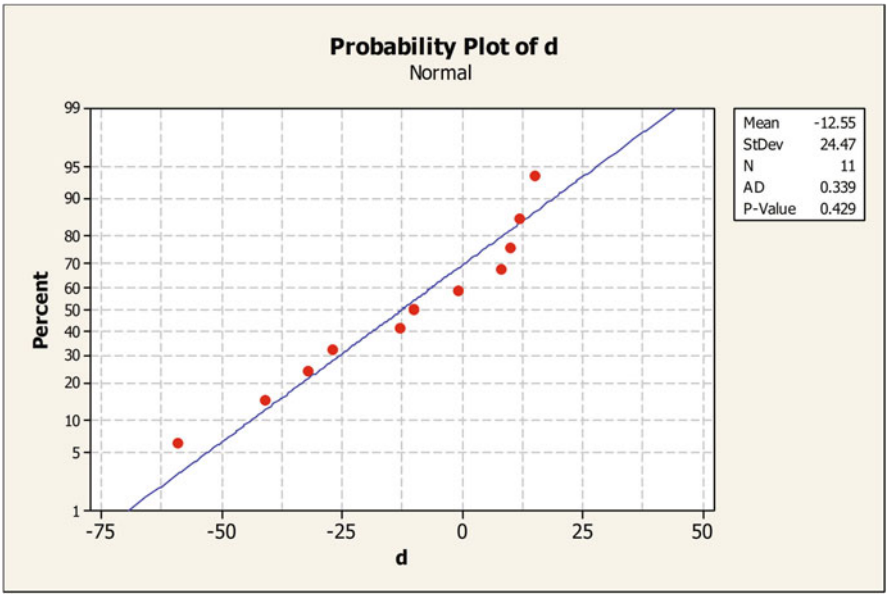


Fig. 5.1 Normal probability plot and test

The test of the assumption of normality is carried out with the normal probability plot in Fig. 5.1. The Anderson–Darling test indicates that these data can be assumed to have come from a normal population.

5.1.5 Confidence Interval for a Population Proportion

To estimate a population proportion, we assume that a sample of size n is drawn from a population of interest and the number of subjects having the characteristic of interest is observed, (say, X). Thus X subjects have the characteristic of interest and therefore $n - X$ do not have the characteristic of interest. An estimate of the population proportion p is therefore $\hat{p} = \dfrac{X}{n}$, $\hat{p}$ is read p-hat. A $100(1-\alpha)\,\%$ for p is therefore given by:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{5.8}$$

Here again, the quantity $z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is called the margin of error.

Example 5.1.4

In evaluating the policy of routine vaccination of infants for whooping cough, adverse reactions were monitored in 339 infants who received their first injection of vaccine. Reactions were noted in 69 of the infants. To construct a 95 % confidence interval for p, the proportion of all infants have reactions to first time injection of the vaccine, we have:

Solution

The sample proportion of infants having reactions to the vaccine is $\hat{p} = \frac{69}{339} = 0.204$. The sample size is large enough in this example. The estimate of the standard error $\sigma_{\hat{p}}$ is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{(0.2035)(0.7965)}{339} = 0.022$. The 95 % confidence interval for p, based on the above data is:

$$0.204 \pm 1.96(0.022)$$
$$0.204 \pm 0.043$$
$$(0.161, 0.247)$$

We are 95 % confident that the proportion of adverse reaction in infants who receive their first injection of the vaccine is between 0.161 and 0.247.

5.1.6 Sample Size Determination for Estimating a Proportion

As in the case for the mean, given a one half of the desired interval, d, the margin of error, then, the sample size required to achieve this desired margin

of error at a $100(1-\alpha)\,\%$ confidence interval is obtained by setting d equal to the margin of error, to obtain n. That is,

$$d = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$d^2 = z^2_{\alpha/2}\,\frac{\hat{p}(1-\hat{p})}{n}, \quad \text{hence}$$
$$n = \frac{z^2_{\alpha/2}\,\hat{p}(1-\hat{p})}{d^2}$$

The above expression calls for $\hat{p}$ which we do not know as this can only be obtained after the sample has been drawn. Thus, if prior information about $\hat{p}$ is available, we would use this instead. Thus, for a known (prior value) p, the above formula becomes:

$$n = \frac{z^2_{\alpha/2}\,p(1-p)}{d^2} \tag{5.9}$$

However, if no prior informed guess of p is available, then we would use the following formula to obtain n, viz:

$$n = \frac{z^2_{\alpha/2}}{4d^2} \tag{5.10}$$

The expression in (5.10) is based on the fact that $p(1-p)$ in (5.9) is largest when $p = 0.5$, and in this case, $p(1-p) = \frac{1}{4}$. Thus a value of n calculated from the expression in (5.10) will be *conservative*, in the sense that it will be large enough.

Example 5.1.5

A hospital administrator wishes to know what proportion of discharged patients are unhappy with the care received during hospitalization. How large a sample should be drawn if we let $d = 0.05$. The confidence coefficient is 0.95, and no other information is available. How large should the sample be if p is approximated by 0.22?

Solution

In the first part of the question, p is unknown, hence we would use the expression in (5.10) to obtain n. Here, $Z_{\alpha/2} = 1.96$ and $d = 0.05$, thus,

$$n = \frac{(1.96)^2}{4(.05)^2} = 384.16 = 385.$$

Thus $n = 385$ in this case. In the second part of the question, p is given as 0.22, hence in this case, we would use the expression for n in (5.9), thus,

$$n = \frac{(1.96)^2 \times 0.22 \times 0.78}{(0.05)^2} = 263.69 = 264.$$

That is, we would require a sample size of $n = 264$ in this case.

5.2 Confidence Interval for the Difference of Two Population Means

5.2.1 Distribution of Difference of Two Means

Suppose we have two populations X_1 and X_2 such that X_1 is distributed normal with mean μ_1 and variance σ_1^2. Similarly, X_2 is distributed normal with mean μ_2 and variance σ_2^2. That is,

$$X_1 \sim N(\mu_1, \sigma_1^2)$$
$$X_2 \sim N(\mu_2, \sigma_2^2)$$

If a random sample of sizes n_1 and n_2 is taken from these populations respectively, it follows that

$$\bar{x}_1 = \frac{\sum X_1}{n_1} \sim N\left[\mu_1, \frac{\sigma_1^2}{n_1}\right]$$

and

$$\bar{x}_2 = \frac{\sum X_2}{n_2} \sim N\left[\mu_2, \frac{\sigma_2^2}{n_2}\right]$$

Since the two samples are random and independent, hence,

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \tag{5.11}$$

that is, the difference of two sample means is also normally distributed with a mean that equals the difference of the two means and a variance that equals the sum of their variances.

For two independent populations I and II with means μ_1, μ_2 and variances σ_1^2 and σ_2^2 respectively, suppose random samples of sizes n_1, n_2 were drawn respectively and sample means $\bar{x}_1$ and $\bar{x}_2$ were appropriately computed. Further, if the populations are normally distributed with known variances σ_1^2 and σ_2^2 respectively, then a $100(1-\alpha)\,\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{5.12}$$

This result stems from the result in (5.11). However, if the populations are not normally distributed, but if the sample sizes n_1 and n_2 are large ($n_1, n_2 \geq 30$), then, a large sample $100(1-\alpha)\,\%$ confidence interval for $\mu_1 - \mu_2$ will now be given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5.13}$$

where s_1^2 and s_2^2 are sample variances computed from the samples respectively.

Example 5.1.6

The bacterial count in the mouths of 10 patients just admitted to hospital was as follows:

1570, 2275, 1194, 7006, 9993,
4034, 8608, 7976, 7280, 6337.

A second group of 12 patients who had spent 6 days in hospital gave the following counts:

9709, 9847, 5292, 7751, 9038, 4030
4011, 7325, 7054, 5877, 8074, 5247

If the bacteria counts are known to have a population standard deviation of 2500 for each group of patients, find the 95 % confidence interval for $(\mu_1 - \mu_2)$.

Solution

Let μ_1 and μ_2 be the means of the two groups. Then,

$$\bar{x}_1 = \frac{\sum x_1}{10} = 5627.3$$

$$\bar{x}_2 = \frac{\sum x_2}{12} = 6937.9$$

Hence, the 95 % confidence interval for $(\mu_1 - \mu_2)$ is computed using expression in (5.12) as,

$$(5627.3 - 6937.9) \pm 1.96\sqrt{\frac{2500^2}{10} + \frac{2500^2}{12}} = -1310.6 \pm 2098.1$$

$$= (-3406.7, 787.5)$$

We are therefore 95 % confident that the true difference $(\mu_1 - \mu_2)$ of bacteria counts in the two groups is between -3406.7 and 787.5. Since the confidence interval includes zero, we conclude that the two population bacteria count means are equal.

5.2.2 Case When $(n_1, n_2) < 30$ and σ_1, σ_2 Are Unknown

Here again, two independent random samples of sizes n_1 and n_2 are each drawn from two independent normal populations with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2) respectively. However, in this situation σ_1^2 and σ_2^2 are unknown. This implies that they would have to be estimated from the samples.

In order to proceed for this analysis, we would need to make the assumption that although σ_1^2 and σ_2^2 are not known, but they are being assumed to be equal, that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, a common unknown population variance. In this situation the populations are said to be homogeneous.

Under these conditions, we have

$$\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right), \quad \bar{x}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right) \tag{5.14}$$

and therefore

$$(\bar{x}_1 - \bar{x}_2) \sim N\left[(\mu_1 - \mu_2), \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right] \tag{5.15}$$

But then, σ^2 still remains unknown, and we can estimate it from each of the two samples thus:

$$s_1^2 = \frac{(\sum x_{1i} - \bar{x}_1)^2}{n_1 - 1}, \qquad s_2^2 = \frac{(\sum x_{2i} - \bar{x}_2)^2}{n_2 - 1}$$

These are two estimates of the same parameter. We can therefore pool these two estimates together to get a unified (pooled) estimate

$$S_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{5.16}$$

S_P^2 is an unbiased estimate of σ^2 and hence, a $100(1-\alpha)\,\%$ confidence interval for $\mu_1 - \mu_2$ will be given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{S_P^2}{n_1} + \frac{S_P^2}{n_2}}$$

where $t_{\alpha/2}$, is the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom. We should note here that the above analysis is contingent on the following assumptions being true:

1. That the populations in which the samples were drawn are homogeneous. That is, $\sigma_1^2 = \sigma_2^2$. In other words, that the population variances are equal.
2. That the samples were drawn from populations that are independently normally distributed.

Example 5.1.8

The activity of an enzyme (units per gram protein) in 12 liver tissues infected with hepatitis and 18 normal liver tissues was as follows:

Hepatitis Liver tissue :

4.15, 4.48, 4.22, 3.94, 4.52, 3.70
4.77, 4.03, 3.90, 4.86, 3.16, 3.33

Normal Liver tissues:

3.15, 4.23, 3.12, 2.70, 3.99, 4.40
3.86, 3.86, 3.16, 4.27, 4.34, 3.79
4.28, 4.63, 4.98, 3.52, 2.77, 3.18

Construct a 95 % confidence interval for the difference in their population means. Does it appear that the mean activity of enzyme in the hepatitis group is higher than the normal group? Why do you reach this conclusion?

Solution

Let μ_1 = mean of Hepatitis and μ_2= mean of Normal liver patients.

$$\sum x_1 = 49.06, \sum x_1^2 = 203.6432 \text{ and } n_1 = 12$$

$$\bar{x}_1 = 4.088, s_1^2 = \frac{\sum x_1^2 - \frac{(\sum x_1)^2}{n_1}}{n_1 - 1} = 0.27915$$

Similarly,

$$\sum x_2 = 68.23, \quad \sum x_2^2 = 266.0687, \quad \text{and } n_2 = 18$$

$$\bar{x}_2 = 3.791, \quad , s_2^2 = 0.4376$$

Hence, Pooled variance S_P^2 equals

$$S_P^2 = \frac{11 \times s_1^2 + 17 \times s_2^2}{28} = 0.3753 \tag{5.17}$$

and $S_P = \sqrt{S_P^2} = 0.6126$.
$t_{\alpha/2}(28) = 2.048$ and hence, the 95 % confidence interval for $(\mu_1 - \mu_2)$ is computed as:

$$(4.088 - 3.791) \pm 2.048\sqrt{\frac{0.3753}{12} + \frac{0.3753}{18}} = (0.297) \pm 0.468$$

$$= (-0.171,\ 0.765) \tag{5.18}$$

```
MTB > PRINT C1-C2

Data Display


 Row  HEPTIT  NORMAL

   1    4.15    3.15
   2    4.48    4.23
   3    4.22    3.12
   4    3.94    2.70
   5    4.52    3.99
   6    3.70    4.40
   7    4.77    3.86
   8    4.03    3.86
   9    3.90    3.16
  10    4.86    4.27
  11    3.16    4.34
  12    3.33    3.79
  13            4.28
  14            4.63
  15            4.98
  16            3.52
  17            2.77
  18            3.18

MTB > TwoSample 'HEPTIT' 'NORMAL';
SUBC>   Pooled.

Two-Sample CI: HEPTIT, NORMAL

Two-sample T for HEPTIT vs NORMAL

          N      Mean     StDev   SE Mean
HEPTIT   12     4.088     0.528      0.15
NORMAL   18     3.791     0.662      0.16

Difference = mu HEPTIT - mu NORMAL
Estimate for difference:  0.298
95% CI for difference: (-0.170, 0.765)
Both use Pooled StDev = 0.613
```

We have assumed for now that the two assumptions above are satisfied for these data set. We shall revisit this example and conduct the necessary tests after we have discussed hypothesis testing.

5.3 Confidence Interval for the Difference of Two Population Proportions

Suppose we have two binomial populations, A and B. If a random sample of size n_1 is taken from population A and the number of successes is denoted by X. Another independent random sample of size n_2 is taken from population B and the number of successes is denoted by Y.

Then, from the sample from population A, we have $\hat{p}_1 = \dfrac{X}{n_1}$. Similarly, from the sample from population B, we also have $\hat{p}_2 = \dfrac{Y}{n_2}$. An unbiased point estimator for the difference between the two population proportions is $\hat{p}_1 - \hat{p}_2$, with standard error,

$$\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}}.$$

Hence, a $100(1-\alpha)\,\%$ confidence interval for $p_1 - p_2$ will be given by

$$(\hat{p}_1 - \hat{p}_2) \pm\ z_{\alpha/2}\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}}.$$

Example 5.3.1

Aronow and Kronzon (1991) identified coronary risk factors among men and women in a long-term health care facility. Of 215 subjects who were black, 58 had diabetes mellitus. Of the 1140 white subjects, 217 had diabetes mellitus. We wish to construct a 90 % confidence interval for the difference between the two population proportions.

Solution

The sample proportion for blacks with diabetes is $\hat{p}_1 = 58/215 = 0.2698$. Similarly, for whites, is $\hat{p}_2 = 217/1140 = 0.1904$. Hence, the 90 % confidence interval is computed as:

$$(0.2698 - 0.1904) \pm 1.645\sqrt{\frac{(0.2698)\,(0.7302)}{215} + \frac{(0.1904)\,(0.8096)}{1140}}.$$

$$0.0794 \pm 0.0533 = (0.0261,\ 0.1327)$$

5.4 Hypothesis Testing

5.4.1 Concepts and Definitions

The concepts and definitions presented here were adapted from a set of notes shared with me almost 30 years ago. I liked the take on this topic; therefore, I have adapted it for this chapter. This source is duly acknowledged at the end of this text.

A hypothesis is a statement about population characteristics. The hypothesis is tested through experimental investigations to ascertain its plausibility. A hypothesis that is relatively well verified and possesses some degree of generality is a *theory*. A theory that has been verified beyond all reasonable doubt at the moment is designated a *law*.

There are two types of scientific researches, namely, empirical and analytical. The former deals with experimental investigations which involve data while the latter deals with laws, axioms, postulates, and definitions in the field of inquiry. Much of our research is empirical in nature as it involves measurement and observations on various characteristics. Empirical facts only substantiate the claim for the hypothesis; they do not prove it.

The object of a scientific investigation is not to prove the scientist correct, but to establish the truth.

A necessary part of research is Inference, which is a process of reasoning. Inference may be deductive or inductive. *Deductive Inference* is the process of determining the implications inherent in a set of propositions, and it is always associated with analytic research.

Aristotle was the first to stress the systemic nature of science and to teach the use of reasoning in the development of science. Syllogism is one form of logical deduction that was used to a large extent; this method of deductive inference begins with two main premises, usually a major premise and a minor premise, or propositions which are so related in thought that a person is able to infer a third proposition from them. The following example(s) illustrate this method of deductive Inference

(i)

Major Premise:	All living plants absorb water (inductive)
Minor Premise:	This tree is a living plant (observation)
Conclusion:	Therefore, this tree absorbs water (deduction)

(ii)

Major Premise:	Human beings are composed of men and women (Inductive)
Minor Premise:	This person is a man (observation)
Conclusion:	Therefore, this person is a human being (deduction)

We can see that knowledge of the past furnishes the major premise, a particular problem or situation supplies the minor premise. The deduction obtained by the psychological process of reasoning constitutes the conclusion.

Inductive Inference

forms a large part of the definition of the subject of Statistics. This type of inference is characterized by the fact that from the sample facts, we draw conclusions concerning population facts, i.e., it is reasoning from the results obtained from a sample, a part or an experiment of characteristics of the

entire population or of all the members of a class. Thus inductions are based on partial evidence and as such are characterized by a degree of uncertainty. The evidence for inductive inference may often be stated on a probability basis.

Two other terms *Sample Survey* and *Experiment* require definition. A Sample Survey is an investigation of what is present in the population. When the sample is a 100 % sample, it is called the census. Any observation that appears in the population could appear in the sample. Any condition not represented in the population will not be observed in a sample or a census. In many investigations however, it is desired to investigate conditions which do not appear in a population. In an *experimental investigation* or *experiment*, the experimenter may, and often does, introduce conditions which do not exist in any naturally occurring population. The investigator controls the conditions in the experiment whereas the conditions in a survey are those that prevail in the population.

5.4.2 Test of Significance

Sometimes one wants to compare an observed result with a prior hypothesis or expectation. If the observed result differs from the hypothesis but is based on a sample, a test of *significance* is needed. This in effect determines whether the difference is possibly real or only due to sampling error.

When studying an observable phenomenon, we often have some prior hypothesis in mind, e.g., from previous studies, theory or feelings. If we measured the whole population in question, the result would either agree with our prior hypothesis or show it to be wrong. The outcome would be clear. But if our data were based only on a sample from that population, and the sample result differed from our prior hypothesis, we would have a problem. Was our prior hypothesis wrong? Or was the sample result differed only because of sampling error, i.e., was it an atypical sample? We could determine the answer taking a much larger sample. But with random sampling, we can avoid the extra work and cost by using a test of significance. This gives the probability that the difference between the sample value and the hypothesized value was only due to a sample error. If the probability is high we accept our prior hypothesis, if it is low we reject it.

The hypothesized value is usually called the null hypothesis. If the test shows the difference to be highly probable, the sample value is called "Statistically Significant," i.e., the difference is probably real, and the null hypothesis will be rejected in this case.

To illustrate, consider the following example. Acute myeloblastic leukemia is among the most deadly of cancers. It is claimed that the time X, in months that a patient survives after initial diagnosis is distributed normal with mean $\mu = 13$ months, and standard deviation $\sigma = 3$ months. A random sample of

16 patients is selected among those diagnosed with this disease at a teaching hospital in Ibadan and an average of 11 months was recorded for their survival. Does this indicate that the mean of survival is not achieved?

The fact that our sample estimate is less than 13 tells us nothing. What we would like to establish is whether the difference of an average of 2 months could reasonably be expected to arise, given that the $\mu = 13$ is correct. If it could, we would not be justified in claiming that the sample indicated a reduction in survival months. To evaluate the significance of this observed difference, we can calculate the probability that our sample result would be 2 months or more below the population figure of 13.

From Chap. 4, we know that

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Hence, to compute $P(\bar{x} < 11)$, we note here that the sampling distribution of $\bar{x}$ has, $\mu_{\bar{x}} = 13$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 3/4 = 0.75$. Thus the required probability is computed as:

$$P(\bar{x} < 11) = P(Z < \frac{11 - 13}{0.75} = P(Z < -2.67) = 0.0038$$

That is, only about 4 samples in every 1000 of similar size ($n = 16$) would give a mean survival time of 11 months or less. We are therefore faced with two alternative decisions:

(a) Either: A very rare event has occurred, and by chance we have got a sample which will only occur about 4 times in every 1000 such samples or
(b) This low sample has arisen because the population mean survival time is less than 13 months which we assumed it to be.

Which of these alternatives we choose depends on how much of a risk we are prepared to take. The first of these alternatives has a probability of 0.0038 of being correct. In rejecting it and accepting the second alternative, therefore, we are running a 0.38 % chance of being wrong. In the long run we would expect to be wrong about 4 times in every 1000 similar situations. It would be perfectly reasonable to discount such a risk.

Suppose, however, that our chance of being wrong had worked out to be 10 % instead of 0.38 %. Now acceptance of the second alternative means that in the long run we shall be wrong 10 times in every 100 similar situations. Is this a reasonable risk to run?

5.4.3 Example Seed

To illustrate, suppose a maize seed order firm stipulates that 80 % of seeds purchased must germinate in 3 days. It is supposed that this degree of germination rate is not being achieved, so a sample is taken to check the situation.

It is discovered that out of a random sample of 100 seeds, only 70 % germinated. Does this indicate that the required rate of germination is not being achieved?

The mere fact that our sample estimate is less than 80 % tells us nothing. What we want to establish is whether the difference of 10 % could reasonably be expected to arise, given that the 80 % is correct. If it could, we would not be justified in claiming that the sample indicated a deterioration in germination rate. To evaluate the significance of this observed difference, we can calculate the probability that our sample result would be 10 % or more below the population figure of 80 %. Here, we are using the fact that:

$$\hat{p} \sim b(p, \frac{pq}{n})$$

This will be discussed further in Sect. 5.7. Before we can calculate this probability, we need to know the parameters of the sampling distribution when $n = 100$ and $p = 80\,\%$. Remember that the mean $= p$ and the standard error (s.e. $= \sqrt{\frac{\hat{p}(100-\hat{p})}{n}}$. Since, we have $\hat{p} = 80\,\%$, hence the s.e. becomes $\sqrt{\frac{80\times 20}{100}} = 4$.

The Z point for the observed 70 % is, using the normal approximation given by

$$z = \frac{\hat{p} - p}{\text{s.e.}} = \frac{70 - 80}{4} = -2.5$$

$P(z < -2.5) = 0.0062$. That is, only about 6 samples in every 1000 of similar size ($n = 100$) would estimate it at 70 % or less. We are therefore faced with two alternative decisions:

(a) Either: A very rare event has occurred, and by chance we have got a sample which will only occur about 6 times in every 1000 such samples or
(b) This low sample has arisen because the population percentage is less than 80 % which we assumed it to be.

Which of these alternatives we choose depends on how much of a risk we are prepared to take. The first of these alternatives has a probability of 0.0062 of being correct. In rejecting it and accepting the second alternative, therefore, we are running a 0.62 % chance of being wrong. In the long run we would expect to be wrong about 6 times in every 1000 similar situations. It would be perfectly reasonable to discount such a risk.

Suppose, however, that our chance of being wrong had worked out to be 10 % instead of 0.62 %. Now acceptance of the second alternative means that in the long run we shall be wrong 10 times in every 100 similar situations. Is this a reasonable risk to run?

What is or is not a reasonable risk will depend on the consequence of being wrong. Through custom and practice, a number of standard risk levels, known as *significance levels*, have become established. By far the most commonly

used of these are the 5 % and 1 % or 0.05 and 0.01 significance levels. The process of calculating if the probability of making a wrong decision is greater or less than the required significance level is known as *significance testing* or simply as *test of significance*.

5.4.4 The Level of Significance

Different levels of significance involve different chances of making an error. Two kinds of error can be considered here. One is to accept the null hypothesis when it is actually false (a type II error). The other is to reject the null hypothesis when it is actually true (a type I error). The possibilities are set out below.

Decision	Null hypothesis is:	
	True	False
Fail to reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Thus an incorrect decision occurs if either a true null hypothesis is rejected or a false null hypothesis is not rejected. The former is called a *type I error* and the latter is similarly called the *type II error*.

We would however like the chances of correct decisions to be as high as possible. But reducing the likelihood of making a type II error (wrongly rejecting the null hypothesis) generally means increasing the chance of making a type I error (wrongly accepting the null hypothesis). One cannot have it both ways. Fortunately, we can reduce both types of risk by increasing the sample size. The more information you have in the sample, the greater will be the ability of the test statistic to reach the correct decision.

With a 5 % level of significance, we have a 5 % chance of committing a type I error; rejecting the null hypothesis even when it is true. With a 1 % significance level, we reduce the chances of committing such an error. But the chance of committing type II error is correspondingly increased.

So, on which side should the analyst err? With a new type of plane one would rather commit a type I error: Send the plane back for further tests even though it might in fact be air-safe. In legal cases we would rather commit a type II error: Let a guilty person off rather than convict an innocent man. In other cases such evaluations are more difficult to make. But in the case of statistical tests of significance the precise probability levels are usually not very important and most of our results are always clear-cut. Thus the general procedure for significant testing can be summarized as follows:

(i) A statistical hypothesis is set up, i.e., initial assumption, which is almost invariably an assumption about the value of a population parameter, e.g., in an example above $\mu = 13$. These hypotheses are usually referred to as null hypotheses, since they almost always state that no change has occurred from known or specified conditions.
(ii) An alternative hypothesis is defined which is to be accepted if the test permits us to reject the null hypothesis. In our example it is "μ is less than 11." "Steps (i) and (ii) must be carried out before the sample is analyzed."
(iii) An appropriate significance level is fixed, e.g., 5 % or 1 %.
(iv) On the assumption that the null hypothesis is true, the appropriate sampling distribution is defined and the area which will lead to rejection of the null hypothesis is identified. The probability that the sample result will fall into this rejection area is made equal to the specified significance level.
(v) The position of the sample result in the sampling distribution is calculated. If it falls in the rejection area, the null hypothesis is rejected and the alternative hypothesis is accepted. The result is then said to be *Statistically Significant.*

5.4.5 Types of Alternative Hypotheses

Consider the following situations for our example earlier where $\mu_0 = 13$ is specified. Generally therefore, the hypotheses can be formulated as:

(i)

$$H_0 : \mu \geq \mu_0$$
$$H_a : \mu < \mu_0 \tag{5.19}$$

(ii)

$$H_0 : \mu \leq \mu_0$$
$$H_a : \mu > \mu_0 \tag{5.20}$$

(iii)

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0 \tag{5.21}$$

The alternative hypothesis H_a in (i) above is often called a left-tailed single-sided alternative. The alternative in (ii) is similarly referred to as the right-tailed single alternative while that in (iii) is referred to as the two-sided alternative. The decisions that would be made depend on the type of alternative we have. It is therefore imperative that both the null and alternative hypotheses be set up prior to the commencement of conducting an experiment or study. We will now consider in turn the procedure for testing each of the alternatives in (i) to (iii) for a single mean or proportion.

5.5 Tests for Means and Proportion

The basic case of testing the difference between an observed sample mean $\bar{x}$ and a hypothesized population mean μ follows from the sampling distribution of the mean.

If we have a population X that is, $N(\mu, \sigma^2)$ then if a random sample size n is taken from this population, it follows from central limit theorem (CLT) that

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Case I: Population σ^2 Known

The hypothesis to be tested is of the form in (5.19), viz:

$$H_0 : \mu \geq \mu_0 \quad \text{a specified value versus}$$
$$H_a : \mu < \mu_0$$

If the hypothesis is correct and the sample size is greater than 30, then from CLT (irrespective of the population disposition or distribution)

$$\bar{x} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$$

Hence,

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

is distributed $N(0,1)$. That is, a standard normal with mean 0 and variance 1.

In order to test the above hypothesis, all we need is to compute the test statistic:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \tag{5.22}$$

and choose a significance level α. In the present case, the nature of our alternative hypothesis indicates that we use a left-tailed test. Hence if $\alpha = 0.05$ is chosen, then our decision rules are given by the following decision rule:

Decision Rule (DR) Reject H_0 if $Z \leq -z_\alpha$. In our case, since $\alpha = 0.05$, $z_{.05} = 1.645$, then our critical rejection region is $R : Z \leq -1.645$.

On the other hand, if the alternative is of type (5.20), then our decision rule would be:

Decision Rule Reject H_0 if $Z \geq z_\alpha$. In our case, since $\alpha = 0.05$, $z_{.05} = 1.645$, then our critical rejection region is $R : Z \geq 1.645$ since the hypotheses in this case are of the form:

$$H_0 : \ \mu \leq \mu_0 \quad \text{(a specified value)} \quad \text{versus}$$
$$H_a : \ \mu > \mu_0$$

If the hypothesis to be tested is of the form in (5.21), that is, a two-sided alternative, then our DR would be to reject H_0 if $\mid Z \mid \geq z_{\alpha/2}$. In our case, since $\alpha = 0.05$, hence, we need $z_{.025}$. The Table in appendix I gives $z_{.025} = 1.96$. Hence, our critical region (rejection region) is $R : \mid Z \mid \geq 1.96$.

5.5.1 p Values

With the advent of high-powered computing, most researches these days often quote p values rather than say that the test statistic is significant or not significant. The p value is the probability of getting a value as extreme or more extreme than the calculated test statistic if the null hypothesis were true. We give below, how to compute p values for each of the alternatives in (5.19)–(5.21).

$$\text{If } (\mathrm{H}_a : \mu < \mu_0) : \ p \text{ value} = P(z < Z^*)$$
$$\text{If } (\mathrm{H}_a : \mu > \mu_0) : \ p \text{ value} = P(z > Z^*)$$
$$\text{If } (\mathrm{H}_a : \mu \neq \mu_0) : \ p \text{ value} = 2\, P(z > |Z^*|)$$

where Z^* is the computed test statistics from (5.22). With p values, the decision rules are given with a specified α as,

$$\text{Reject } \mathrm{H}_0 \text{ if } \quad p \text{ value} \leq \alpha \tag{5.23}$$

Example 5.5.1

The mean and variance of the number of contaminants in a bacterial vaccine preserved with phenol are 60 and 25 respectively. A sample of 20 bacterial vaccines in a different preservative gave the contamination counts presented earlier in Table 5.1. Assuming that the counts are normally distributed, has the preservative significantly changed the contamination counts? Use $\alpha = 0.05$.

Here the population has $N(60, 25)$ with $\mu = 60$ and $\sigma^2 = 25$, and hence $\sigma = 5$. The hypothesis to be tested is

$$H_0 : \ \mu = 60$$
$$H_a : \ \mu \neq 60$$

From the data, $\bar{x} = 57.3$, $n = 20$, and $\sigma = 5$ are given. Hence we can compute the test statistic Z^* as:

$$Z^* = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} = \frac{\sqrt{20}(57.3 - 60)}{5} = -2.41$$

The rejection region is $|Z^*| \geq 1.96$. Since $\mid -2.41 \mid = 2.41 > 1.96$, i.e., we would reject H_0 and conclude that the preservative has significantly changed the number of counts at the 5 % significance level.

The corresponding p value is computed as:

$$p \text{ value} = 2P(Z > 2.41) = 2(1 - 0.9920) = 0.016$$

The decision rule is that we would reject H_0 if p value ≤ 0.05. Since, $0.016 << 0.05$, thus, we would strongly reject H_0. This result leads to the same conclusion we obtained using the other decision rule. In any case both should give the same conclusion. The MINITAB implementation of this is presented below:

```
Data Display

Counts
     67     62     52     55     54     61     51     59     54     57
     57     60     50     66     68     54     53     52     58     56

MTB > OneZ 'Counts';
SUBC>   Sigma 5;
SUBC>   Test 60.

One-Sample Z: Counts

Test of mu = 60 vs mu not = 60
The assumed sigma = 5

Variable          N      Mean     StDev    SE Mean
Counts           20     57.30      5.32       1.12

Variable             95.0% CI            Z       P
Counts        (   55.11,   59.49)    -2.41  0.016
```

Notice that the 95 % confidence interval does not include 60, hence H_0 can not be true. That is, we would reject H_0.

Case II: Population Variance Unknown

In this case, since the population variance is unknown, it means that we will have to estimate σ^2 from the sample data.
An unbiased estimate of the population variance σ^2 is:

$$s^2 = \frac{(\sum x - \bar{x})^2}{n-1} = \frac{\sum x^2 - n\bar{x}^2}{n-1}$$

We also estimate μ by $\bar{x} = \dfrac{\sum x_i}{n}$. From CLT, we know that $\bar{x} \sim N(\mu, \sigma^2/n)$. But here, we do not know σ^2, and hence we replace σ^2 in the expression for the test statistic by s^2. That is, the expression in (5.22) becomes

$$\frac{\sqrt{n}(\bar{x} - \mu_0)}{s}$$

but the expression is now no longer distributed as a standardized normal variate but as a Student's t distribution with $n-1$ degrees of freedom. Tables of Student's distribution are given in Table 2 of the appendix. That is, the test statistic is now

$$T = \frac{(\bar{x} - \mu_0)}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} \tag{5.24}$$

Example 5.5.2

The mean time for mice to die when injected with 1000 leukemia cells is known to be 12.5 days. When the injection was doubled in a sample of 10 mice, the survival times were

10.5, 11.2, 12.9, 12.7, 10.3, 10.4, 10.9, 11.3, 10.6, 11.7

If the survival times are normally distributed do the results suggest that the increased dosage has decreased survivorship?

The hypothesis is formulated below as

$$H_0 : \mu \geq 12.5$$
$$H_a : \mu < 12.5 \quad \text{(one-tailed test)}$$

Here, σ^2 of the population is not known, only the mean $\mu = 12.5$ is given. Hence we would have to estimate σ^2 from the data.

$$\bar{x} = \frac{\sum x}{n} = \frac{112.5}{10} = 11.25$$

$\sum x^2 = 1273.39$ and hence

$$s^2 = \frac{1273.39 - \frac{(112.5)^2}{10}}{10 - 1} = \frac{7.765}{9} = 0.8628$$

Hence $s = \sqrt{0.8628} = 0.9289$ and the test statistic becomes:

$$T = \frac{\sqrt{10}(11.25 - 12.5)}{0.9289} = -4.2554$$

From the Student's t distribution in Table 2 in the Appendix, we have $\nu = 9$ and if $\alpha = 0.05$, the tabulated t value equals 1.833. The decision rule for

the problem is: Reject H_0 if $T < -t_\alpha = -1.833$. Clearly $-4.25 << -1.833$, thus we would reject H_0, and the results suggest that a large injection dose reduces the life expectancy of mice. The MINITAB implementation for this example is displayed below.

```
Data Display

Time
   10.5   11.2   12.9   12.7   10.3   10.4   10.9   11.3   10.6   11.7

MTB > OneT 'Time';
SUBC>   Test 12.5;
SUBC>   Alternative -1.

One-Sample T: Time

Test of mu = 12.5 vs mu < 12.5

Variable          N      Mean     StDev   SE Mean
Time             10    11.250     0.929     0.294

Variable      95.0% Upper Bound        T      P
Time                     11.788   -4.26  0.001
```

In this example, the computed p value is $0.001 < 0.05$. Hence we would strongly reject H_0, again leading to the same conclusion as above. Had we not assumed that the sample was drawn from a normal population, a normality test would have been necessary. This test is presented in Fig. 5.2, and indicates that the p value for the Andersen–Darling test is 0.144 indicating that the data indeed could be assumed to have come from a normally distributed population.

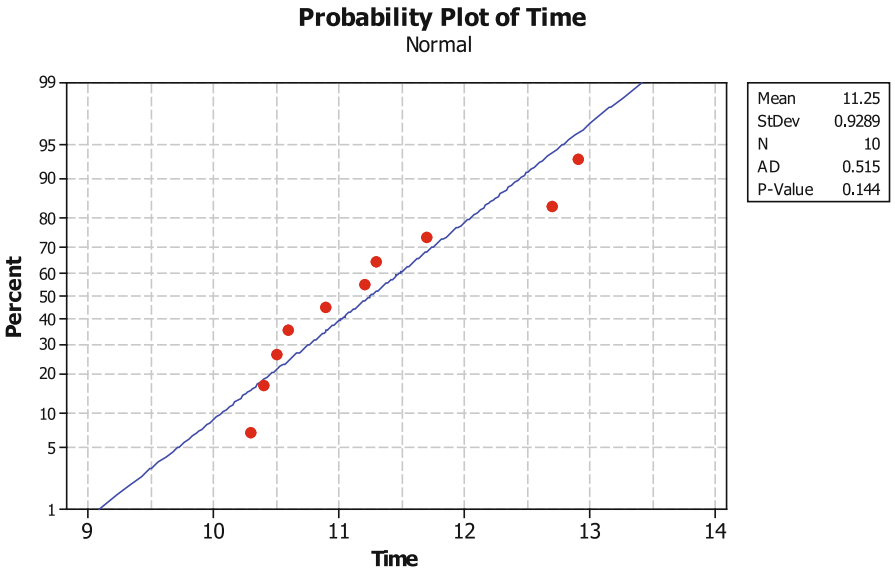


Fig. 5.2 Normal probability plot and test for this example

Example 5.5.3

Suppose in our Example 5.5.1, the variance of the population was not given. Then from the data, $n = 20$, $\sum x = 1146$, $\sum x^2 = 66,204$, hence,

$$s^2 = \frac{66,204 - \frac{(1146)^2}{20}}{20 - 1} = 28.33$$

and the test statistic is computed as:

$$T = \frac{\sqrt{20}(57.3 - 60)}{\sqrt{28.3}} = -2.27$$

Thus $\mid T \mid = 2.27$. The tabulated $t_{\alpha/2} = t_{.025}$ with 19 degrees of freedom is 2.093. Since $2.27 > 2.093$, we would still need to reject H_0 because 2.27 falls in the critical region and the same conclusion holds at $\alpha = 0.05$ level of significance. The MINITAB output in this case is presented below.

```
MTB > OneT  'Counts';
SUBC>   Test 60.

One-Sample T: Counts

Test of mu = 60 vs mu not = 60

Variable          N      Mean     StDev   SE Mean
Counts           20     57.30      5.32      1.19

Variable             95.0% CI            T      P
Counts        (   54.81,   59.79)    -2.27  0.035
```

Example 5.5.4

The number of wing beats per second of 16 male house flies were as follows:

194.7, 191.5, 187.0, 189.7, 190.0, 197.0, 189.9, 188.9

197.2, 191.4, 193.1, 186.9, 189.3, 185.2, 193.1, 196.6

If the mean number of wing beats per second of female flies is 190, do the wings of the males beat with a different frequency?

Let us assume that the number of beats per second follows a normal distribution, and the hypothesis of interest is

$$H_0 : \mu = 190$$

$$H_a : \mu \neq 190 \quad \text{(i.e., two-tailed test)}$$

From the data, $n = 16$, $\sum x = 3061.7$, $\sum x^2 = 586,079.41$, hence,

$$\bar{x} = \frac{3061.7}{16} = 191.36$$

$$s^2 = \frac{586,079.41 - \frac{(3061.7)^2}{16}}{16 - 1} = 13.5986$$

hence $s = \sqrt{13.5986} = 3.69$ and the calculated test statistic T is:

$$T = \frac{\sqrt{16}(191.36 - 190)}{3.69} = 1.47$$

From the t table, the tabulated t value with $\nu = 15$ degrees of freedom at $\alpha = 0.05$ is $t_{.025} = 2.131$.

Since $1.47 < 2.131$, i.e., we would fail to reject H_0 and conclude that there is no difference in wing speeds between male and female house flies at the 0.05 level of significance. The MINITAB implementation is presented below. The Anderson–Darling normality test gives a p value of 0.559 which is not less than 0.05. Hence, we can say that the sample was drawn from a normally distributed population.

```
Data Display

Beats
   194.7   191.5   187.0   189.7   190.0   197.0   189.9   188.9   197.2
   191.4   193.1   186.9   189.3   185.2   193.1   196.6

MTB > OneT  'Beats';
SUBC>   Test 190.

One-Sample T: Beats

Test of mu = 190 vs mu not = 190

Variable          N     Mean     StDev   SE Mean
Beats            16  191.344     3.692     0.923

Variable             95.0% CI            T      P
Beats         ( 189.376, 193.311)     1.46  0.166
```

5.5.2 Testing for a Binomial Proportion

Suppose we wish to test the hypothesis

$$H_0 : p = p_0$$
$$H_a : p \neq p_0$$

Let n (usually very large) denote the number of trials and let X denote the number of success in the n trials, then $\hat{p} = \dfrac{X}{n}$ is approximately normally distributed with mean and standard errors respectively,

$$\mu_{\hat{p}} = p, \quad \text{and} \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Hence under H_0:

$$\hat{p} \sim N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

Consequently, the relevant test statistic is:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{5.25}$$

and Z is distributed $N(0,1)$. Since the alternative hypothesis is two-sided, we would therefore reject H_0 at a significance level α if $\mid Z \mid \geq z_{\alpha/2}$.

An equivalent formulation of the test statistic Z is also given by:

$$Z = \frac{\frac{X}{n} - p_0}{\sqrt{\frac{p(1-p)}{n}}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \tag{5.26}$$

Example 5.5.5

In a certain cross of two varieties of peas, genetic theory led the investigator to expect one-half of the seeds produced to be wrinkled and the remaining one half to be smooth. In order to test this hypothesis, a final-year student conducted the experiment with 40 seeds of the cross and observed that 30 are wrinkled and 10 are smooth. Assuming $\alpha = 0.01$, is the genetic theory right?

Here the hypotheses are

$$H_0 : p = 0.5$$
$$H_a : p \neq 0.5$$

$n = 40$, $X = 30$ and $p_0 = 0.5$.

Thus,

$$Z = \frac{30 - 40 \times 0.5}{\sqrt{40 \times 0.5 \times 0.5}} = \frac{30 - 20}{\sqrt{10}} = 3.162$$

$z_{\alpha/2} = 2.58$ and since $3.162 > 2.58$, i.e., we would reject H_0 and conclude that the results of the experiment do not support the claim of the genetic theory. The corresponding p value is calculated as,

$$p \text{ value} = 2P(Z > 3.16) = 2(1 - 0.9992) = 0.0016$$

Since $0.0016 \lll 0.05$, we would again strongly reject H_0 in this example.

5.6 Tests Concerning Two Population Means

We have shown in an earlier section that given two populations X_1 and X_2 distributed normally with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively, then from (5.11)

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

and hence,

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{5.27}$$

will be distributed as a standardized normal variate. That is, $N(0,1)$.

5.6.1 Testing Differences Between Two Population Means

Case I: σ_1^2 and σ_2^2 Known

If the variances σ_1^2 and σ_2^2 of two independent populations are known and samples of sizes n_1 and n_2 are drawn respectively from these two populations, situations do arise when we are interested in testing whether significant differences do exist between the means of the two populations. The hypothesis of interest could be:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

Under H_0, the test statistic is given by:

$$\frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

But under $H_0 : \mu_1 = \mu_2$ implies that $\mu_1 - \mu_2 = 0$, thus the test statistic reduces to

$$Z = \frac{(\bar{x_1} - \bar{x_2})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{5.28}$$

The test is valid for all sample sizes if the probability distribution of x_1 and x_2 is normal.

Example 5.6.1

The data for this example were presented in Example 5.1.6 and relate to bacteria counts in the mouth of two groups of patients admitted to a hospital. The first group has 10 patients (and had just been admitted in the hospital) and the second group has 12 patients (who have spent 6 days in the hospital). The bacteria counts are known to have a population standard deviation of 2500 for each group of patients. Do the data for the two groups indicate that the mean bacterial count is influenced by a stay in hospital?

Let μ_1 and μ_2 be the means of the two groups. Then, the hypothesis of interest can be formulated as:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

$\bar{x}_1 = \frac{\sum x_1}{10} = 5627.3$. Similarly, $\bar{x}_2 = \frac{\sum x_2}{12} = 6937.9$. Hence, the statistic Z in (5.28) becomes

$$Z = \frac{(5627.3 - 6937.9}{\sqrt{\frac{2500^2}{10} + \frac{2500^2}{12}}} = -1.224$$

Since the test is a two-tailed one, $\mid Z \mid = 1.224$. At a 5 % significance level, the tabulated z value is 1.96 and since $1.224 < 1.96$, i.e., we would fail to reject H_0 and conclude that no evidence is found in support of a period in hospital influencing bacterial counts in the patients at the 0.05 level of significance. The corresponding p value is computed as $2P(Z > 1.22) = 2(1 - 0.8888) = 0.2224$. Since $0.2224 \not< 0.05$, we would therefore fail to reject H_0. This again leads to the same conclusion. If n_1 and n_2 are large (> 30) and σ_1 and σ_2 are unknown, we can use the large sample test statistic

$$\frac{(\bar{x_1} - \bar{x_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{5.29}$$

where s_1^2 and s_2^2 are the respective sample variances from the samples.

Case II: σ_1^2 and σ_2^2 Not Unknown

By far the most common case encountered in research studies is this case. Here two independent random samples of sizes n_1 and n_2 are each drawn from two independent normal populations with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2) respectively. The hypothesis of interest is again given by:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

However, in this situation σ_1^2 and σ_2^2 are unknown. This implies that they would have to be estimated from the samples.

In order to test these hypotheses, we would need to make the assumption that although σ_1^2 and σ_2^2 are not known, but they are being assumed to be equal, that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, a common unknown population variance. In this situation the population are said to be homogeneous.

Under these conditions, we have

$$\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right), \quad \bar{x}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right)$$

and therefore

$$(\bar{x}_1 - \bar{x}_2) \sim N\left((\mu_1 - \mu_2), \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

But then, σ^2 still remains unknown, and we can estimate it from each of the two samples, thus:

$$s_1^2 = \frac{(\sum x_{1i} - \bar{x}_1)^2}{n_1 - 1}, \quad s_2^2 = \frac{(\sum x_{2i} - \bar{x}_2)^2}{n_2 - 1}$$

These are two estimates of the same parameter. We can therefore pool these two estimates together to get a unified (pooled) estimate

$$S_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

S_P^2 is an unbiased estimate of σ^2 and the above test statistic becomes

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_P^2}{n_1} + \frac{S_P^2}{n_2}}}$$

However, this will now be distributed as Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom. Under H_0, the test statistic reduces to:

$$T = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{S_P^2}{n_1} + \frac{S_P^2}{n_2}}} \tag{5.30}$$

Example 5.6.2

The activities of an enzyme (units per gram protein) in 12 liver tissues infected with hepatitis and 18 normal liver tissues were presented in Example 5.2.1 earlier. Is there a significant difference is enzyme activity at $\alpha = 0.05$ level of significance?

Solution

Let μ_1 = mean of Hepatitis and μ_2 = mean of Normal liver patients.

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

We reproduce the calculations carried out earlier for this example below.

$$\sum x_1 = 49.06, \sum x_1^2 = 203.6432 \text{ and } n_1 = 12$$
$$\bar{x}_1 = 4.088, s_1^2 = \frac{\sum x_1^2 - \frac{(\sum x_1)^2}{n_1}}{n_1 - 1} = 0.27915$$

Similarly,

$$\sum x_2 = 68.23, \quad \sum x_2^2 = 266.0687, \quad \text{and } n_2 = 18$$
$$\bar{x}_2 = 3.791, \quad , s_2^2 = 0.4376$$

Hence, Pooled variance S_P^2 equals

$$S_P^2 = \frac{11 \times s_1^2 + 17 \times s_2^2}{28} = 0.3753 \tag{5.31}$$

and $S_P = \sqrt{S_P^2} = 0.6126$ and the test statistics is computed as:

$$T = \frac{4.088 - 3.791}{S\sqrt{\frac{1}{12} + \frac{1}{18}}} = 1.301$$

$t_{\alpha/2}(28) = 2.048$ and since $1.301 < 2.048$, i.e, we would fail to reject H_0 and conclude that there is no significant difference in enzyme activities in the two groups.

```
MTB > TwoSample 'HEPTIT' 'NORMAL';
SUBC>   Pooled.

Two-Sample T-Test and CI: HEPTIT, NORMAL


Two-sample T for HEPTIT vs NORMAL

         N      Mean     StDev   SE Mean
HEPTIT  12     4.088     0.528      0.15
NORMAL  18     3.791     0.662      0.16

Difference = mu HEPTIT - mu NORMAL
Estimate for difference:  0.298
95% CI for difference: (-0.170, 0.765)
T-Test of difference = 0 (vs not =): T-Value = 1.30  P-Value = 0.203  DF = 28
Both use Pooled StDev = 0.613
```

As discussed earlier, the above analysis is contingent upon the following assumptions being true.

1. That the populations in which the samples were drawn are homogeneous. That is, $\sigma_1^2 = \sigma_2^2$. In other words, that the population variances are equal.
2. That the samples were drawn from populations that are independently normally distributed.

A test of equality of variances of the form:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_a : \sigma_1^2 \neq \sigma_2^2$$

is provided by computing the statistic:

$$F* = \frac{s_1^2}{s_2^2} \tag{5.32}$$

where s_1^2 and s_2^2 are the sample variances and we usually allow the bigger of the two to be the numerator. Our decision rule is therefore to reject H_0 if $F* \geq F_{(n_1-1,n_2-1)}(1-\alpha)$. Here F is the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom at α level of significance. In our example,

$$F* = 0.4376/0.2791 = 1.57$$

Here sample 2 has the bigger variance and is used as the numerator. Hence, $F_{(18-1,12-1)} = F_{(17,11)}(0.95) = 2.68$. Since $1.57 < 2.68$, we would therefore fail to reject H_0. That is, $\sigma_1^2 = \sigma_2^2$ and this assumption is satisfied. The equality of variances test is implemented in MINITAB as follows:

```
MTB > %VarTest 'NORMAL' 'HEPTIT';
SUBC>   Unstacked.

Test for Equal Variances

Level1     NORMAL
Level2     HEPTIT
ConfLvl    95.0000

Bonferroni confidence intervals for standard deviations

  Lower     Sigma     Upper     N  Factor Levels

0.477372  0.661509  1.05658    18  NORMAL
0.357214  0.528253  0.97648    12  HEPTIT


F-Test (normal distribution)
Test Statistic: 1.568
P-Value       : 0.452


Levene's Test (any continuous distribution)
Test Statistic: 1.014
P-Value       : 0.323

Test for Equal Variances: NORMAL vs HEPTIT
```

The F test computed is 1.568 with a p value of 0.452, indicating again that the null hypothesis is plausible.

The test of normality is similarly carried out and the Anderson–Darling tests for both data give respectively, p value of 0.913 for the hepatitis group and p value of 0.386 for the normal group. In both cases, the p values are not less than 0.05, hence we will say that there are no evidence to suggest that the two samples were not drawn from two independent normally distributed populations. Hence our analysis for this example is valid as the two assumptions are satisfied.

Example 5.6.2

A feeding test is conducted on a herd of 25 milking cows to compare two diets, A and B. A sample of 12 cows randomly selected from the herd is fed diet B (dewatered alfalfa), the remaining 13 cows are fed diet A (field-wilted alfalfa). From observations made over a 3-week period the average daily milk production is given in Table 5.4:

Table 5.4 Data for this example

Diet A:	44, 44, 56, 46, 47, 38, 58, 49, 35, 46, 30, 53, 41
Diet B:	35, 47, 55, 29, 40, 39, 32, 41, 42, 57, 51, 39

Do the data strongly indicate that milk yield is less with diet B than with diet A (test at $\alpha = 0.05$)?

Here too there is no information on the variances of the populations from which the samples were drawn and the hypotheses are of the form:

$$H_0 : \mu_1 \leq \mu_2$$
$$H_a : \mu_1 > \mu_2$$

The following initial computations of means and sample variances yield the following for both diets.

$$\text{Diet A: } \bar{x}_1 = 45.15, \quad \sum (x_{1i} - \bar{x}_1)^2 = 767.69, \quad s_1^2 = 64.0$$
$$\text{Diet B: } \bar{x}_2 = 42.25, \quad \sum (x_{2i} - \bar{x}_2)^2 = 840.25, \quad s_2^2 = 76.4$$

The pooled variance S_P^2 equals,

$$S_P^2 = \frac{12s_1^2 + 11s_2^2}{23} = 69.9, \quad \text{and} \quad S = 8.36.$$

Hence, the test statistic T is computed as,

$$T = \frac{(45.15 - 42.25)}{8.36\sqrt{\frac{1}{13} + \frac{1}{12}}} = 0.87$$

The d.f. $= 12 + 13 - 2 = 23$ and $t_{.05}(23) = 1.714$ since this is a one-tailed test. But $0.87 < 1.714$; therefore, we would fail to reject H_0 and conclude

that there is no evidence that diet A significantly yields better than diet B. This test is conducted in MINITAB with the following:

```
Data Display

 Row   Diet A  Diet B

   1       44      35
   2       44      47
   3       56      55
   4       46      29
   5       47      40
   6       38      39
   7       58      32
   8       49      41
   9       35      42
  10       46      57
  11       30      51
  12       53      39
  13       41

MTB > TwoSample 'Diet A' 'Diet B';
SUBC>   Pooled;

SUBC>   Alternative 1.

Two-Sample T-Test and CI: Diet A, Diet B


Two-sample T for Diet A vs Diet B

         N      Mean     StDev   SE Mean
Diet A  13     45.15      8.00       2.2
Diet B  12     42.25      8.74       2.5

Difference = mu Diet A - mu Diet B
Estimate for difference:  2.90
95% lower bound for difference: -2.83
T-Test of difference = 0 (vs >): T-Value = 0.87  P-Value = 0.197 DF = 23
Both use Pooled StDev = 8.36
```

The p value for the test is $0.197 \not< 0.05$. Therefore, we would fail to reject H_0, which agrees with the earlier conclusion reached.

```
MTB > %VarTest 'Diet A' 'Diet B';
SUBC>   Unstacked.

Test for Equal Variances

Level1     Diet A
Level2     Diet B

F-Test (normal distribution)

Test Statistic: 0.838
P-Value       : 0.762

Test for Equal Variances: Diet A vs Diet B
```

A test of equality of variances of the two populations gives a p value of $0.762 \not< 0.05$. Hence we would fail to reject the null hypothesis that $H_0 : \sigma_1^2 = \sigma_2^2$. In other words, the two populations are homogeneous.

The second assumption that the sample be drawn from independent normal populations yield Anderson–Darling Normality test p values of 0.910 and 0.585 for diets A and B respectively. In both cases, the normality assumption will be satisfied as the Anderson–Darling p values indicate that there is no evidence to suggest that the samples could not have come from normally distributed populations. Hence our analysis in this example is valid.

5.7 The Mann–Whitney *U*-Test

For the two-sample pooled t-test discussed earlier, it was assumed that (i) the variances of the two populations are equal (that is, homogeneous) and that (ii) the two samples are drawn from normally distributed independent populations. In reality, these two assumptions could prove to be too restrictive. As an alternative, the Mann–Whitney U-test, which is a nonparametric test (does not assume any formal distribution for the populations) but however requires that the two populations have continuous type identical distributions. Consider the following example adapted from Samuels and Witmer (2006). The data involve the studying of breathing patterns in an experimental and control groups. The variable of interest is the total ventilation measurements (liters of air per minute per square meter of body area).

Experimental	Control
5.32	4.50
5.60	4.78
5.74	4.79
6.06	4.86
6.32	5.41
6.34	5.70
6.79	6.08
7.18	6.21

To employ the Mann–Whitney test, we first rank the data $n_1 + n_2 = 16$ data values in order of magnitude. These are presented below:

Experimental	Rank (R_1)	Control	Rank (R_2)
5.32	5	4.50	1
5.60	7	4.78	2
5.74	9	4.79	3
6.06	10	4.86	4
6.32	13	5.41	6
6.34	14	5.70	8
6.79	15	6.08	11
7.18	16	6.21	12
	89		47

The rank sums are:

$$R_1 = 5 + 7 + \cdots + 15 + 16 = 89$$
$$R_2 = 1 + 2 + \cdots + 11 + 12 = 47$$

A test statistic that is based on R_1 and R_2 is obtained by first calculating:

$$U_1 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$
$$U_2 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

The Mann–Whitney test employs either U_1 or U_2 but often, the smaller of the two, denoted by T is usually employed. Thus, the Mann–Whitney U-test for $(\bar{\mu}_1 - \bar{\mu}_2)$ for $n_1 \geq 10$ and $n_2 \geq 10$ is given by:

$$z = \frac{U_1 - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}} \tag{5.33}$$

The hypotheses

$$H_0 : (\bar{\mu}_1 - \bar{\mu}_2) = 0$$
$$H_a : (\bar{\mu}_1 - \bar{\mu}_2) < 0$$

are tested by rejection H_0 if $z < -Z_\alpha$. Similar rejection regions are equivalent to what we have discussed in previous sections for the other two alternative hypotheses (b) $H_a : (\bar{\mu}_1 - \bar{\mu}_2) > 0$ or (c) $H_a : (\bar{\mu}_1 - \bar{\mu}_2) \neq 0$. We present the analysis of the data using MINITAB.

```
MTB > Mann-Whitney 95.0 'Exptal' 'Control';
SUBC>   Alternative 0.

Mann-Whitney Test and CI: Exptal, Control

         N  Median
Exptal   8   6.190
Control  8   5.135


Point estimate for ETA1-ETA2 is 0.895
95.9 Percent CI for ETA1-ETA2 is (0.110,1.560)
W = 89.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0313
```

The alternative employed here is the two-tailed alternative (c). The p value for this test is 0.0313 which indicated that we would reject H_0 at the 5 % significance level. We conclude therefore that there is a significant difference in the median ventilation measurements of the two groups. We also note here

that the MINITAB automatically computes the 95 % confidence intervals. This can be changed to say, 90 % at will in MINITAB.

5.8 Comparison of Two Binomial Proportions

Suppose we have two binomial populations, A and B. If a random sample of size n_1 is taken from population A and the number of successes is denoted by X. Another independent random sample of size n_2 is taken from population B and the number of successes is denoted by Y. The relevant hypotheses can take one of the three forms I, II, and III as indicated in Table 5.5 of the form, for example:

Table 5.5 The three possible alternative hypotheses

I	II	III
$H_0 : p_1 \leq p_2$	$H_0 : p_1 \geq p_2$	$H_0 : p_1 = p_2$
$H_a : p_1 > p_2$	$H_0 : p_1 < p_2$	$H_0 : p_1 \neq p_2$
Right-tailed test	Left-tailed test	Two-tailed test

Under H_0, $p_1 = p_2 = p$?, where p denotes the unspecified equality population proportion. Then,

$$\hat{p}_1 = \frac{X}{n_1}, \quad E(\hat{p}_1) = p_1, \quad \text{Var}(\hat{p}_1) = \frac{p_1(1-p_1)}{n_1}$$

Similarly,

$$\hat{p}_2 = \frac{Y}{n_2}, \quad E(\hat{p}_1) = p_1, \quad \text{Var}(\hat{p}_2) = \frac{p_2(1-p_2)}{n_2}$$

and

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2, \quad \text{and} \quad \text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Thus under H_0, $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed with

$$E(\hat{p}_1 - \hat{p}_2) = 0$$

$$\text{Var}(\hat{p}_1 - \hat{p}_2) = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

since in this case $p_1 = p_2 = p$. Hence, a pooled estimate of p is given by:

$$\hat{p} = \frac{X+Y}{n_1+n_2} = \frac{n_1\,\hat{p}_1 + n_2\,\hat{p}_2}{n_1+n_2}$$

with estimated variance that equals

$$\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

and the test statistic is given by:

$$Z=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Example 5.7.1

A study is conducted to detect the effectiveness of mammographies. From 31 cases of breast cancer detected in women in the 40- to 49-year-old age group, 6 were found by the use of mammography alone. In older women, 38 of 101 cancers detected were found by mammography alone. Is this evidence at the $\alpha = 0.05$ level that the probability of detecting cancer by mammography alone is higher with older women than with younger?

Solution

Let the population of younger women be characterized by p_1 and the corresponding of older women by p_2. Then the hypotheses of interest here are type I, that is,

$$H_0:\ p_2 \le p_1$$
$$H_a:\ p_2 > p_1$$

From the younger women, we have $\hat{p}_1 = \frac{6}{31} = 0.194$. Similarly, from the older women, we also have $\hat{p}_2 = \frac{38}{101} = 0.376$. Thus here $X + Y = 6 + 38 = 44$ and $n = n_1 + n_2 = 31 + 101 = 132$. Hence, the pooled estimate for p is

$$\hat{p}=\frac{X+Y}{n}=\frac{44}{132}=0.333$$

Thus the test statistic is

$$\frac{\hat{p}_2-\hat{p}_1}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}=\frac{0.376-0.194}{\sqrt{0.333(1-0.333)\left(\frac{1}{31}+\frac{1}{101}\right)}}=1.88$$

The corresponding p value is computed as $P(Z > 1.88) = 0.0301$. Our decision rule rejects H_0 if p value is ≤ 0.05. Since $0.0301 < 0.05$, we would therefore reject H_0. Therefore, it is the case that the proportion of cancer detected by mammography alone is higher in older women than in younger women.

5.9 Paired *T*-test

The hypotheses involving two populations that we have discussed in the previous sections in this chapter assume that the populations of interest are independent. Consequently, we have assumed all along that the random samples from the two independent populations are themselves independent. Often however, situations do arise in which the random samples are not independent, where each observation in one sample is paired either by design or naturally (left hand versus right hand of the same person) with observation in the other sample. Examples are left and right gripping strengths of individuals, pair of twins studies in Medicine, pre- and posttest, same subjects observed before and after receiving a treatment, and litter mates of the same sex being assigned randomly to receive two different treatments, etc. Data arising from these kind of studies are often called matched-pair data and the corresponding hypothesis tests based on them have been described as *paired comparisons* tests.

Example 5.8.1

A study was conducted to investigate the effect of physical training on the triglyceride level. Eleven subjects participated in the study. Prior to training, blood samples were taken to determine the triglyceride level of each subject. Then the subjects were put through a training program that centered on daily running and jogging. At the end of the training period, blood samples were taken again and a second reading on the triglyceride level was obtained. The data obtained are presented in Table 5.6.

In this example, the two readings are not independent, being from the same subjects taken at different times and therefore are paired by subject. If we denote observations from the first sample by $x_i, i = 1, 2, \cdots, n$ and observations from the second sample similarly by $y_i, i = 1, 2, \cdots, n$ with corresponding population means μ_x and μ_y respectively. The possible hypotheses are presented in Table 5.7.

The hypotheses in Table 5.8 are equivalent to those presented in Table 5.7 respectively:

Since the two samples are not independent, we often have to base our analysis on the differences $d_i = x_i - y_i$ or could alternatively be defined as $d_i = y_i - x_i$. Either way, we just bear in mind that we need to formulate the alternative hypothesis correctly. Suppose in the above example, we wish to test the hypothesis that the training program is effective in increasing the mean level of triglyceride. Suppose we use the differences defined by $d_i = x_i - y_i$. Then,

$$\mu_d = \mu_x - \mu_y$$

Hence, our hypothesis of interest here is:

Table 5.6 Pre- and posttraining readings (in milligrams of triglyceride per 100 mL of blood

Subject	Pretraining	Posttraining
1	68	95
2	77	90
3	94	86
4	73	58
5	37	47
6	131	121
7	77	136
8	24	65
9	99	131
10	629	630
11	116	104

Table 5.7 Possible hypotheses of interest

I	II	III
$H_0 : \mu_x \geq \mu_y$	$H_0 : \mu_x \leq \mu_y$	$H_0 : \mu_x = \mu_y$
$H_a : \mu_x < \mu_y$	$H_a : \mu_x > \mu_y$	$H_a : \mu_x \neq \mu_y$

Table 5.8 Equivalent hypotheses to those in Table 5.7

I	II	III
$H_0 : \mu_x - \mu_y \geq 0$	$H_0 : \mu_x - \mu_y \leq 0$	$H_0 : \mu_x - \mu_y = 0$
$H_a : \mu_x - \mu_y < 0$	$H_a : \mu_x - \mu_y > 0$	$H_a : \mu_x - \mu_y \neq 0$

$$H_0 : \mu_x - \mu_y \geq 0 \quad \text{or} \quad H_0 : \mu_d \geq 0 \tag{5.34}$$

$$H_a : \mu_x - \mu_y < 0 \quad \text{or} \quad H_a : \mu_d < 0 \tag{5.35}$$

To test this hypothesis therefore, we obtain $\sum d_i$ and hence $\bar{d}$. We also obtain the standard deviation of d, and denote this as s_d. For these data, $\bar{d} = -12.55$ and $s_d = 24.47$. The test statistic for the above hypotheses is the one-sample t-test discussed earlier with $\mu_0 = 0$. That is,

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \tag{5.36}$$

For our data therefore, we have,

$$t = \frac{-12.55 - 0}{24.47/\sqrt{11}} = \frac{-12.55}{7.38} = -1.70$$

The decision rule rejects H_0 if $t \leq -t_\alpha(10 \text{ d.f.})$. That is if $-1.70 \leq -1.8125$. Since $-1.70 \not< -1.8125$, we would therefore fail to reject H_0 and conclude

that the data do not support the claim that the training program increases the mean level of triglyceride in subjects. This paired t-test is implemented in MINITAB with the following:

```
MTB > Let C3=C1-C2
Data Display

 Subjects    PRE   POST    d

    1         68     95    -27
    2         77     90    -13
    3         94     86      8
    4         73     58     15
    5         37     47    -10
    6        131    121     10
    7         77    136    -59
    8         24     65    -41
    9         99    131    -32
   10        629    630     -1
   11        116    104     12

MTB > OneT 'd';
SUBC>   Test 0;
SUBC>   Alternative -1.

One-Sample T: d

Test of mu = 0 vs mu < 0

Variable         N      Mean     StDev   SE Mean
d               11    -12.55     24.47      7.38

Variable      95.0% Upper Bound        T      P
d                          0.83    -1.70  0.060
```

Alternatively, we could use the paired t-test procedure in MINITAB to obtain a similar result.

```
MTB > Paired  'PRE' 'POST';
SUBC>   Alternative -1.

Paired T-Test and CI: PRE, POST

Paired T for PRE - POST

                 N      Mean     StDev   SE Mean
PRE             11     129.5     168.5      50.8
POST            11     142.1     164.4      49.6
Difference      11    -12.55     24.47      7.38

95% upper bound for mean difference: 0.83
T-Test of mean difference = 0 (vs < 0): T-Value = -1.70  P-Value = 0.060
```

A 95 % confidence interval for μ_d is computed as:

$$\begin{aligned}\bar{d} \pm t_{.025}(10 \text{ d.f.}) \left(\frac{s_d}{\sqrt{n}}\right) &= -12.55 \pm 2.2281(7.38) \\ &= -12.55 \pm 16.44 \\ &= (-28.99, 3.89)\end{aligned}$$

As observed before, the above one-sample test assumes normality for the differences d_i. A test of normality is presented in Fig. 5.3 and the Anderson–Darling test gives a p value of 0.429. Hence, our assumption of the normality of the differences d_i is satisfied and our analysis is valid.

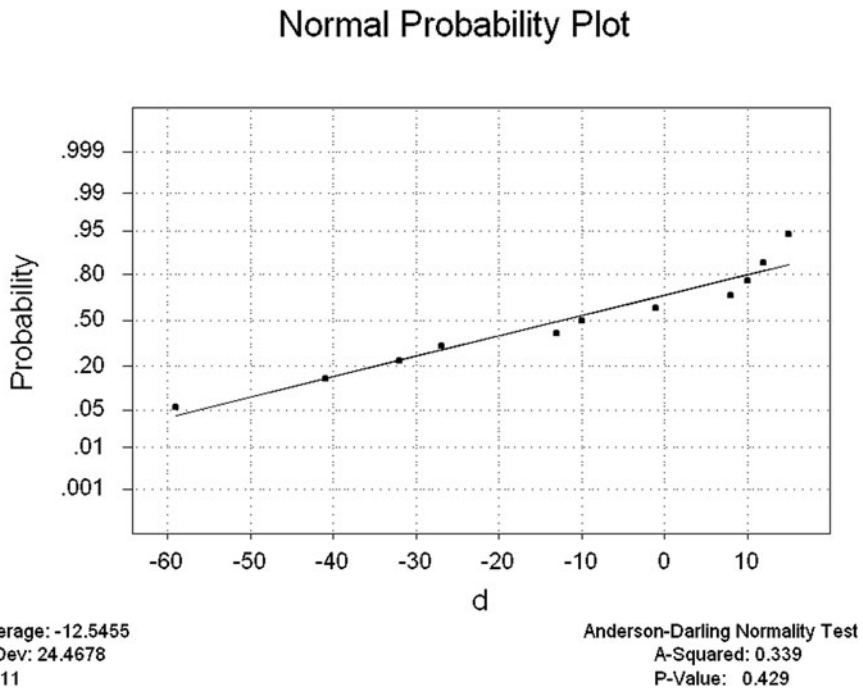


Fig. 5.3 Normality test for the differences d_i

5.10 Chapter Summary

We summarize in Table 5.9 the various formulae and underlying conditions necessary for the rightful application of the computation of confidence intervals and hypotheses testing in one and two samples.

Table 5.9 Summary of expressions for test statistics and confidence intervals

Situation	Parameter(s) of interest	Hypothesis test statistic	Confidence interval
(i)	Population mean μ	$Z = \dfrac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$	$\bar{x} \pm z_{\alpha/2}\left(\dfrac{\sigma}{\sqrt{n}}\right)$
(ii)	Population mean μ	$T = \dfrac{\sqrt{n}(\bar{x} - \mu_0)}{s}$	$\bar{x} \pm t_{\alpha/2}\left(\dfrac{\sigma}{\sqrt{n}}\right)$
(iii)	Population proportion p	$Z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$\hat{p} \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
(iv)	Poplns $(\mu_1 - \mu_2)$	$Z = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
(ivb)	$\vdots$	$Z = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$
(v)	Poplns $(\mu_1 - \mu_2)$	$T = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{S_P^2}{n_1} + \dfrac{S_P^2}{n_2}}}$	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\dfrac{S_P^2}{n_1} + \dfrac{S_P^2}{n_2}}$
(vi)	Poplns $(p_1 - p_2)$	$Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

where:

(i) refers to a situation where a random sample is drawn from a normal population with σ known and the $z_{\alpha/2}$ are the z value from a standard normal. Typical values are:

Area $1-\alpha$	$z_{\alpha/2}$
0.80	1.28
0.90	1.645
0.95	1.96
0.99	2.575

(ii) refers to the case of a random sample drawn from a normal population with σ unknown but will be estimated from available data and the $t_{\alpha/2}$ is the critical value for a t distribution with $n-1$ degrees of freedom which gives an area of $\alpha/2$ to the right of the distribution.

(iii) This refers to a large random sample from a population with a proportion p of a certain characteristic or attribute.

(iv) refers to two random samples drawn from two normally distributed populations with known variances (or standard deviations) σ_1^2 and σ_2^2 respectively.

(ivb) refers to case (iv) above but with variances unknown, but the samples sizes are large ($n_1, n_2 \geq 30$). In this case, s_1^2 and s_2^2 are respectively estimated from the samples.

(v) is the case to two random samples drawn from two independently normal populations with both σ_1^2 and σ_2^2 unknown but assumed equal to a common value (homogeneity) $S_P^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ and t is the Student's t distribution with n_1+n_2-2 degrees of freedom. Again, both s_1^2 and s_2^2 are respectively estimated from the samples.

(vi) This is the case of two large random samples drawn from two populations having p_1 and p_2 proportions of a certain attribute.

The corresponding standard errors for each of our statistics in Table 5.9 are the denominators in each of the expressions for the test statistics. That is:

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}}$$

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\text{s.e.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \quad \text{if } \sigma_1^2, \sigma_2^2, \text{ are known}$$

$$\text{s.e.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}, \quad \text{if } \sigma_1^2, \sigma_2^2, \text{ are unknown}$$

$$\text{s.e.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

5.11 Exercises

1. Find $z_{\alpha/2}$ for the following levels of α,

   (a) $\alpha = 0.04$ (b) $\alpha = 0.08$ (c) $\alpha = 0.06$ (d) $\alpha = 0.14$.

2. Find the $t_{\alpha/2}(n-1)$ for the following levels of α and n.

   (a) $\alpha = 0.05,\ n = 17$
   (b) $\alpha = 0.01,\ n = 10$
   (c) $\alpha = 0.10,\ n = 15$
   (d) $\alpha = 0.05,\ n = 9$

3. Find the $z_{\alpha/2}$ for the following confidence levels.

   (a) 96 %
   (b) 88 %
   (c) 94 %
   (d) 85 %
   (e) 92 %

4. A milkman surveyor wishes to know how many liters of milk people over 20 drink each week. Their consultant thinks that the mean is 6.2 L of milk per week with a standard deviation of 1.1. How large a sample would be required in order to estimate the average number of liters of milk per week consumed by people over 20 at the 90 % confidence level with an error of at most 0.07 L of milk?
5. A botanist observed 150 seedlings for the purpose of studying chlorophyll inheritance in corn. The seed came from self-fertilized heterozygous green plants. Hence green and yellow seedlings were expected in proportions of 3 green to 1 yellow. The sample showed 120 green and 30 yellow seedlings. Is this sample in agreement with expectations?
6. A metropolitan newspaper was considering a change to tabloid form. A random sample of 900 of its daily readers was polled to secure readership reaction for such a change. Of this sample, 541 persons opposed the change in format for the paper.

   (a) Is it likely that more than 50 % of the readers are in favor of the change?
   (b) Describe two or more procedures for obtaining confidence limits for the population proportion opposed to the change.

7. A farmer claims that the average yield of corn of variety A exceeds the average yield of variety B by at least 12 bushels/acre. To test this claim,

50 acres of each variety are planted and grown under similar conditions. Variety A yielded, on average, 86.7 bushels/acre with a standard deviation of 6.28 bushels/acre, while variety B yielded, on the average, 77.8 bushels/acre with a standard deviation of 5.61 bushels/acre. Test the farmer's claim using a 0.05 level of significance.

8. The following data represent the running times of films produced by two different motion-picture companies:

	Time in minutes						
Company 1	81	165	97	134	92	87	114
Company 2	102	86	98	109	92		

Test the hypothesis that the average running time of films produced by company 1 exceeds the average running time of films produced by company 2 by 10 min against the one-sided alternative that the difference is more than 10 min. Use a 0.1 level of significance and assume the distributions of times to be approximately normal.

9. For each of the following combinations of the p value and α, decide whether to accept or fail to reject the null hypothesis

   (a) p value $= 0.06$, $\alpha = 0.10$
   (b) p value $= 0.005$, $\alpha = 0.05$
   (c) p value $= 0.07$, $\alpha = 0.02$
   (d) p value $= 0.05$, $\alpha = 0.03$

10. Compute p values for the following tests and draw your conclusions at $\alpha = 0.01$

    (a) $H_0 : \mu = 20$ vs $H_1 : \mu > 20$, $z = 1.98$
    (b) $H_0 : \mu = 2.5$ vs $H_1 : \mu < 2.5$, $z = -1.36$
    (c) $H_0 : \mu = 50$ vs $H_1 : \mu \neq 50$, $z = 2.64$
    (d) $H_0 : \mu = 8.5$ vs $H_1 : \mu \neq 8.5$, $z = -1.80$

11. In clinical trials, there is what we call the placebo effect whereby patients often will report that they feel better even though the placebo contains nothing more than an harmless composition. Suppose that in a given study, 400 random subjects were given a placebo and the percentage of patients reporting improvement was 35 %. What would be a 90 % confidence interval for the true proportion of patients in the population who exhibit the placebo effect? Compute the same for 95 % and 99 % confidence intervals. Compare the three results.

12. For the following small sample tests for the population mean, compute p values for each test and draw your conclusions at $\alpha = 0.01$

    (a) $H_0 : \mu = 20$ vs $H_1 : \mu > 20$, $t = 2.7$, $n = 15$
    (b) $H_0 : \mu = 2.5$ vs $H_1 : \mu < 2.5$, $t = -2.75$, $n = 18$
    (c) $H_0 : \mu = 50$ vs $H_1 : \mu \neq 50$, $t = 2.4$, $n = 5$
    (d) $H_0 : \mu = 8.5$ vs $H_1 : \mu \neq 8.5$, $t = 2.0$, $n = 15$

13. Percentages of ideal body weight were determined for 18 randomly selected insulin-dependent diabetics and are shown below. A percentage of 120 means that an individual weighs 20 % more than his or her ideal weight; a percentage of 95 means that the individual weighs 5 % less than the ideal.

```
107   119    99   114   120   104   88   114   124
116   101   121   152   100   125   114   95   117
```

    (a) Estimate μ, σ^2 and σ.
    (b) Construct a 95 % confidence interval for the population mean, μ, the percentage of ideal body weight for this group.
    (c) Does this confidence interval contain the value 100 %? What does the answer to this question tell you?
    (d) What assumption must be satisfied for the above analysis to be valid? Conduct such a test.

14. The following data represent resting systolic blood pressure (SBP) of a group of children having one hypertensive parent (group 1) and a group of children both of whose parents have normal blood pressure (group 2).

```
GROUP 1:

             100  102  96  106  110
             110  120 112  112   90

GROUP 2:

            104  88  100  98  102
             92  96  100  96   96
```

    (a) Compute the mean SBP for groups 1 and 2 children.
    (b) Obtain a 95 % confidence interval for the difference, $\mu_1 - \mu_2$, between means of SBP for group 1 and group 2 respectively.
    (c) Use your result in (b) to test the hypotheses $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ at $\alpha = 0.05$
    (e) What assumptions are necessary for the above analysis? Test these assumptions from the information provided.

15. The mean carbon dioxide concentration in the air is 0.035 %. It is thought that the concentration immediately above the soil surface is higher than this.

    (a) Set up the null and alternative hypotheses required to gain statistical support for this contention.
    (b) Suppose 144 randomly selected air samples taken from within 1 foot of the soil were analyzed. A sample mean $\bar{x}$, of 0.09 % and sample standard deviation, s, of 0.25 % resulted. Calculate the necessary test statistic.

(c) Using p values, can H_0 be rejected at the $\alpha = 0.10$ level?
(d) Can H_0 be rejected at the $\alpha = 0.05$ level?
(e) Based on your decision in (d), do you think that the stated contention has been supported statistically? Use $\alpha = 0.05$.

16. A zoologist measured tail length in 86 individuals, all in the 1-year age group, of the deermouse *Peromyscus*. The mean length was 60.43 mm and the standard deviation was 3.06 mm. A 95 % confidence interval for the mean is (59.77, 61.09). True or false and say why:

(a) We are 95 % confident that the average tail length of the 86 individuals is between 59.77 and 61.09 mm.
(b) We are 95 % confident that the average tail length of all the individuals in the population is between 59.77 mm and 61.09 mm.

17. Acute myeloblastic leukemia is among the most deadly of cancers. Consider variable X, the time in months that a patient survives after the initial diagnosis of the disease. Assume that X is normally distributed with a standard deviation of 3 months. Studies indicate that $\mu = 13$ months. Consider the sample mean $\bar{X}$ based on a random sample of size 16. If the above information is correct, what are the numerical values of

(a) $E[\bar{X}]$
(b) Var $\{\bar{X}\}$
(c) the standard error of the mean?

18. The drug Anturane, marketed since 1959 for the treatment of gout, is being studied for use in preventing sudden deaths from a second heart attack among patients who have already suffered a first attack. In the study, 733 patients received Anturane and 742 were given a placebo. After 8 months, it was found that of out of 42 deaths from a second heart attack, 29 had occurred in the placebo group and 13 in the Anturane group.

(a) Use these data to estimate the difference in the proportion of sudden deaths among Anturane users and among patients not receiving the drug.
(b) Construct a 99 % confidence interval on the difference in the proportion of sudden deaths among Anturane users and among patients not on the drug.
(c) If 90 % confidence interval was constructed based on the same data, which interval would be longer? Why? Verify your answer.
(d) Have you gained evidence to support the statement that the death rate from second attacks is lower among patients on the drug than among patients not on the drug? Explain.

19. The following sample of 16 measurements was selected from a population with mean μ and standard deviation σ.

```
 91   80   99  110    95   106    78   121
106  100   97   82   100    83   115   104
```

(a) Construct an 80 % confidence interval for the population mean. You may assume that ($\sum x = 1567$, $\sum x^2 = 155867$). What assumptions must be satisfied?
(b) Interpret your result in (a) and explain what would happen to the confidence width if the confidence level is increased to 95 %?

20. A football league reported that the average number of touchdowns per game in 1988 was eight. The number of touchdowns per game for 37 randomly selected games played this year is:

```
5 8 6 4 8 5 4 4 4 6 5 6 5 8 6 4 4 5 8
8 7 6 4 4 8 6 5 6 6 4 8 5 4 4 5 7 5
```

Do these touchdown totals suggest that the average number of touchdowns per game, μ, for this year has decreased from the 1988 mean of eight? Formulate the null and the alternative hypotheses for this problem and test at the 10 % significance level. Assume $\sigma = 0.80$ and that the sum of all the data, $\sum x = 207$. Also compute the corresponding p value for this test and draw your conclusion.

21. In a recent poll, college students were asked if they supported the state lottery. The poll showed that 710 of the 1360 students surveyed did support the state lottery.

(a) Determine a 99 % confidence interval for the true proportion, π, of all college students who support the state lottery.
(b) Interpret your result in (a) and explain what would happen to the confidence width if the sample size were increased from 1360 to 2000? (no calculations needed!)
(c) What is the margin of error in (a)?

22. Using the MINITAB system, we get the following results:

```
TEST OF MU=155.000 VS MU > 155.000
THE ASSUMED SIGMA=25.9

        N     MEAN    STDEV    SE MEAN      Z   P VALUE
DAYS   40   166.73    25.89       4.09   2.86    0.0021
```

From the above printout, determine:

(a) the null and alternative hypotheses
(b) the smallest significance level at which the null hypothesis can be rejected
(c) Show how the P value was computed.
(d) How were 2.86 and 4.09 obtained?

23. In a 1993 study conducted by the American Management Association, 630 randomly selected major US firms were polled on their drug-testing policies. According to the report, " ... 85 % of the firms surveyed now test employees, applicants, or both." At the 5 % significance level, do the data provide sufficient evidence to conclude that the percentage of major US firms that drug-tested in 1993 exceeds the 1992 figure of 74 %. Also computer the corresponding p value for the test and once again draw your conclusions.
24. The mean carbon dioxide concentration in the air is 0.035 %. It is thought that the concentration immediately above the soil surface is higher than this.
    (a) Set up the null and alternative hypotheses required to gain statistical support for this contention.
    (b) Suppose 144 randomly selected air samples taken from within 1 foot of the soil were analyzed. A sample mean $\bar{x}$, of 0.09 % and sample standard deviation, s, of 0.25 % resulted. What is the p value for this test?
    (c) Based on your decision in (b), do you think that the stated contention has been supported statistically? Use $\alpha = 0.05$.
25. Among patients with lung cancer, usually 90 % or more die within 3 years. As a result of new forms of treatment, it is felt that this rate has been reduced. In a recent study of 150 patients diagnosed with lung cancer, 128 died within 3 years.
    (a) Calculate a point estimate of p, the true proportion of lung cancer patients who died within 3 years.
    (b) Set up the null and alternative hypotheses needed to support the above contention.
    (c) Can H_0 be rejected at the $\alpha = 0.10$ level?
    (d) Can H_0 be rejected at the $\alpha = 0.05$ level?
    (e) Do you think that there is sufficient evidence to claim that the new methods of treatment are more effective than the old? Explain.
26. An investigator randomly selected 36 nerve cells from a certain region of the brain of male guinea pigs. The counted number of dendritic branch segments emanating from each selected cell are as follows:

38	42	25	35	35	33	48	53	17
24	26	26	47	28	24	35	38	26
38	29	49	26	41	26	35	38	44
25	45	28	31	46	32	39	59	53

    The mean $\bar{x}$ for these counts is 35.67 and the sample standard deviation s is 9.99.

(a) Obtain the standard error for the mean $\bar{x}$.
(b) Construct a 95 % confidence interval for the population mean, μ, the number of dendritic segment counts that can emanate from the body of a male pig nerve cell.
(c) Test the hypothesis that:

$$H_0 : \mu \leq 30$$
$$H_a : \mu > 30$$

Use $\alpha = 0.05$. Obtain the p value for this test and draw your conclusions.

The number of wing beats per second of 16 male house flies were as follows:

```
194.7 191.5 187.0 189.7 190.0 189.9 188.9 197.0
197.2 191.4 193.1 186.9 189.3 185.2 193.1 196.6
```

Use MINITAB or R to analyze to answer the following questions:

(a) Estimate μ, σ^2 and σ.
(b) Construct a 95 % confidence interval for the population mean, μ, the number of beats of male flies.
(c) If the mean number of wing beats of female flies is 190, do the data above support the claim that the wings of the males beat with a different frequency? Explain your answer on the basis of the confidence interval obtained.
(d) What assumption must be satisfied for the above analysis to be valid? Conduct such a test.
(e) What is the standard error of the mean?

The activity of an enzyme (units per gram protein) in 12 liver tissues infected with hepatitis and 18 normal liver tissues were as follows:

```
HEPATITIS LIVER TISSUES:

            4.15 4.48 4.22 3.94 4.52 3.70
            4.77 4.03 3.90 4.86 3.16 3.33

NORMAL LIVER TISSUES:

            3.15 4.23 3.12 2.70 3.99 4.40
            3.86 3.86 3.16 4.27 4.34 3.79
            4.28 4.63 4.98 3.52 2.77 3.18
```

Use MINITAB or R to conduct a two-sample T test for the data.

(a) Test the hypotheses that:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

Use $\alpha = 0.05$.

(b) What assumptions are necessary for the above analysis? Conduct these tests.

In running a white cell count, a drop of blood is smeared thinly and evenly on a glass slide, stained with Wright's stain, and examined under a microscope. Of 200 white cells counted, 125 were neutrophils, a white cell produced in the bone marrow whose function, in part, is to take up infective agents in the blood.

(a) Find a point estimate for p, the proportion of neutrophils found among the white cells of this individual.
(b) Construct a 95 % confidence interval for p.
(c) In a normally healthy individual, the percentage of neutrophils among the white cells is 60–70 %. Based on the interval obtained in part (b), is there clear evidence of a neutrophil imbalance in this individual? Explain.
(d) How large a sample should be drawn if we let the margin of error d = 0.05, the confidence coefficient is 0.95, and no estimate of p is available from previous studies?

In a study of water usage in a small town, a random sample of 25 homes is obtained. The variable of interest is X, the number of gallons of water utilized per day. The following observations are obtained on a randomly selected weekday.

175	185	186	168	158
150	190	178	137	175
180	200	189	200	180
172	145	192	191	181
183	169	172	178	210

(a) Estimate μ, σ^2 and σ.
(b) Construct a 90 % confidence interval for μ.
(c) The reservoir is large enough to handle an average usage of 160 gallons/day. Does there appear to be a water shortage problem in the town? Explain your answer on the basis of the confidence interval obtained.
(d) What assumptions must be satisfied for the above analysis to be valid?

Using specimens obtained from 10 individuals, determinations of percent calcium content of sound teeth gave the following results:

36.39	36.19	34.20	35.15	35.47
35.22	36.11	35.63	36.63	35.59

If the mean percentage of calcium content of sound teeth in all individuals is 35, do the data above support the hypothesis that average calcium content in such individuals is different from 35?

(a) Test the above hypothesis at $\alpha = 0.05$ level of significance.
(b) What assumption must be satisfied for the above test to be valid? Conduct such a test.
(c) What is the value of the test statistic?
(d) How was the standard error for the mean obtained?
(e) Obtain a 95 % confidence interval for the population mean, μ, the average calcium content of sound teeth for all individuals.

An experimental flock of 200 chickens is inoculated with an organism suspected of causing a set of clinical signs observed in several commercial flocks in recent months.Within 14 days after inoculation, 137 of the birds have exhibited the characteristic signs. Suppose the chickens are housed and handled in such a way that cross infection is not a problem and that independence can be assumed. Find:

(a) A point estimate $\hat{p}$ of the proportion of inoculated chickens showing the signs within 14 days.
(b) Find $\sigma_{\hat{p}}$.
(c) Obtain a 92 % confidence interval for the population proportion of inoculated chickens showing signs within 14 days.
(d) Interpret your result.

Scientific method or scientific inquiry is the procedure whereby knowledge is acquired. Science is an attempt to extend our range of knowledge.

The evaluated knowledge is then used to formulate a *hypothesis* which is a tentative or a postulated explanation of a phenomenon.

Chapter 6
Analysis of Variance (ANOVA)

6.1 Introduction

In the last chapter, we were able to compare two population means using either the large sample Z test or the two-sample t test. Analysis of variance (that is, analysis based on the variation in the data) often written as "ANOVA" concerns how to test the means of more than two populations. Suppose we have k populations each distributed independently with mean μ_i and variance σ_i^2, $i = 1, 2, \cdots, k$. Then the hypotheses of interest here are:

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = \cdots = \mu_k \\ H_a &: \quad \text{at least two of these means are not equal} \end{aligned} \tag{6.1}$$

ANOVA therefore is a powerful tool for testing such hypotheses as in (6.1). It also allows us to test for interaction effects among factors as we will discuss in later chapters.

To carry out the above hypotheses in (6.1), suppose we take simple random samples of equal sizes (we will consider the case of unequal sample sizes later) r (equal replication) from each of these populations. Then we can compute $\bar{y}_i$ and s_i^2, the sample mean and sample variance for each sample, where:

$$\bar{y}_i = \frac{\sum y_{ij}}{r}, \quad s_i^2 = \sum y_{ij}^2 - r\bar{y}_i;\ i = 1, 2, \cdots, k;\ j = 1, 2, \cdots, r$$

and a typical layout for such a data is presented in Table 6.1.

Table 6.1 Table of observations for one-way ANOVA

Treatments	Observations				Total
1	y_{11}	y_{12}	$\cdots$	y_{1r}	Y_{1+}
2	y_{21}	y_{22}	$\cdots$	y_{2r}	Y_{2+}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kr}	Y_{k+}
Total	Y_{+1}	Y_{+2}	$\cdots$	Y_{+k}	Y_{++}

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_6,

6.1.1 Analysis of Variance of Table 6.1

There are $r \times k = rk$ observations in the data and if we let $G = \sum_{i=1}^{k} \sum_{j=1}^{r} y_{ij} = Y_{++}$, then,

$$\text{The correction factor (C.F.)} = \frac{Y_{++}^2}{rk} = \frac{G^2}{rk}.$$

The total sums of squares (Total SS) is computed as:

$$\text{Total SS} = y_{11}^2 + y_{12}^2 + \cdots + y_{kr}^2 - CF = \sum_i \sum_j y_{ij}^2 - \frac{Y_{++}^2}{rk} = TSS$$

and is based on $(rk - 1)$ degrees of freedom. That is, the total number of observations minus one.

The treatments sum of squares (TSS) is similarly computed as:

$$\text{TSS} = \frac{Y_{1+}^2}{r} + \frac{Y_{2+}^2}{r} + \cdots + \frac{Y_{k+}^2}{r} - CF = \sum_i \frac{Y_{i+}^2}{r} - \frac{Y_{++}^2}{rk}$$

and is also based on $(k-1)$ degrees of freedom. Again, this is obtained as the total number of treatments minus one. For brevity, if we denote the treatment totals as $T_1, T_2, \cdots, T_k$, then, the treatment SS (or between samples sum of squares) can be computed as:

$$\text{TSS} = \frac{T_1^2}{r} + \frac{T_2^2}{r} + \cdots + \frac{T_k^2}{r} - CF$$

The error sum of squares is obtained by subtraction as Total SS−Treatment SS, or as:

$$SSE = \sum_i \sum_j y_{ij}^2 - \frac{Y_{i+}^2}{r}$$

and is based on *Total d.f.*−*Treatment d.f.* $= rk-1-(k-1) = k(r-1)$ degrees of freedom.

We can put all these more succinctly in what is called the analysis of variance table which is displayed in Table 6.2.

Table 6.2 Analysis of variance table

Source of variation	d.f.	SS	MS	F
Treatments	$k-1$	TSS	$\frac{\text{TSS}}{k-1} = A$	$\frac{A}{S^2}$
Error	$k(r-1)$	SSE	$\frac{\text{SSE}}{k(r-1)} = S^2$	
Total	$rk-1$	Total SS		

Where:

(a) The MS = Mean Squares obtained by dividing the SS with their corresponding degrees of freedom. Thus for instance, the Treatments MS $= \frac{\text{TSS}}{k-1}$. Similarly for the error line, we have, $\frac{\text{SSE}}{k(r-1)} = \text{MSE}$ (MSE = mean square error). Please note that we did not have a mean square value for the Total SS line.
(b) The error mean square (EMS) is a pooled variance estimate of the population variance σ^2 (unknown) and is often denoted by S^2.
(c) The F values (F ratios) are computed as the ratio of (Treatments MS)/(EMS) and are distributed as the F distribution with $(k-1)$ and $(k(r-1))$ degrees of freedom.
(d) We note here that:
- Total SS = TSS + Error SS
- Total d.f. = Treatment d.f. + Error d.f.

The appropriate test procedure for the hypotheses in (6.1) is derived from the analysis of variance table in Table 6.2. The value under the F-column $\frac{A}{S^2}$ is, when H_0 is true, distributed as F distribution with $k-1$ and $k(r-1)$ degrees of freedom. This value can be compared with the tabulated F value with the corresponding pairs of degrees of freedom at a specified α level (Table 4 in the Appendix). The one-way analysis of variance (it is one way because, there is only one partition above the Total SS line in addition to the Error SS line) validity is contingent on the following assumptions being true.

6.1.2 Assumptions of the One-Way ANOVA Model

The following are some of the assumptions underlying the one-way analysis of variance.

(i) The populations from which the samples are drawn are assumed to be approximately normally distributed and independent.
(ii) The populations are homogeneous. That is, we will assume that the population variances are equal, viz.:

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2.$$

For a given data set, these assumptions must therefore be tested to ensure the validity of our analysis. We give below an example of one-factor ANOVA analysis and the subsequent tests for the validity of the above assumptions.

6.1.3 An Example

Carbon dioxide is known to have a critical effect on microbiological growth. Small amounts of CO_2 stimulate the growth of many organisms, while high concentrations inhibit the growth of most. The latter effect is used commercially when perishable food products are stored. A study is conducted to investigate the effect of CO_2 on the growth rate of *Pseudomonas fragi*, a food spoiler. Carbon dioxide is administered at five different atmospheric pressures. The response noted was the percentage change in cell mass after 1-h growing time. Ten cultures are used at each level. The following are the data so obtained (Table 6.3).

Table 6.3 Factor (CO_2 pressure in atmospheres) level

	Atmospheric pressure levels				
Sample	0.0	0.083	0.29	0.50	0.86
1	62.6	50.9	45.5	29.5	24.9
2	59.6	44.3	41.1	22.8	17.2
3	64.5	47.5	29.8	19.2	7.8
4	59.3	49.5	38.3	20.6	10.5
5	58.6	48.5	40.2	29.2	17.8
6	64.6	50.4	38.5	24.1	22.1
7	50.9	35.2	30.2	22.6	22.6
8	56.2	49.9	27.0	32.7	16.8
9	52.3	42.6	40.0	24.4	15.9
10	62.8	41.6	33.9	29.6	8.8
$\sum y$	591.4	460.4	364.5	254.7	164.4
$\bar{y}$	59.14	46.04	36.45	25.47	16.44

For the above data, $k = 5$ and $r = 10$. Hence, there are $rk = 50$ observations in the data. The Total SS is computed as:

$$\text{Total SS} = 62.6^2 + 50.9^2 + \cdots + 29.6^2 + 8.8^2 - \frac{(1835.4)^2}{50} = 12,522.3568$$

where $\text{G} = \sum y_{ij} = 62.6 + 50.0 + \cdots + 29.6 + 8.8 = 1835.4$.

Similarly, the Treatments SS, TSS, is computed as:

$$\begin{aligned}\text{TSS} &= \frac{591.4^2}{10} + \frac{460.4^2}{10} + \frac{364.5^2}{10} + \frac{254.7^2}{10} + \frac{164.4^2}{10} - \frac{1835.4^2}{500} \\ &= 11,274.3188\end{aligned}$$

Hence, the Error SS is computed by subtraction as:

$$\text{Error SS} = \text{Total SS} - \text{TSS} = 12,522.3568 - 11,274.3188 = 1248.0380.$$

The resulting ANOVA table is presented in Table 6.4.

Table 6.4 Analysis of variance table for the example

Source of variation	d.f.	SS	MS	F
Treatments	4	11,274.3188	2818.5797	101.63
Error	45	1248.0380	$27.7342 = S^2$	
Total	49	12,522.3568		

From Table 4 in the appendix, $F(4, 45, \alpha = 0.05) = 2.61$. Since the calculated F-value from our ANOVA table is $101.63 >> 2.61$, we would therefore strongly reject the null hypothesis. Consequently, we can at this point, conclude that there are significant differences in the means of the atmospheric levels at 0.05 level of significance.

We can implement the above analysis in MINITAB in two ways. Here, we have chosen to read the data in unstacked way (that is by levels). The ANOVA table for this analysis is also provided by MINITAB with results presented to only one decimal point. The results are consistent with what we obtained by hand calculations. The p value for the test of the hypotheses is $0.0000 << 0.05$. Since we reject when p value < 0.05, i.e., we will strongly reject H_0.

```
MTB > read c1-c5
DATA> 62.6 50.9 45.5 29.5 24.9
DATA>  59.6 44.3 41.1 22.8 17.2
DATA>  64.5 47.5 29.8 19.2 7.8
DATA>  59.3 49.5 38.3 20.6 10.5
DATA>  58.6 48.5 40.2 29.2 17.8
DATA>  64.6 50.4 38.5 24.1 22.1
DATA>  50.9 35.2 30.2 22.6 22.6
DATA>  56.2 49.9 27.0 32.7 16.8
DATA>  52.3 42.6 40.0 24.4 15.9
DATA>  62.8 41.6 33.9 29.6 8.8
DATA> end
10 rows read.
MTB > AOVOneway 'P1' 'P2' 'P3' 'P4' 'P5';
SUBC>   Tukey 5;
SUBC>   GBoxplot;
SUBC>   GNormalplot;
SUBC>   NoDGraphs.

One-way ANOVA: P1, P2, P3, P4, P5

Source  DF       SS      MS       F      P
Factor   4  11274.3  2818.6  101.63  0.000
Error   45   1248.0    27.7
Total   49  12522.4

S = 5.266   R-Sq = 90.03%   R-Sq(adj) = 89.15%


                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level   N    Mean  StDev  -+---------+---------+---------+--------
P1     10  59.140  4.805                                (-*--)
P2     10  46.040  5.053                         (--*-)
P3     10  36.450  5.934                  (-*--)
P4     10  25.470  4.483           (-*-)
P5     10  16.440  5.894  (-*-)
                          -+---------+---------+---------+--------
                          15         30         45         60
```

```
Pooled StDev = 5.266


Bartlett's Test (Normal Distribution)
Test statistic = 1.07, p-value = 0.899

Levene's Test (Any Continuous Distribution)
Test statistic = 0.19, p-value = 0.941
```

We notice that there are significant differences between the five levels. At this point, we can not fully determine which level is necessarily the best until we consider multiple comparison procedure at a later chapter, however, we see from the generated means from MINITAB above that there are differences between the means which range from 16.44 to 59.140. The box plot of these means is presented in Fig. 6.1. It is clear that level 5 seems to be the best but whether it is better than level 4 or 3 at this point we do not know. We shall investigate this further later. However, we can demonstrate at this point, the analysis of variance (ANOVA) procedure is an extension of the two-sample t test that we considered in the previous chapter. We demonstrate this by computing the pooled estimate of the unknown population variances (that are assumed equal).

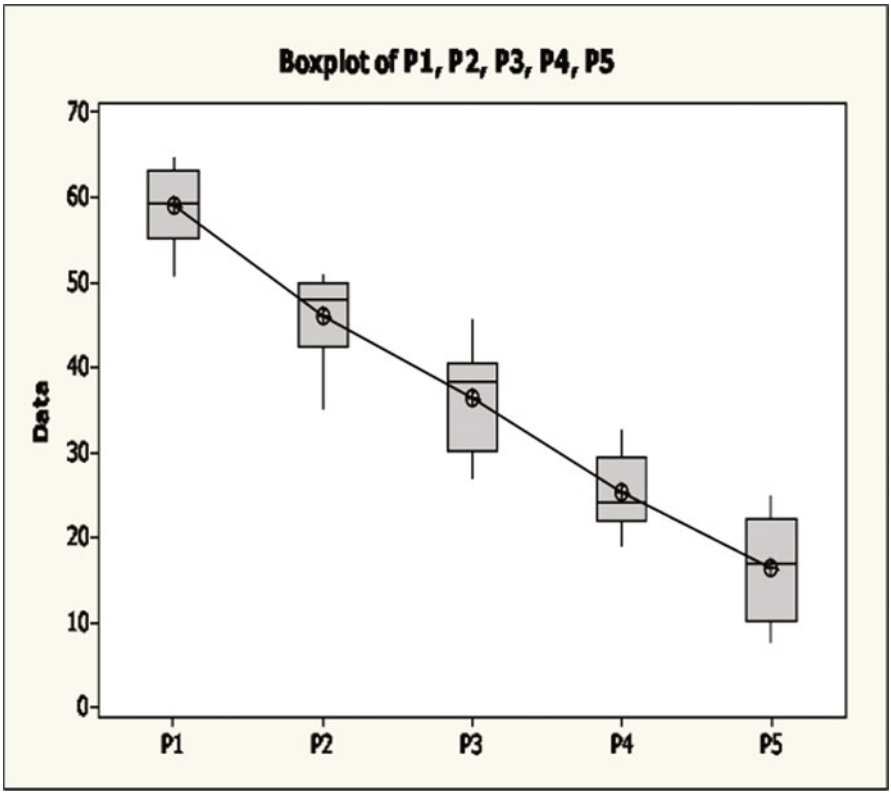


Fig. 6.1 Box plots of the pressure level means

6.1.4 Extending the Two-Sample t Test

For each of our samples, we noted that we would obtain $\bar{y}_i$ and s_i^2 for $i = 1, 2, \cdots, k$. Thus, under the assumption that the populations are homogeneous, then, an estimate of the pooled estimate of the common variance σ^2 is obtained as:

$$S_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1)} \tag{6.2}$$

If $n_i = n$, $\forall\, i$, (for all i) then, the above reduces to:

$$S_P^2 = \frac{(n-1)s_1^2 + (n-1)s_2^2 + \cdots + (n-1)s_k^2}{(n-1) + (n-1) + \cdots + (n-1)} = \frac{1}{k}\sum_{i=1}^{k} s_i^2$$

For our data in the above example, $n = 10 = r$ and $k = 5$. Here, $s_1^2 = 23.0849$, $s_2^2 = 25.5293$, $s_3^2 = 35.2117$, $s_4^2 = 20.1001$, and $s_5^2 = 34.7449$. Thus,

$$S_P^2 = \frac{23.0849 + 25.5293 + \cdots + 34.7449}{5} = \frac{138.6709}{5} = 27.7342$$

We notice that this computed value of $S_P^2 = 27.7342$ is exactly what we obtained as the Error mean square (EMS) under the analysis of variance in Table 6.4. We therefore see that the analysis is a general extension of the two-sample t test to k means.

Without further knowledge about the five levels, the making of particular comparisons between pairs of levels is rather dangerous. However, it is clear that level 5 is the best atmospheric pressure as it has the lowest percentage change.

6.2 Multiple Comparisons Procedures

We have seen that although the F test above indicates that there are significant differences among the level (treatment) means, however, this does not tell or indicate which specific means are significantly different. To do this, we have to conduct paired comparisons. In this example, since we have five population means, we would therefore need to make $\binom{5}{2} = 10$ such paired comparisons, leading to what is often known as *multiple comparison tests* because we are making more comparisons than allowed by the treatment degrees of freedom (Here, $10 > 4$). We consider some of the procedures for conducting multiple comparisons in what follows.

6.2.1 Fisher's Least Significant Difference (LSD)

For any two means, say, μ_i and μ_j, with $i \neq j$, the hypothesis of interest is

$$H_0 : \mu_i = \mu_j$$
$$H_a : \mu_i \neq \mu_j, \quad i \neq j.$$

To conduct the test, the standard error for the difference between the two sample means is $\sqrt{\frac{S^2}{r_1} + \frac{S^2}{r_2}}$. In our example, this equals $\sqrt{\frac{2S^2}{10}}$ since $r_1 = r_2 = 10$. The least significance difference is

$$\text{LSD} = \sqrt{\frac{2S^2}{10}} \times t_{0.025,45} = 2.3552 \times 2.0141 = 4.7436,$$

where $t_{0.025,45} = 2.0141$ is the Students' t-distribution percentile based on the error degrees of freedom of 45 at the $\alpha = 0.05$ level of significance. In other words, the two means μ_i and μ_j will be significantly different at $\alpha = 0.05$ level of significance if

$$|\bar{y}_{i.} - \bar{y}_{j.}| \geq 4.7436, \quad \text{for } i \neq j. \tag{6.3}$$

The implementation of this in MINITAB is presented in a summarized result below and is accomplished in MINITAB by specifying after the model statement *means level/lsd*, where *level* is the factor name. The result is given in the previous section.

Alternatively, we could use confidence interval approach to implement the same. Here, 95 % confidence intervals for the ten pairs of comparisons can be obtained as follows:

$$(\bar{y}_{i.} - \bar{y}_{j.}) \pm t_{0.025,45} \times \sqrt{\frac{2S^2}{10}} = (\bar{y}_{i.} - \bar{y}_{j.}) \pm 4.7436.$$

For levels 1 and 2 for instance, this becomes

$$(59.14 - 46.04) \pm 4.7436 = 13.10 \pm 4.7436 = (8.3564, 17.8436).$$

Since this interval does not include zero, therefore, we can conclude that there is a significant difference between level means 1 and 2. The procedure is then repeated for the other nine paired comparisons. This can readily be accomplished in MINITAB by specifying after the model statement, the following: MINITAB will conduct both Tukey and LSD tests. The LSD results are presented below.

```
Grouping Information Using Fisher Method

     N    Mean  Grouping
P1  10  59.140  A
P2  10  46.040    B
P3  10  36.450      C
P4  10  25.470        D
P5  10  16.440          E

Means that do not share a letter are significantly different.


Fisher 95% Individual Confidence Intervals
All Pairwise Comparisons

Simultaneous confidence level = 72.40%


P1 subtracted from:

      Lower   Center    Upper  ---------+---------+---------+---------+
P2  -17.844  -13.100   -8.356             (-*-)
P3  -27.434  -22.690  -17.946          (-*-)
P4  -38.414  -33.670  -28.926     (-*)
P5  -47.444  -42.700  -37.956  (-*-)
                              ---------+---------+---------+---------+
                                     -2          0         25         50


P2 subtracted from:

      Lower   Center    Upper  ---------+---------+---------+---------+
P3  -14.334   -9.590   -4.846             (-*-)
P4  -25.314  -20.570  -15.826          (-*-)
P5  -34.344  -29.600  -24.856       (-*-)
                              ---------+---------+---------+---------+
                                     -25         0         25         50


P3 subtracted from:

      Lower   Center    Upper  ---------+---------+---------+---------+
P4  -15.724  -10.980   -6.236             (-*-)
P5  -24.754  -20.010  -15.266          (-*-)
                              ---------+---------+---------+---------+
                                     -25         0         25         50


P4 subtracted from:

      Lower  Center   Upper  ---------+---------+---------+---------+
P5  -13.774  -9.030  -4.286               (-*-)
                            ---------+---------+---------+---------+
                                   -25         0         25         50
```

Other statistical softwares, such as SAS, often warns us with *NOTE: This test controls the Type I comparison-wise error rate, not the experiment-wise error rate.* It is therefore important that we discuss briefly what is meant by *experiment-wise error rate.*

6.2.2 Experiment-Wise Error Rate (EER)

With several comparisons on the means, the experiment-wise error rate (EER) is the probability that one or more of the comparison tests result in a Type I error (that is, the probability of rejecting at least one correct null hypothesis under several other competing null hypotheses—which may be true or false). If the comparisons are independent, then the experiment-wise error rate is

$$\alpha^* = 1 - (1 - \alpha)^h,$$

where α^* is the experiment-wise error rate, α is the per-comparison error rate or specified level of significance, and h is the total number of comparisons. In our example for instance, where there are 10 independent comparisons to be made at the 0.05 level of significance each, then the probability that at least one of them would result in a Type I error is

$$1 - (1 - 0.05)^{10} = 0.4013.$$

Clearly, a Type I error rate of 0.4013 is unacceptable. However, if the comparisons are not independent then the experiment-wise error rate is less than $1 - (1 - \alpha)^h$ and regardless of whether the comparisons are independent, $\alpha^* \leq h\,\alpha$. In our example for instance, $0.4013 < 10(0.05) = 0.50$.

Although Fisher's LSD does not control the experiment-wise error rate, the results obtained indicate that there are significant differences in the ten pairs. Thus level 5 gives the lowest mean, while level 1 gives the highest rate of change of bacteria. Since lowest is best here, we would recommend level 5 as the best.

We now consider multiple comparisons procedures that endeavor to control the experiment-wise error rates in the following sections.

6.2.3 The Tukey Test

The Tukey test procedure uses the critical value $q_\alpha(k, \nu)$ which is obtained from tables of significant studentized ranges (two-tailed Table 6 in the Appendix) having k treatments, in the expression in (6.4). Here, α is the upper tail of the q distribution and ν is the number of degrees of freedom on which the mean square error (MSE) is based and b_i is the number of observations on which the means are based. That is, we need to obtain

$$\text{Significance difference (SD)} = q_\alpha(k, \nu) \times \sqrt{\frac{MSE}{b_i}}. \tag{6.4}$$

In our example, $b_i = 10$, and if $\alpha = 0.05$ and $\sqrt{MSE/10} = \sqrt{27.7342/10} = 1.6654$, hence $q_{0.05}(5, 45) = 4.025$ where $k = 5$ is the number of means to be

compared. Now SD $= 1.6654 \times 4.025 = 6.7032$, approximately. When testing differences between the various means, if the difference between any two means is larger than the SD $= 6.7032$, then the means are assumed to be significantly different. That is, if

$$|\bar{y}_{i.} - \bar{y}_{j.}| \geq 6.7032, \quad \text{for } i \neq j. \tag{6.5}$$

The implementation of this in MINITAB is presented in a summarized result below and is accomplished in MINITAB by specifying after the model statement *means level/tukey lines*.

In general, to conduct Tukey's test, we would do the following:

(a) Calculate the SD for a specified α level as in (6.4).
(b) Rank the treatment means from smallest to largest.
(c) For those treatment means not indicating significance, place a bar under those pairs. Any pair not connected by an underbar implies significant difference in the population means.

We may note that Tukey's procedure ensures that all comparisons are made at the specified α significance value. For our example, the results indicate significance differences between all means.

Results of Tukey's test

$\bar{y}_{5.}$	$\bar{y}_{4.}$	$\bar{y}_{3.}$	$\bar{y}_{2.}$	$\bar{y}_{1.}$
16.44	25.47	36.45	46.04	59.14

We present below the MINITAB output for conducting Tukey's test on the data in Table 6.3.

```
                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level   N    Mean  StDev    -+---------+---------+---------+--------
P1     10  59.140  4.805                                    (-*--)
P2     10  46.040  5.053                           (--*-)
P3     10  36.450  5.934                    (-*--)
P4     10  25.470  4.483           (-*-)
P5     10  16.440  5.894    (-*-)
                            -+---------+---------+---------+--------
                            15          30          45          60

Pooled StDev = 5.266


MTB > AOVOneway 'P1' 'P2' 'P3' 'P4' 'P5';
SUBC>   Tukey 5;
SUBC>   Fisher 5.
```

```
Grouping Information Using Tukey Method

     N    Mean  Grouping
P1  10  59.140  A
P2  10  46.040    B
P3  10  36.450      C
P4  10  25.470        D
P5  10  16.440          E

Means that do not share a letter are significantly different.


Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons

Individual confidence level = 99.33%


P1 subtracted from:

      Lower   Center    Upper    +---------+---------+---------+---------
P2  -19.795  -13.100   -6.405                (--*-)
P3  -29.385  -22.690  -15.995           (--*--)
P4  -40.365  -33.670  -26.975        (--*-)
P5  -49.395  -42.700  -36.005    (--*--)
                                  +---------+---------+---------+---------
                               -50       -25         0        25


P2 subtracted from:

      Lower   Center    Upper    +---------+---------+---------+---------
P3  -16.285   -9.590   -2.895               (--*--)
P4  -27.265  -20.570  -13.875            (--*-)
P5  -36.295  -29.600  -22.905         (--*--)
                                  +---------+---------+---------+---------
                               -50       -25         0        25


P3 subtracted from:

      Lower   Center    Upper    +---------+---------+---------+---------
P4  -17.675  -10.980   -4.285               (--*-)
P5  -26.705  -20.010  -13.315            (--*--)
                                  +---------+---------+---------+---------
                               -50       -25         0        25


P4 subtracted from:

      Lower  Center   Upper    +---------+---------+---------+---------
P5  -15.725  -9.030  -2.335                (-*--)
                              +---------+---------+---------+---------
                           -50       -25         0        25
```

6.3 Contrasts

We will discuss other multiple comparison procedures in Chap. 10. However here, we will define *contrasts* and the concept of orthogonal contrasts.

Definition of Contrast

For k treatments having population means $\mu_1, \mu_2, \ldots, \mu_k$, a contrast L is a linear combination of the k treatment means, that is

$$L = c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k = \sum_{i=1}^{k} c_i\mu_i, \tag{6.6}$$

where $c_1, c_2, \ldots, c_k$ are constants such that $\sum_{i=1}^{k} c_i = 0$.

Consider for example four treatments A, B, C, and D. One might be interested in the following comparisons (contrasts):

$$L_1 = \mu_A - \frac{\mu_B + \mu_C + \mu_D}{3},$$

$$L_2 = \mu_B - \frac{\mu_C + \mu_D}{2},$$

$$L_3 = \mu_C - \mu_D.$$

(a) In L_1, $c_A = 1, c_B = c_C = c_D = -\frac{1}{3}$. Hence, L_1 is a linear combination of the four means and $\sum c_i = 0$. Thus, L_1 is a contrast.
(b) In L_2, $c_A = 0, c_B = 1, c_C = c_D = -\frac{1}{2}$. Hence, L_2 is a linear combination of the four means and again, $\sum c_i = 0$. Thus, L_2 is also a contrast.
(c) Similarly, in L_3, $c_A = 0, c_B = 0, c_C = 1, c_D = -1$. Again, this indicates that L_3 is a linear combination of the means and that $\sum c_i = 0$ in this case.

Alternatively, we may decide not to work with fractions and rewrite the contrasts, say, L_1 as $L_1 : \ 3\mu_A - \mu_B - \mu_C - \mu_D$. The results of removing fractions are displayed in the following table.

Contrasts	μ_A	μ_B	μ_C	μ_D
L_1	3	-1	-1	-1
L_2	0	2	-1	-1
L_3	0	0	1	-1

Definition of Orthogonal Contrasts

Two contrasts, say,

$$L_1 = \sum_{i=1}^{k} c_i\mu_i \quad \text{and} \quad L_2 = \sum_{i=1}^{k} d_i\mu_i,$$

where $c_1, c_2, \ldots, c_k$ and $d_1, d_2, \ldots, d_k$ are constants with $\sum_{i=1}^{k} c_i = 0$ and $\sum_{i=1}^{k} d_i = 0$ are said to be orthogonal if and only if

$$\sum_{i=1}^{k} c_i \, d_i = 0.$$

For instance, in the four treatments example above, the pairs of contrasts (L_1, L_2), (L_1, L_3), and (L_2, L_3) are orthogonal. That is, L_1, L_2, and L_3 are pairwise orthogonal. This has implications for partitioning the treatment SS into the three components represented by the contrasts. We shall examine an example of this in a later section in this chapter.

6.4 Partitioning the Treatments SS

For our example, the treatment degrees of freedom is 4, hence we can form at most four contrasts each based on 1 degree of freedom, in this study. Suppose the contrasts so formed are presented in the following table.

Contrasts	μ_1	μ_2	μ_3	μ_4	μ_5	$\sum_i c_i^2$
L_1	−1	−1	−1	−1	4	20
L_2	−1	−1	−1	3	0	12
L_3	−1	−1	2	0	0	6
L_4	−1	1	0	0	0	2

The above contrasts correspond respectively to the following null hypotheses:

$$H_0 : L_1 = \mu_5 - \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} = 0$$

$$H_0 : L_2 = \mu_4 - \frac{\mu_1 + \mu_2 + \mu_3}{3} = 0$$

$$H_0 : L_3 = \mu_3 - \frac{\mu_1 + \mu_2}{2} = 0$$

$$H_0 : L_4 = \mu_2 - \mu_1 = 0$$

These contrasts can be estimated and tested in MINITAB with the following code statements and corresponding partial output using the GLM procedure. Note how the contrasts are declared as covariates.

```
MTB > code (1) -1 (2) -1 (3) -1 (4) -1 (5) 4 c7 c8
MTB >  code (1) -1 (2) -1 (3) -1 (4) 3 (5) 0 c7 c9
MTB >  code (1) -1 (2) -1 (3) 2 (4) 0 (5) 0 c7 c10
MTB >  code (1) -1 (2) 1 (3) 0 (4) 0 (5) 0 c7 c11

MTB > GLM 'Y' = L1 L2 L3 L4;
SUBC>   Covariates 'L1' 'L2' 'L3' 'L4';
SUBC>   Brief 2 .

General Linear Model: Y versus

Factor  Type  Levels  Values


Analysis of Variance for Y, using Adjusted SS for Tests

Source  DF   Seq SS  Adj SS  Adj MS       F      P
L1       1   5134.9  5134.9  5134.9  185.15  0.000
L2       1   3544.7  3544.7  3544.7  127.81  0.000
L3       1   1736.7  1736.7  1736.7   62.62  0.000
L4       1    858.1   858.1   858.1   30.94  0.000
Error   45   1248.0  1248.0    27.7
Total   49  12522.4


S = 5.26632   R-Sq = 90.03%   R-Sq(adj) = 89.15%


Term         Coef  SE Coef       T      P
Constant  36.7080   0.7448   49.29  0.000
L1        -5.0670   0.3724  -13.61  0.000   5 vs others
L2        -5.4350   0.4807  -11.31  0.000   4 vs 1, 2 & 3
L3        -5.3800   0.6799   -7.91  0.000   3 vs 2 & 1
L4         -6.550    1.178   -5.56  0.000   2 vs 1
```

We note the following from the above partial MINITAB output for this example:

1. First, because the design is balanced (each treatment being equally replicated), the Type I and adjusted Type III SS are the same. This is often not the case for unbalanced designs.
2. The total of all contrasts SS equals the original Treatment (level) SS of 11,274.3188. That is,

$$5134.898 + 3544.707 + 1736.664 + 858.050 = 11,274.319$$

The sum of squares for the contrasts add up to the original treatment SS in this case because the contrasts are pairwise orthogonal. That is, for constants c_i and d_j, we have

$$\sum_{i=1}^{5}\sum_{j=1}^{5} c_i\, d_j = 0, \quad \text{for all pairs of contrasts.}$$

Calculating Contrasts SS For any contrast $L = \sum_{i=1}^{k} c_i\mu_i$, the SS is obtained as

$$\text{SS}(L) = \frac{\left(\sum_{i=1}^{k} c_i\, y_{i.}\right)^2}{b_i \sum c_i^2}, \tag{6.7}$$

where $y_{i.}$ is the total for treatment i. For example, for the L_1 contrast, $b_i = 10$, $\sum c_i^2 = 4^2 + 4 = 20$, hence, SS($L_1$) is

$$\frac{[(591.4)(-1) + (460.4)(-1) + (364.5)(-1) + (254.7)(-1) + (164.4)(4)]^2}{(20)(10)}$$

$$= \frac{(-1013.4)^2}{120} = 5134.898 = 5134.898$$

Similarly, SS(L_2) is

$$\frac{[(591.4)(-1) + (460.4)(-1) + (364.5)(-1) + (254.7)(3)]^2}{(12)(10)} = \frac{(-652.2)^2}{120}$$

$$= 3544.707$$

SS(L_3) is computed as

$$\frac{[(591.4)(-1) + (460.4)(-1) + (364.5)(2)]^2}{(6)(10)} = \frac{(-322.8)^2}{60} = 1736.664.$$

Finally, SS(L_4) is similarly calculated as

$$\frac{[(591.4)(-1) + (460.4)(1)]^2}{(2)(10)} = \frac{(-131)^2}{20} = 858.05.$$

Each of the above calculated contrasts SS is based on 1 degree of freedom. To obtain corresponding estimates of means using the MINITAB approach, we note that for contrast L_1 for instance, this equals

$$\frac{4\bar{y}_{5.} - (\bar{y}_{1.} + \bar{y}_{2.} + \bar{y}_{3.} + \bar{y}_{4.})}{20} = \frac{4(16.44) - (59.14 + 46.04 + 36.45 + 25.47)}{20}$$

$$= -5.067$$

with corresponding standard error calculated as

$$\sqrt{\frac{\frac{16S^2}{10}+\frac{4S^2}{10}}{20^2}}=\sqrt{\frac{20\times 27.7342}{20^2\times 10}}=0.3724.$$

Similarly for the contrast L_2, the mean estimate equals

$$\frac{3\bar{y}_{4.}-(\bar{y}_{1.}+\bar{y}_{2.}+\bar{y}_{3.})}{12}=\frac{3(25.47)-(59.14+46.04+36.45)}{12}=-5.435$$

with corresponding standard error calculated as

$$\sqrt{\frac{\frac{9S^2}{10}+\frac{3S^2}{10}}{12^2}}=\sqrt{\frac{12\times 27.7342}{12^2\times 10}}=0.4807.$$

Also for L_3, the mean estimate and corresponding standard error are computed as:

$$\bar{L}_3=\frac{2\bar{y}_{3.}-(\bar{y}_{1.}+\bar{y}_{2.})}{6}=\frac{2(36.45)-(59.14+46.04)}{6}=-5.380$$

$$\text{s.e.}=\sqrt{\frac{\frac{4S^2}{10}+\frac{2S^2}{10}}{6^2}}=\sqrt{\left(\frac{6\times 27.7342}{6^2\times 10}\right)}=0.6799.$$

Finally, the mean estimate and corresponding standard error for L_4 are computed as:

$$\bar{L}_4=\frac{\bar{y}_{2.}-\bar{y}_{1.}}{2}=\frac{46.04-59.14}{2}=-6.550$$

$$\text{s.e.}=\sqrt{\frac{\frac{S^2}{10}+\frac{S^2}{10}}{2^2}}=\sqrt{\left(\frac{2\times 27.7342}{2^2\times 10}\right)}=1.1778.$$

These calculated results are consistent with those generated using MINITAB which were displayed earlier. We may observe that denominators in all the cases are the number of replications r multiplied by the appropriate $\sum_i c_i^2$ in the contrast formulation. Of course in this example, we have only four contrasts that are all very significant, indicating again that level 5 gives the lowest growth rate of the bacteria and hence the best in this example.

6.5 Tests of Homogeneity of Variances

At this point, we need to test the assumptions underlying the analysis. The test that the data come from normal populations with constant variance σ^2 is sometimes referred to as *Bartlett's test of homogeneity of variances.*

The implication of the constant variance assumption under the assumptions above is that all treatments came from the same population. What this in essence means is that the treatments are assumed to have equal variances. If this were not so, then all the inferences, t-tests etc., are invalid. It is therefore important to check whether this assumption of equality of variances or homogeneity is invalidated and in order to do this, we use Bartlett's test of homogeneity. We are interested in testing the hypotheses,

$$\begin{aligned} H_0 &: \quad \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \\ H_a &: \quad \text{at least two of these are not equal} \end{aligned} \tag{6.8}$$

where k is the number of treatments or number of factor levels. To implement Bartlett's test, we perform the following calculations for the data in Table 6.3. To accomplish the above hypotheses, we will employ Bartlett's homogeneity test which is based on the following computations:

(1) Compute the pooled variance for the three sites as

$$S_P^2 = \frac{1}{N-k} \sum_{i=1}^{k} (r_i - 1) s_i^2, \quad i = 1, 2, \cdots, k$$

(2) Compute

$$q = (N-k) \log_{10} S_p^2 - \sum_{i=1}^{k} (r_i - 1) \log_{10} s_i^2$$

(3) Compute

$$c = 1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^{k} (r_i - 1)^{-1} - (N-k)^{-1} \right]$$

(4) Then,

$$\chi_0^2 = 2.3026 \frac{q}{c} \sim \chi_{k-1}^2$$

$N = \sum_{i=1}^{k} r_i$ and s_i^2 is the sample variance of the ith treatment. We would therefore reject H_0, whenever $\chi_0^2 > \chi_{\alpha,k-1}^2$.

6.5.1 Bartlett's Test for Data in Table 6.3

For the data in Table 6.3, we have $s_1^2 = 23.0849$, $s_2^2 = 25.5293$, $s_3^2 = 35.2117$, $s_4^2 = 20.1001$, $s_5^2 = 34.7449$, and $S_P^2 = 27.7342$. Here $k = 5$, $n_1 = n_2 = n_3 = n_4 = n_5 = 10$ (number of replications). Hence, $N = \sum r_i = 5 \times 10 = 50$ and therefore,

$$\begin{aligned} q &= 45 \log 27.7342 - 9(\log 23.0849 + \log 25.5293 + \log 35.2117 \\ &\quad + \log 20.1001 + \log 34.7449) \\ &= 64.9350 - 9(7.1611) \\ &= 0.4858 \end{aligned}$$

That is, $q = 0.4858$.

$$\begin{aligned} c &= 1 + \frac{1}{3 \times (5-1)} \left[\frac{5}{9} - \frac{1}{45} \right] \\ &= 1 + \frac{2}{45} \\ &= 1.0444 \end{aligned}$$

Hence,

$$X_0^2 = 2.3026 \left(\frac{0.4858}{1.0444} \right) = 1.0710 \tag{6.9}$$

But, $\chi^2_{0.05,4} = 7.35$ and since $1.0710 \ll 7.35$, we would fail to reject H_0 and conclude that indeed, the treatment populations all have the same variance at the 5 % significance point. Bartlett's test is implemented in MINITAB with the following statements and modified output (Fig. 6.2).

```
Test for Equal Variances: Y versus LEVELS

MTB > Vartest 'Y' 'LEVELS';
SUBC>   Confidence 95.0.

Test for Equal Variances: Y versus LEVELS

95% Bonferroni confidence intervals for standard deviations

LEVELS   N    Lower    StDev    Upper
     1  10  2.96775  4.80467  10.9432
     2  10  3.12092  5.05266  11.5080
     3  10  3.66527  5.93394  13.5152
     4  10  2.76925  4.48331  10.2113
     5  10  3.64090  5.89448  13.4253


Bartlett's Test (Normal Distribution)
Test statistic = 1.07, p-value = 0.899


Levene's Test (Any Continuous Distribution)
Test statistic = 0.19, p-value = 0.941
```

The calculated statistic of 1.07 under Bartlett's test agrees exactly with our result. The p value indicates that we would fail to reject H_0. It should be pointed out here that Bartlett's test is often used when it is assumed that

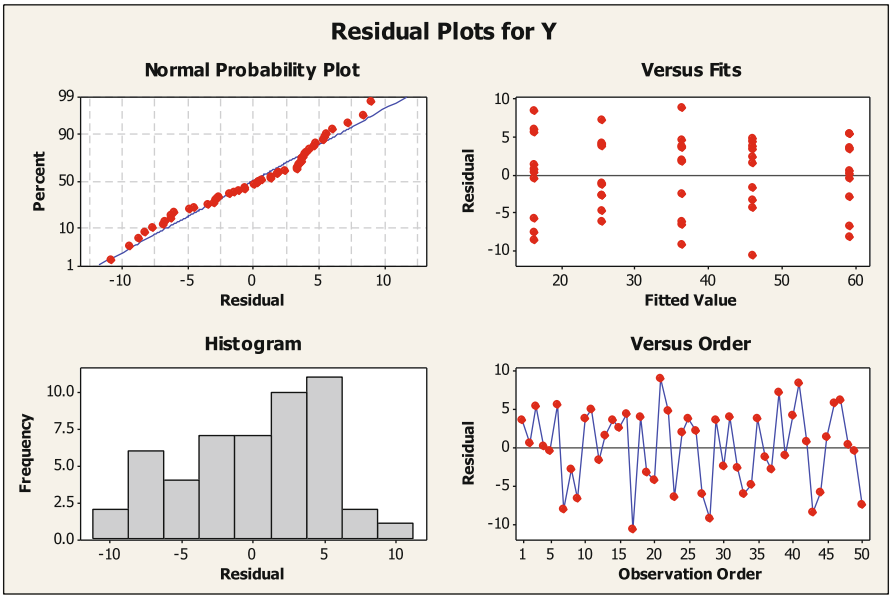


Fig. 6.2 Four MINITAB plots for residuals

the errors follow a normal distribution. When this is not the case, Levene's test is most appropriate.

The residual plots for the data indicate that the normality assumption is justified. A more formal normality test is provided in MINITAB as follows:

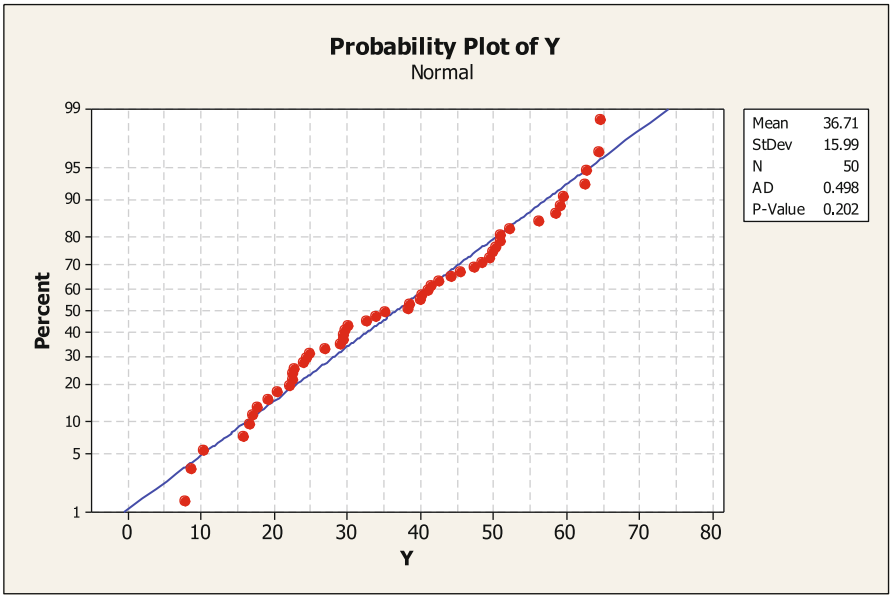


Fig. 6.3 Normality test for the residuals

The Anderson–Darling normality test in Fig. 6.3 gives a p value of 0.498 indicating that we would fail to reject the hypothesis that the data came from a normal population.

6.6 Nonparametric Test

In situations where the homogeneity assumption or normality assumption fails, the *Kruskal–Wallis test* (K–W) provides an alternative. It is simply, "analysis of variance by ranks" and can be more powerful than the traditional ANOVA test, especially if any of the assumptions were violated. The Kruskal–Wallis test statistic for testing the hypotheses in (6.10)

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = \cdots = \mu_k \\ H_a &: \quad \text{at least two of these are equal} \end{aligned} \tag{6.10}$$

is given by:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1). \tag{6.11}$$

where n_i is the number of observations in group or treatment i, $n = \sum_{i=1}^{k} n_i$ is the total number of observations for the k groups, and R_i is the sum of the ranks of the n_i observations in group i. The hypothesis is rejected when $H > \chi_\alpha^2$ with $(k-1)$ degrees of freedom.

Example

A completely randomized design produced the following sample results (Blaisdell 1993).

Sample 1	Sample 2	Sample 3
69 (13)	63 (8)	51 (1)
73 (15)	64 (9)	56 (2)
70 (14)	60 (5)	59 (4)
68 (12)	61 (6)	58 (3)
74 (16)	65 (10)	62 (7)
	67 (11)	
$n_1 = 5$	$n_2 = 6$	$n_3 = 5$
$R_1 = 70$	$R_2 = 49$	$R_3 = 17$

Here, $n = 5 + 6 + 5 = 16$ and hence,

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{16(17)}\left[\frac{70^2}{5} + \frac{49^2}{6} + \frac{17^2}{5}\right] - 3(17)$$
$$= \frac{12}{272}(1437.9667) - 51$$
$$= 12.440$$

The K–W test is implemented in MINITAB as follows; The data are first read in as factors and response variables.

```
Data Display

Row  Factors    Resp
  1        1      69
  2        1      73
  3        1      70
  4        1      68
  5        1      74
  6        2      63
  7        2      64
  8        2      60
  9        2      61
 10        2      65
 11        2      67
 12        3      51
 13        3      56
 14        3      59
 15        3      58
 16        3      62

MTB > Kruskal-Wallis 'Resp' 'Factors'.

Kruskal-Wallis Test: Resp versus Factors

Kruskal-Wallis Test on Resp

Factors   N  Median  Ave Rank      Z
1         5   70.00      14.0   3.12
2         6   63.50       8.2  -0.22
3         5   58.00       3.4  -2.89
Overall  16               8.5

H = 12.44  DF = 2  P = 0.002
```

The computed value from MINITAB agrees with our computed value above. The p value suggests that the null hypothesis will be strongly rejected in this case when the computed H statistic is compared to a chi-squared distribution with 2 degrees of freedom. In the above example, the ranks are not tied. If there are ties in the rankings, there is a small adjustment that needs to be made. This is automatically done in MINITAB and we do not plan to give an example of this here.

6.7 ANOVA with Unequal Replication

We are interested here in the case where $r_i \neq r_j$ for some i and j. We demonstrate the analysis with the following example. This example from Zar (1993) relates to 19 pigs that are assigned at random among four experimental groups. Each group is fed a different diet. The data are pig body weights in kilograms. Are there significant differences in the effects of the diets on the pigs?

	Feed 1	Feed 2	Feed 3	Feed 4
	60.8	68.7	102.6	87.9
	57.0	67.7	102.1	84.2
	65.0	74.0	100.2	83.1
	58.6	66.3	96.5	85.7
	61.7	69.8		90.3
n_i	5	5	4	5
T_i	303.1	346.5	401.4	431.2

In this example, the sample sizes are not all equal. The total number of observations is $n* = 5 + 5 + 4 + 5 = 19$. $y_{++} = 1482.2$. The total sum of squares is computed as:

$$\text{Total SS} = 60.8^2 + 68.7^2 + 69.8^2 + 90.3^2 - \frac{(1482.2)^2}{19} = 4354.698$$

Similarly, the Feeds sum of squares is computed as:

$$\begin{aligned}\text{Feeds SS} &= \frac{(303.1)^2}{5} + \frac{(346.5)^2}{5} + \frac{(401.4)^2}{4} + \frac{(431.2)^2}{5} - \frac{(1482.2)^2}{19} \\ &= 4226.348\end{aligned}$$

and the error sum of squares is obtained by subtraction. That is,

$$\text{Error SS} = 4354.698 - 4226.348 = 128.350.$$

The resulting analysis of variance table is presented in Table 6.5. The hypotheses,

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_a &: \quad \text{at least two of these means are not equal} \end{aligned} \tag{6.12}$$

Table 6.5 Analysis of variance table for the example

Source of variation	d.f.	SS	MS	F
Feeds	3	4226.348	1408.783	164.64
Error	15	128.350	8.557	
Total	18	4354.698		

are tested by comparing the computed F value of 164.64 with an $F(3,15, 0.05) = 3.29$. Since $164.64 >> 3.29$, we would strongly reject H_0. There are significant differences between the feeds as they affect the pigs at the 0.05 level of significance (Fig. 6.4).

The standard error for comparing any two feeds means is $\sqrt{8.557\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$. For two means with equal replications, say feed 1 and feed 2 for example, the standard error becomes $\sqrt{8.557\left(\frac{1}{5}+\frac{1}{5}\right)} = 1.1850$. In general, this is given by $\sqrt{\frac{2S^2}{n}}$, where $S^2 = \text{EMS}$.

On the other hand, for two means with unequal replications, such as feed 1 and feed 3, the standard error is computed as: $\sqrt{8.557\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = \sqrt{8.557\left(\frac{1}{5}+\frac{1}{4}\right)} = 1.9623$. The ANOVA for this example is analyzed in MINITAB as follows:

```
MTB > set C1
DATA> 1 1 1 1 1 2 2 2 2 2 3 3 3 3 4 4 4 4 4
DATA> end
MTB > set c2
DATA> 60.8 57 65 58.6 61.7 68.7 67.7 74 66.3 69.8
DATA> 102.6 102.1 100.2 96.5 87.9 84.2 83.1 85.7 90.3
DATA> end
MTB > print c1 c2

Data Display

Row  FEEDS    Wgt
  1      1   60.8
  2      1   57.0
  3      1   65.0
  4      1   58.6
  5      1   61.7
  6      2   68.7
  7      2   67.7
  8      2   74.0
  9      2   66.3
 10      2   69.8
 11      3  102.6
 12      3  102.1
 13      3  100.2
 14      3   96.5
 15      4   87.9
 16      4   84.2
 17      4   83.1
 18      4   85.7
 19      4   90.3

MTB > Oneway 'Wgt' 'FEEDS';
SUBC>   Tukey 5;
SUBC>   GBoxplot;
SUBC>   GNormalplot;
SUBC>   NoDGraphs.

One-way ANOVA: Wgt versus FEEDS
```

```
Source  DF       SS       MS       F      P
FEEDS    3  4226.35  1408.78  164.64  0.000
Error   15   128.35     8.56
Total   18  4354.70

S = 2.925   R-Sq = 97.05%   R-Sq(adj) = 96.46%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level  N    Mean  StDev  --+---------+---------+---------+-------
1      5   60.62   3.06  (--*-)
2      5   69.30   2.93         (--*-)
3      4  100.35   2.77                                  (--*-)
4      5   86.24   2.90                        (-*-)
                         --+---------+---------+---------+-------
                           60        72        84        96

Pooled StDev = 2.93


Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of FEEDS

Individual confidence level = 98.87%


FEEDS = 1 subtracted from:

FEEDS   Lower  Center   Upper  -----+---------+---------+---------+----
2       3.343   8.680  14.017                  (-*--)
3      34.069  39.730  45.391                             (-*-)
4      20.283  25.620  30.957                       (-*-)
                           -----+---------+---------+---------+----
                              -25         0         25        50


FEEDS = 2 subtracted from:

FEEDS   Lower  Center   Upper  -----+---------+---------+---------+----
3      25.389  31.050  36.711                        (-*--)
4      11.603  16.940  22.277                   (-*-)
                           -----+---------+---------+---------+----
                              -25         0         25        50


FEEDS = 3 subtracted from:

FEEDS    Lower   Center   Upper  -----+---------+---------+---------+----
4      -19.771  -14.110  -8.449        (-*--)
                                 -----+---------+---------+---------+----
                                    -25         0         25        50

Bartlett's Test (Normal Distribution)
Test statistic = 0.03, p-value = 0.998


Levene's Test (Any Continuous Distribution)
Test statistic = 0.02, p-value = 0.995
```

The means are all significantly different from one another with B being the best. None of the computed confidence intervals includes zero.

6.8 One Factor with Quantitative Levels

In most experimental designs, the levels of the treatments are sometimes equally spaced. This allows us to explore the relationship between the response variables and the qualitative variables more intimately in terms of examining linear, quadratic, and cubic effects if the quantitative has four levels. We give an example below (Table 6.6).

```
MTB > read c1-c3
DATA> 18 1 33.6
DATA> 24 1 31.1
DATA> 30 1 33.0
DATA> 36 1 28.4
DATA> 42 1 31.4
DATA> 18 2 37.1
DATA> 24 2 34.5
DATA> 30 2 29.5
DATA> 36 2 29.9
DATA> 42 2 28.3
DATA> 18 3 34.1
DATA> 24 3 30.5
DATA> 30 3 29.2
```

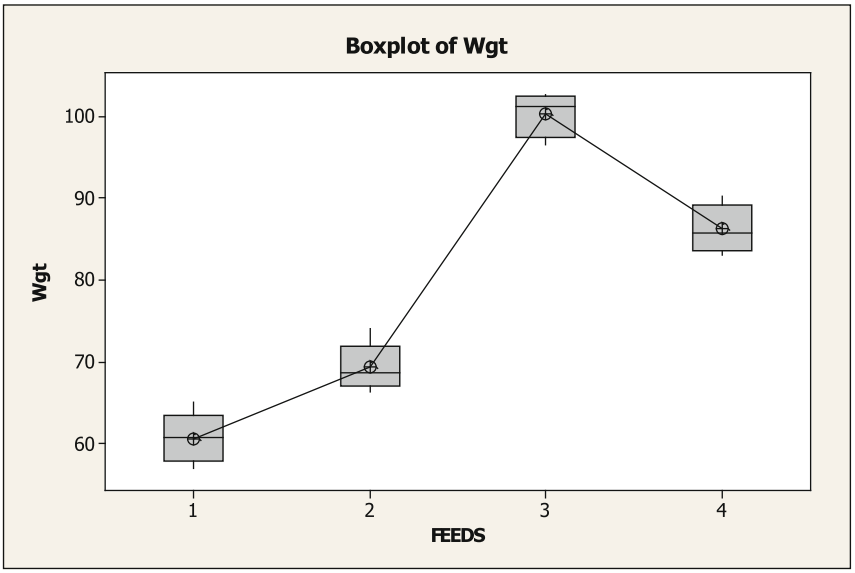


Fig. 6.4 Box plot of feed means

Table 6.6 Data on row spacing on yield of soybean

	Row spacings (inches)				
Block	18	24	30	36	42
1	33.6	31.1	33.0	28.4	31.4
2	37.1	34.5	29.5	29.9	28.3
3	34.1	30.5	29.2	31.6	28.9
4	34.6	32.7	30.7	32.3	28.6
5	35.4	30.7	30.7	28.1	29.6
6	36.1	30.3	27.9	26.9	33.4

```
DATA> 36 3 31.6
DATA> 42 3 28.9
DATA> 18 4 34.6
DATA> 24 4 32.7
DATA> 30 4 30.7
DATA> 36 4 32.3
DATA> 42 4 28.6
DATA> 18 5 35.4
DATA> 24 5 30.7
DATA> 30 5 30.7
DATA> 36 5 28.1
DATA> 42 5 29.6
DATA> 18 6 36.1
DATA> 24 6 30.3
DATA> 30 6 27.9
DATA> 36 6 26.9
DATA> 42 6 33.4
DATA> end
30 rows read.
MTB > GLM 'Y' = Blocks Row;
SUBC>   Brief 2 .

General Linear Model: Y versus Blocks, Row

Factor  Type   Levels  Values
Blocks  fixed       6  1, 2, 3, 4, 5, 6
Row     fixed       5  18, 24, 30, 36, 42


Analysis of Variance for Y, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS  Adj MS     F      P
Blocks   5    5.410    5.410   1.082  0.29  0.911
Row      4  125.661  125.661  31.415  8.50  0.000
Error   20   73.919   73.919   3.696
Total   29  204.990


S = 1.92248   R-Sq = 63.94%   R-Sq(adj) = 47.71%
```

Clearly, there are significant differences between the level means for row spacing. Now, let us partition the treatment (Spacing) SS into 4 single degree of freedom components using orthogonal polynomial contrasts. For a five-level treatment, we may use the table of orthogonal polynomials in the appendix. To implement this in MINITAB, we use the code command to accomplish the same.

```
MTB >  code (18) -2 (24) -1 (30) 0 (36) 1 (42) 2 c1 c4
MTB >  code (18) 2 (24) -1 (30) -2 (36) -1 (42) 2 c1 c5
MTB >  code (18) -1 (24) 2 (30) 0 (36) -2 (42) 1 c1 c6
MTB >  code (18) 1 (24) -4 (30) 6 (36) -4 (42) 1 c1 c7
MTB > print c1-c7

Data Display

Row  Row  Blocks     Y  RL  RQ  RC  RQT
  1   18       1  33.6  -2   2  -1    1
  2   24       1  31.1  -1  -1   2   -4
  3   30       1  33.0   0  -2   0    6
  4   36       1  28.4   1  -1  -2   -4
  5   42       1  31.4   2   2   1    1
  6   18       2  37.1  -2   2  -1    1
  7   24       2  34.5  -1  -1   2   -4
  8   30       2  29.5   0  -2   0    6
  9   36       2  29.9   1  -1  -2   -4
 10   42       2  28.3   2   2   1    1
 11   18       3  34.1  -2   2  -1    1
 12   24       3  30.5  -1  -1   2   -4
 13   30       3  29.2   0  -2   0    6
 14   36       3  31.6   1  -1  -2   -4
 15   42       3  28.9   2   2   1    1
 16   18       4  34.6  -2   2  -1    1
 17   24       4  32.7  -1  -1   2   -4
 18   30       4  30.7   0  -2   0    6
 19   36       4  32.3   1  -1  -2   -4
 20   42       4  28.6   2   2   1    1
 21   18       5  35.4  -2   2  -1    1
 22   24       5  30.7  -1  -1   2   -4
 23   30       5  30.7   0  -2   0    6
 24   36       5  28.1   1  -1  -2   -4
 25   42       5  29.6   2   2   1    1
 26   18       6  36.1  -2   2  -1    1
 27   24       6  30.3  -1  -1   2   -4
 28   30       6  27.9   0  -2   0    6
 29   36       6  26.9   1  -1  -2   -4
 30   42       6  33.4   2   2   1    1
```

The created columns c4–c7 are respectively the Linear (RL), Quadratic (RQ), Cubic (RC), and Quartic (RQT) components. The complete analysis is presented with the following MINITAB commands and corresponding output.

```
MTB > GLM 'Y' = Blocks RL RQ RC RQT;
SUBC>   Covariates 'RL' 'RQ' 'RC' 'RQT';
SUBC>   Brief 2 .

General Linear Model: Y versus Blocks

Factor  Type   Levels  Values
Blocks  fixed       6  1, 2, 3, 4, 5, 6
```

```
Analysis of Variance for Y, using Adjusted SS for Tests

Source  DF   Seq SS  Adj SS  Adj MS      F      P
Blocks   5    5.410   5.410   1.082   0.29  0.911
RL       1   91.267  91.267  91.267  24.69  0.000
RQ       1   33.693  33.693  33.693   9.12  0.007
RC       1    0.504   0.504   0.504   0.14  0.716
RQT      1    0.197   0.197   0.197   0.05  0.820
Error   20   73.919  73.919   3.696
Total   29  204.990

S = 1.92248   R-Sq = 63.94%   R-Sq(adj) = 47.71%


Term         Coef  SE Coef      T      P
Constant  31.3033   0.3510  89.18  0.000
RL        -1.2333   0.2482  -4.97  0.000
RQ         0.6333   0.2098   3.02  0.007
RC        -0.0917   0.2482  -0.37  0.716
RQT       0.02167  0.09381   0.23  0.820
```

Both the linear and quadratic components are significant, hence the anticipated model would be:

$$Y_i = 31.3033 - 1.2333x + 0.633x^2 \tag{6.13}$$

The plot of the response is presented in Fig. 6.5.

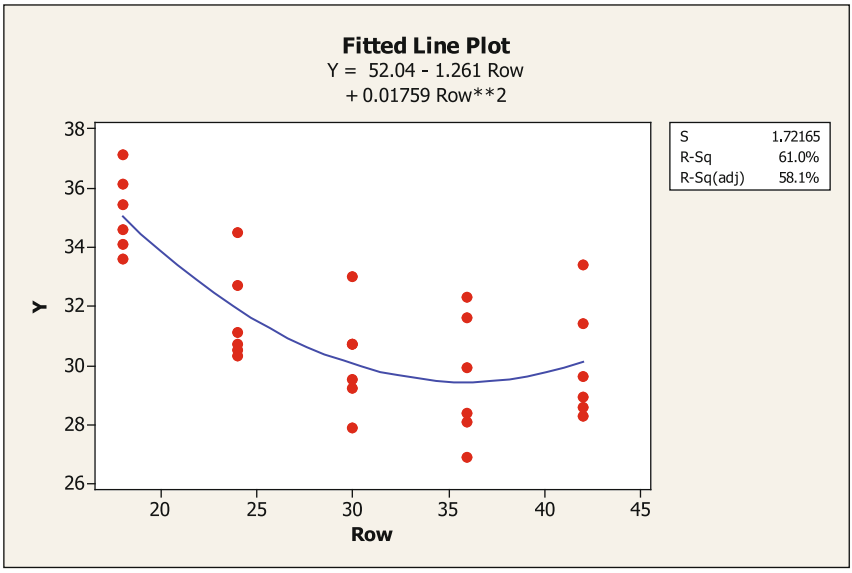


Fig. 6.5 Quadratic response plot of effects of sowing

6.9 Orthogonal Polynomials for Unequal Spacing

The table of orthogonal polynomial coefficient in the appendix is only suitable for the cases when the levels of factors are equally spaced. When this is not the case, it might lead to simple complications, but we can generally generate our own orthogonal polynomials that satisfy all the conditions set forth for orthogonal polynomials. Suppose we have a fertilizer experiment with three levels {0, 10, 30} kg/ha. We see that orthogonal polynomial procedure can not be readily applied in this case because the levels are not equally spaced. Let us see how we can generate an appropriate orthogonal polynomial coefficients for these levels. We accomplish this as follows:

1. First, we code the levels to the smallest integers by dividing with ten to get the following

Levels	Coded X	X^2	X^3
0	0	0	0
10	1	1	1
30	3	9	27
$\sum$	4	10	28

2. Form the following three equations with three unknowns (a, b, c)

$$\begin{aligned} L_i &= a + X_i \\ Q_i &= b + cX_i + X_i^2 \\ L_iQ_i &= (a + X_i)(b + cX_i + X_i^2) \end{aligned} \tag{6.14}$$

3. Summing the equations in 6.14, we have

$$\begin{aligned} \sum L_i &= 3a + \sum X_i = 0 \\ \sum Q_i &= 3b + c\sum X_i + \sum X_i^2 = 0 \\ \sum L_iQ_i &= 3ab + (ac + b)\sum X_i + (a + c)\sum X_i^2 + \sum X_i^3 = 0 \end{aligned} \tag{6.15}$$

4. The above leads to

$$\begin{aligned} 3a + 4 &= 0 \\ 3b + 4c + 10 &= 0 \\ 3ab + 4(ac + b) + 10(a + c) + 28 &= 0 \end{aligned} \tag{6.16}$$

5. Solving the system of equations in (6.16), we have

$$a = -\frac{4}{3}$$

$$b = \frac{6}{7}$$
$$c = -\frac{22}{7}$$

6. Substituting these values back in the set of equations in (6.14) for $i = 1, 2, 3$, we have

$$L_1 = a + X_1 = -\frac{4}{3} + 0 = -\frac{4}{3}$$
$$L_2 = a + X_2 = -\frac{4}{3} + 1 = -\frac{1}{3}$$
$$L_3 = a + X_3 = -\frac{4}{3} + 3 = +\frac{5}{3}$$

Similarly,

$$Q_1 = b + cX_1 + X_1^2 = \frac{6}{7} - \frac{22}{7}0 + 0 = +\frac{6}{7}$$
$$Q_2 = b + cX_2 + X_2^2 = \frac{6}{7} - \frac{22}{7}1 + 1 = -\frac{9}{7}$$
$$Q_3 = b + cX_2 + X_2^2 = \frac{6}{7} - \frac{22}{7}3 + 9 = +\frac{3}{7}$$

7. Hence, the orthogonal polynomial coefficients are

	Levels of factor		
Linear	$-\frac{4}{3}$	$-\frac{1}{3}$	$\frac{5}{3}$
Quadratic	$\frac{6}{7}$	$-\frac{9}{7}$	$\frac{3}{7}$

Dropping the common denominators, we have the final orthogonal coefficients

	Levels of factor		
Linear	-4	-1	5
Quadratic	6	-9	3

8. Of course if we have four or more levels for the factor, the procedure is exactly the same except that we would have more constants to determine from the systems of equations. Again, I want to reiterate here that if the levels are equally spaced, then we do not have any problem and the tables of orthogonal polynomials can be used to handle this situation.

6.10 Two-Factor Analysis of Variance

A two-factor experiment comprises, say, two factors A and B each having a and b levels respectively. The resulting ab treatment combinations can be laid out in a completely randomized design. If all treatment combinations are equally replicated, then we say that the design is balanced. Otherwise, the design is unbalanced or nonorthogonal. We will however, start our discussion of two-factor experiments by considering a balanced case with interaction present. When interaction is present, the general model can be written as

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}, \quad i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, b, \tag{6.17}$$

with $\sum_i \alpha_i = 0 = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij}$.
Thus,

$$\alpha_i = \mu_{i.} - \mu_{..}, \quad \beta_j = \mu_{.j} - \mu_{..}, \quad \text{and} \quad (\alpha\beta)_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..},$$

where in (6.17),

- μ is the overall mean
- α_i is the main effect of the ith level of factor A
- β_j is the main effect of the jth level of factor B
- $(\alpha\beta)_{ij}$ is the interaction effect between level i of factor A and level j of factor B
- $\varepsilon_{ij} - N(0, \sigma^2)$.

6.10.1 An Example: Bacteria Counts

The data in the following table are taken from the bacteria count data submitted by Binnie, N. S. to the Journal of Statistics Education Data Archive. The data contain measurement of bacteria counts of *Staphylococcus aureus* (strain 1), the concentration of tryptone (a nutrient) and the temperature (in °C) of incubation. In order to test for the effects of concentration, temperature, and their interactions, an analysis of variance model is proposed (Table 6.7).

(a) Obtain an estimated treatment means plot. Does it appear that concentration and temperature have any effects? Explain.
(b) Obtain the ANOVA model for the data.
(c) Set up the ANOVA table.
(d) Test for the interaction effects. Use $\alpha = 0.05$. State the null and the alternative hypotheses, your decision, and conclusion. State the p value of the test.

Table 6.7 Replicated observations for two-factor data

	Temperature		
Concentration	27	35	43
0.6	9, 97	66, 110	98, 123
0.8	16, 123	93, 149	82, 146
1.0	22, 132	147, 189	120, 106
1.2	30, 263	199, 263	148, 232
1.4	27, 145	168, 197	132, 163

(e) Test to see if concentration and temperature have significant effects on the number of counts at $\alpha = 0.05$. State the null and the alternative hypotheses, your decision, and conclusion. State the p value of the test.

The Total SS is obtained as:

$$\text{Total SS} = (9)^2 + (97)^2 + (66)^2 + \cdots + (132)^2 + (163)^2 - \frac{(3795)^2}{30} = 133108$$

To obtain the SS for concentrations and temperature, first we form the two-way table of totals as follows:

	Temperature			
Concentration	27	35	43	Totals
0.6	106	176	221	503
0.8	139	242	228	609
1.0	154	336	226	716
1.2	293	462	380	1135
1.4	172	365	295	832
Totals	864	1581	1350	3795

Hence, Concentrations and Temperature SS are computed as:

$$\text{Conc. SS} = \frac{(503)^2}{6} + \frac{(609)^2}{6} + \cdots + \frac{(832)^2}{6} - \frac{(3795)^2}{30} = 39,432$$

$$\text{Temp. SS} = \frac{(864)^2}{12} + \frac{(1581)^2}{12} + \frac{(1350)^2}{12} - \frac{(3795)^2}{30} = 26,788$$

The Interaction SS is computed as:

$$\text{Conc.} \times \text{Temp. SS} = \frac{(106)^2}{2} + \frac{(176)^2}{2} + \cdots + \frac{(295)^2}{2} - \frac{(3795)^2}{30} - \text{Conc. SS} - \text{Temp. SS} = 4781$$

The resulting analysis of variance table is therefore presented as in Table 6.8.

Table 6.8 ANOVA table for the data in Table 6.7

Source of variation	d.f.	SS	MS	F
Concentration	4	39,432	9857.9	2.38
Temperature	2	26,788	13,394	3.23
C × T	8	4781	597.6	0.14
Error	15	62,107	4140.4	
Total	29	133,108		

Our results indicate that at the 5 % point, neither the concentration, temperature means are significantly different, nor is the interaction effect significant. The MINITAB implementation is presented below.

```
Row  CONC  TEMP    Y
  1   0.6    27    9
  2   0.6    27   97
  3   0.6    35   66
  4   0.6    35  110
  5   0.6    43   98
  6   0.6    43  123
  7   0.8    27   16
 ...................
 25   1.4    27   27
 26   1.4    27  145
 27   1.4    35  168
 28   1.4    35  197
 29   1.4    43  132
 30   1.4    43  163

MTB > Twoway 'Y' 'CONC' 'TEMP';
SUBC>   Means 'CONC' 'TEMP'.

Two-way ANOVA: Y versus CONC, TEMP

Source       DF      SS       MS     F      P
CONC          4   39432   9857.9  2.38  0.098
TEMP          2   26788  13394.1  3.23  0.068
Interaction   8    4781    597.6  0.14  0.995
Error        15   62107   4140.4
Total        29  133108

S = 64.35   R-Sq = 53.34%   R-Sq(adj) = 9.79%
```

6.11 Replication and Sample Size Determination

In this chapter, we will consider sample size determination based on hypothesis testing approach for

1. One-sample t and z tests
2. Two-sample t and z tests
3. One-way ANOVA design.

If the parameter of interest is μ, and the hypotheses have a significance level α, and if we let d be the true difference that the hypothesis is desired to test,

and it is also desired that the test rejects the null hypothesis $H_0 : \mu = \mu_0$ with a probability of at least $1-\beta$, then, $1-\beta$ will be referred to as the power of the test. Usually in any investigative experimental design, the values of α, d and $1-\beta$ are often specified a priori by the investigator and our goal is to find the minimum sample size that will satisfy these requirements. We present in the following sections the calculations required to determine the necessary sample sizes for varying values of α, d and $1-\beta$.

Most statistical softwares these days have the capability of computing sample sizes required for various types of cases and therefore the need for extensive use of formulae is no longer necessary for the understanding of sample size calculations at this level. Hence, we have employed MINITAB to generate the required sample sizes for us given the required parameters. We have therefore demonstrated its use in a few examples in this section.

6.11.1 One- or Two-Sample t Test

For one or two sample tests for population means, if we let r be the number of observations per sample required to guarantee the requirements imposed by the three parameters above, then, r is the smallest integer such that:

$$r \geq a[t(\nu, b\alpha) + t(\nu, \beta)]^2 \left(\frac{\sigma}{d}\right)^2, \tag{6.18}$$

where

(i) σ is the population standard deviation
(ii) ν denotes the degree of freedom for the t-test
(iii) and

$$a = \begin{cases} 1 & \text{for a one-sample test} \\ 2 & \text{for a two-sample test;} \end{cases} \qquad b = \begin{cases} 1 & \text{for a one-sided test} \\ \frac{1}{2} & \text{for a two-sided test.} \end{cases}$$

A problem often associated with the sample size calculation as expressed in (6.18) is that σ is often unknown and it therefore calls for a good estimate of this parameter. Previous or past experience with similar designs is often very useful in getting a good estimate for σ. Of course, if the extent of departure from the mean is measured in terms of σ, that is, if for instance $d = k\sigma$, where k is specified, then, we can see that the expression in (6.18) will not involve the unknown σ at all.

Another result from (6.18) is that the sample size r depends only on the ratio $\frac{\sigma}{d}$. In other words, the sample size obtained from $d = 8$ and $\sigma = 2$ will be the same as that for the case when $d = 12$ and $\sigma = 3$.

Example

In a paired test experiment on weight gained by rats after exercise, the weight change can be tested as one-sample t test with $H_0 = 0$ and $H_a \neq 0$. How

large a sample is needed if we wish to test at $\alpha = 0.05$ with a 90 % chance of detecting a population mean different from $\mu_0 = 0$ by as little as 1.0 g., if it assumed that $s = 1.2523$? MINITAB is employed for this problem. The ensuing power curve is presented in Fig. 6.6 with the minimum sample size required being calculated as $r = 19$.

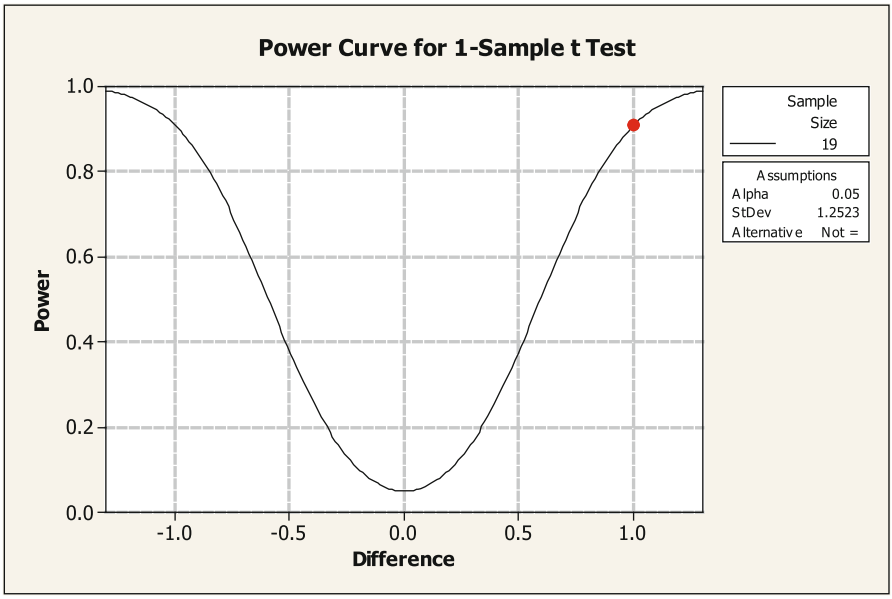


Fig. 6.6 Power curve for this example

```
MTB > Power;
SUBC>   TOne;
SUBC>     Difference 1;
SUBC>     Power .90;
SUBC>     Sigma 1.2523;
SUBC>   GPCurve.

Power and Sample Size

1-Sample t Test

Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0.05  Assumed standard deviation = 1.2523


            Sample  Target
Difference    Size   Power  Actual Power
         1      19     0.9      0.908217
```

6.11.2 One- and Two-Sample Z Test Example

In healthy males, the CD4 T-lymphocytes are normally distributed with a mean of $\mu = 1500$ cells/mm^3. In males with HIV infection present but with no diagnosis of AIDS, the CD4 cells are normally distributed with $\mu = 600$ and $\sigma = 150$. If a drug could increase the mean cell count to 700 cells/mm^3 and maintained that level, then the drug would be of value. What sample size is needed to detect a difference of 100 cells at $\alpha = 0.05$, with power $1 - \beta = 0.90$, if it is assumed that $H_0 = 600$ and $H_a = 700$? We can use MINITAB to solve this problem as follows:

```
Power Curve for 1-Sample Z Test

MTB > Power;
SUBC>   ZOne;
SUBC>     Difference 100;
SUBC>     Power .90;
SUBC>     Sigma 150;
SUBC>     Alternative 1;
SUBC>   GPCurve.

Power and Sample Size

1-Sample Z Test

Testing mean = null (versus > null)
Calculating power for mean = null + difference
Alpha = 0.05  Assumed standard deviation = 150


            Sample  Target
Difference    Size   Power  Actual Power
       100      20     0.9      0.909319
```

The power curve for the problem is presented in Figs. 6.7 and 6.8. Our calculations show that we would require a sample size of 20 for this problem.

Suppose the alternative hypothesis had been, $H_a : \mu \neq 600$, then we would have

```
Power Curve for 1-Sample Z Test

MTB > Power;
SUBC>   ZOne;
SUBC>     Difference 100;
SUBC>     Power .90;
SUBC>     Sigma 150;
SUBC>   GPCurve.

Power and Sample Size

1-Sample Z Test

Testing mean = null (versus not = null)


Calculating power for mean = null + difference
Alpha = 0.05  Assumed standard deviation = 150


            Sample  Target
Difference    Size   Power  Actual Power
       100      24     0.9      0.904228
```

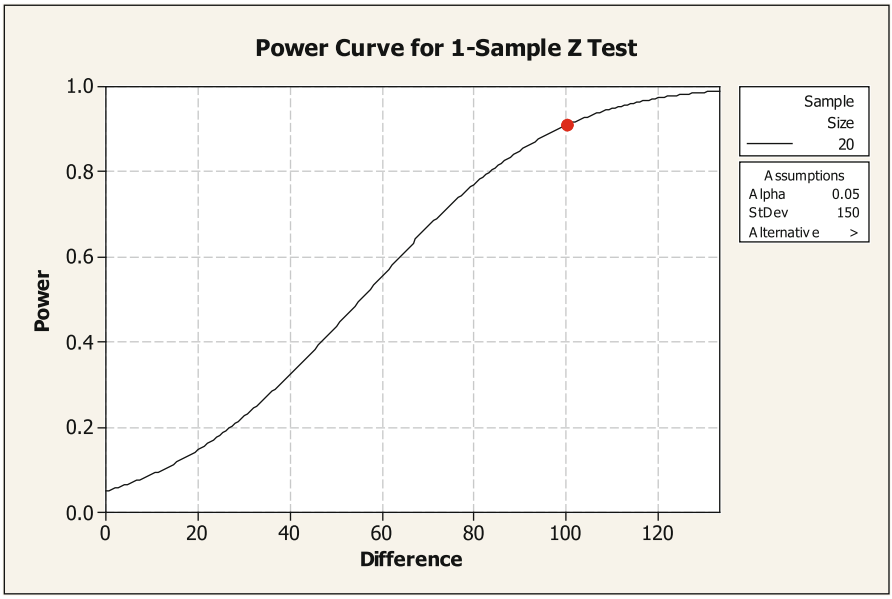


Fig. 6.7 Power curve for this example

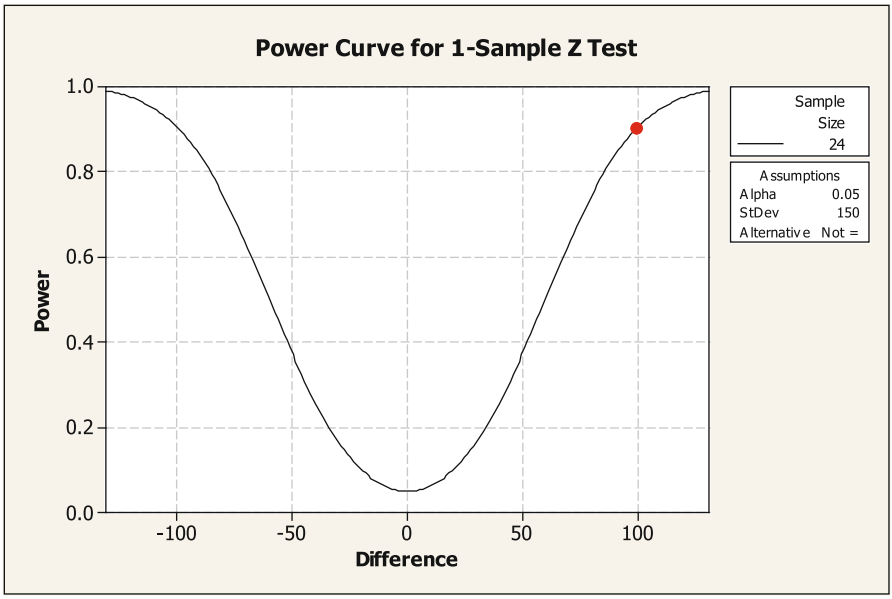


Fig. 6.8 Power curve for this example

In this case, the minimum sample size required will be $r = 24$.

6.11.3 Two-Sample t Test Example

This example is adapted from Zar (1999). It is desired to test for significant difference between the mean blood clotting times of persons using two different drugs. Suppose we wish to detect a true difference between the two means of 1 min and to have 96.6 % probability of detecting this difference, and testing at $\alpha = 0.05$ with an assumed $S_p = 0.7206$.

```
MTB > Power;
SUBC>   TTwo;
SUBC>     Difference 1;
SUBC>     Power .966;
SUBC>     Sigma 0.7206;
SUBC>   GPCurve.

Power and Sample Size

2-Sample t Test

Testing mean 1 = mean 2 (versus not =)
Calculating power for mean 1 = mean 2 + difference
Alpha = 0.05  Assumed standard deviation = 0.7206


            Sample  Target
Difference    Size   Power  Actual Power
         1      16   0.966      0.966869

The sample size is for each group.
```

We see that in this example, we require sample sizes of 16 to achieve our desired goals or objectives and the power curve is presented in Fig. 6.9.

6.11.4 Sample Size in One-Factor ANOVA

Of interest to researchers is the question of how many replications should we have per treatment level. Of course, it is often recommended that we use equal replication, that is, $n_1 = n_2 = \cdots, n_t$, assuming that we have t levels for the factor in question. Suppose the means of these levels are $\mu_1, \mu_2, \cdots, \mu_t$ respectively. MINITAB can generate the required sample sizes.

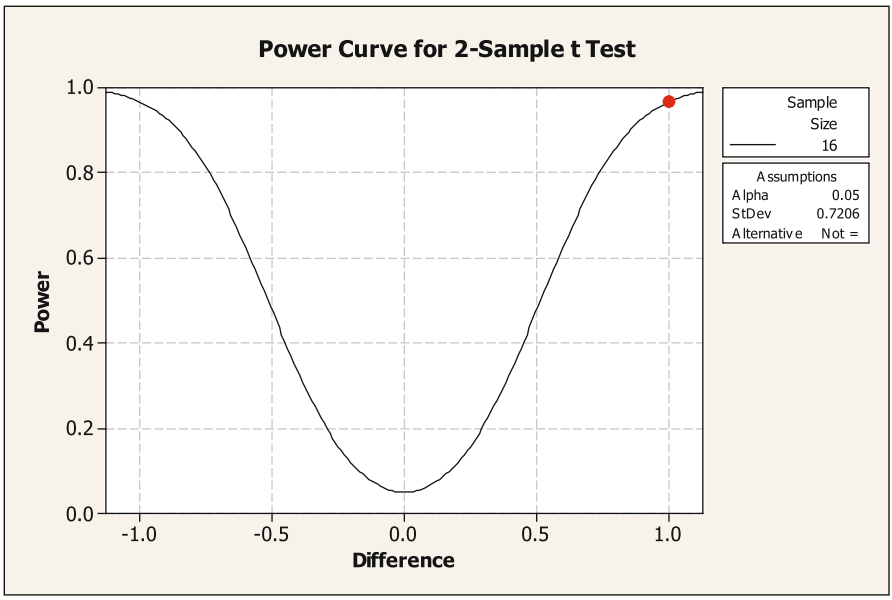


Fig. 6.9 Power curve for the two-sample t example

Consider the following example relating to an experiment to determine whether the development time for insect embryos (measured in days as elapsed from eggs laying to hatching) is the same at four different experimental temperatures. Suppose we wish to have a 90 % probability of detecting a difference between population means as small as 2 days, testing at $\alpha = 0.05$ level of significance with an assumed $\hat{\sigma}^2 = s^2 = 1.6550$.

```
MTB > Power;
SUBC>   OneWay 4;
SUBC>     MaxDifference 2;
SUBC>     Power .9;
SUBC>     Sigma 1.2865;
SUBC>   GPCurve.

Power and Sample Size

One-way ANOVA

Alpha = 0.05  Assumed standard deviation = 1.2865

Factors: 1  Number of levels: 4


   Maximum  Sample  Target
Difference    Size   Power  Actual Power
         2      13     0.9      0.906489

The sample size is for each level.
```

Our analysis indicates that the researcher would require 13 replicates each for each treatment level for this study. The power curve for this example is presented in Fig. 6.10. If the power of detection was reduced from 90 to 80 % for instance, then we would require ten replications each for the experiment, that is, a total of 40 experimental units (eggs) in this case.

Suppose in the last example, we have only 40 eggs available and six temperatures that might be tested. We still wish to detect a difference of at least 2 days, and power $(1 - \beta) > 0.75$ and $s^2 = 1.6550$, then we would have the following table of results:

t	r	Power
6	6	0.4378
5	8	0.6324
4	10	0.7973

where, if $t = 6$, the number of replicates would be $40/6 = 6.67$, that is 6. The computed power for this case is $0.4375 < 0.75$. Therefore, if we reduce the number of temperature levels we can use, from six to five (with eight replicates each), again, the computed power in this case is now 0.6324 which is still less than the desired power of minimum 0.75. Now if we choose only four of the six temperatures, the computed power is now 0.7973 which now meets our desired objective of minimum power of 0.75. Thus, we conclude that no more than 4 of the temperature levels be used in this experiment if we are limited to 40 experimental units (eggs).

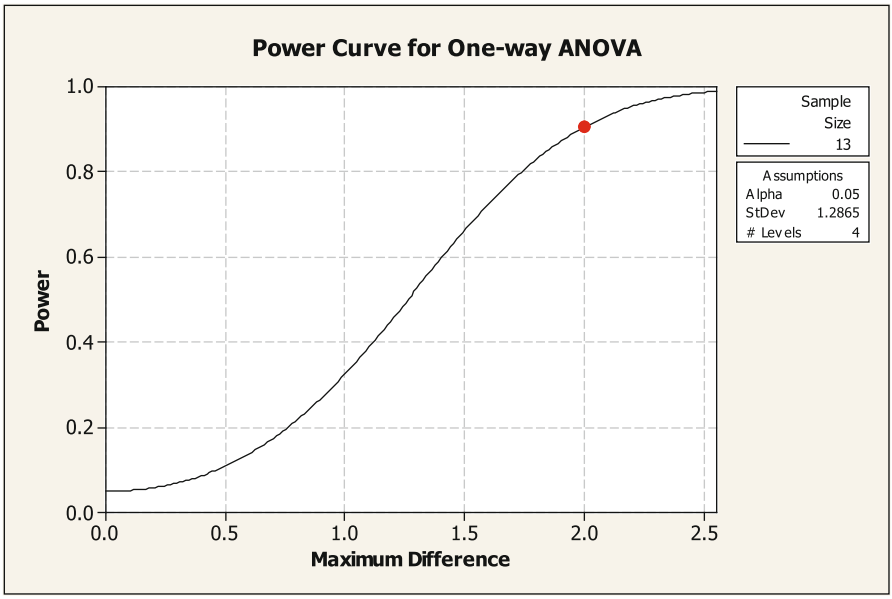


Fig. 6.10 Power curve for the ANOVA example

For equally replicated factor levels, the standard error of the difference between two means is $\sqrt{\frac{2\sigma^2}{r}}$. Thus, at a given significance level α, the observed difference for a given true difference of magnitude d, the number of replications required will be given by:

$$r = \frac{2t_0^2\sigma^2}{d^2} = 2t_0^2\left(\frac{\sigma}{d}\right)^2 = 2[t_\alpha + \Phi^{-1}(p)]^2\left(\frac{\sigma}{d}\right)^2 \approx 2[2.00 + \Phi^{-1}(p)]^2\left(\frac{\sigma}{d}\right)^2 \tag{6.19}$$

where t_0 is the students' t critical value, σ^2 is the unit plot variance, Φ is the normal distribution function and p is the the probability of obtaining an observed difference that will be considered significant. In most situations, we often set $t_{0.05} = 2.00$ and hence the approximate result on the far RHS. We can of course always replace σ^2 by a priori estimate of s^2. Results obtained above indicate that the number of replications to be employed in an experiment depends on (i) the resources available (as in the case of 40 insect eggs above), (ii) the treatment structure or simply put the number of treatments in the experiment, (iii) the size of the difference between the treatment means, and (iv) the relative importance of different comparisons. For example, consider an experiment having four new treatments and a control (or standard treatment). Suppose 30 plots are available for this experiment. We can have the following two possible designs.

(i) Six blocks of five plots each with each of the four new treatments and the control randomized within each block of size 5.
(ii) Five blocks of size 6 with each of the four new treatments and the control applied twice and the resulting randomized within each block of size 6.

For the above two designs, the standard errors for comparing (a) any two treatments, (b) a treatment with the control, and (c) new treatments means and control are computed as follows:

For *design (i)*, we have, the following standard errors:

$$(a)\ \sqrt{\frac{2s^2}{6}} = 0.577s, \qquad (b)\ \sqrt{\frac{2s^2}{6}} = 0.577s, \qquad (c)\ \sqrt{\frac{5s^2}{24}} = 0.456s.$$

Similarly for **design (ii)**, we have, the following standard errors:

$$(a)\ \sqrt{\frac{2s^2}{5}} = 0.632s, \qquad (b)\ \sqrt{\frac{3s^2}{10}} = 0.548s, \qquad (c)\ \sqrt{\frac{3s^2}{20}} = 0.387s.$$

Clearly, if the comparison of the control and treatment is of importance here, then, design (ii) will be preferable as it has a lower standard error for these comparisons than design (i). We may note here that in this case, the treatments will not be equally replicated. Unequal replications are sometimes very useful in experiments in which materials for all treatments may not go round as in breeding experiments. Unequal replication in experimental designs is no longer problematic in terms of analysis since the advent of powerful computer software can take care of unbalanced designs and the estimate of the error variance σ^2 can also be adequately estimated by pooling estimates of variance based on the assumption that estimate are the same for each observation. A problem may arise if this assumption were not true, and then in that case, the methods of transformations of data discussed in Chap. 10 will be employed.

6.12 Exercises

1. The MINITAB printout for an experiment utilizing a completely randomized design is shown below:

	ANOVA Table			
Source	d.f.	SS	MS	F
Factor	3	57,258	19,086	14.80
Error	34	43,836	1289	
Total	37	101,094		

a. How many treatments are involved in the experiment? What is the total sample size?
b. Conduct a test of the null hypothesis that the treatment means are equal. Use $\alpha = .01$.
c. What assumptions must be satisfied before the analysis above can be valid? (state three only).

2. The partially completed ANOVA table given here is for a one-way experiment

Source	d.f.	SS	MS	F
Treatments	–	3047.64	1015.88	–
Error	–	–	–	
Total	14	9966.75		

a. Give the number of levels for the treatments.
b. How many observations were collected for each treatment-level?
c. Complete the ANOVA table.
d. Test to determine whether the treatment means differ. Use $\alpha = 0.10$.

3. Dangerous chemicals from industrial wastes linked to cancer and other diseases can enter the food chain through their presence in lake sediments. The amounts of DDT were determined for samples of trout from four lakes and are given below.

```
Lake    Levels of DDT in Parts per million
---------------------------------------
1      1.7   1.4   1.9   1.1   2.1   1.8
2      0.3   0.7   0.5   0.1   1.1   0.9
3      2.7   1.9   2.0   1.5   2.6
4      1.2   3.1   1.9   3.7   2.8   3.5
---------------------------------------
```

(a) Do the data provide sufficient evidence at $\alpha = 0.05$ level to conclude that there is a difference in mean DDT levels for the four lakes?
(b) Compare the mean DDT levels in lakes 2 and 3 by constructing a 90 % confidence interval.
(c) Suppose it was hypothesized before collecting the data for this experiment that the mean DDT levels were not the same for lakes 1, 3, and 4. Is there sufficient evidence to support this belief?
(d) Conduct the tests necessary for the validation of your analysis.

4. The table below relates to the outcome of an experiment in which five subjects were assigned at random to each of the four dosages of a drug. The dependent variable y is a physiological measure that presumably is influenced by the amount of the drug administered.

Dosage levels			
0	5	10	15
10	9	14	17
8	13	13	15
12	12	11	14
11	10	12	18
9	11	15	16
50	55	65	80

The experiment was designed to see if there is a linear trend in the response.

(a) Complete the following analysis of variance table, explaining how the treatment SS of 105 was obtained.

Source	d.f.	SS	MS	F
Treatments	–	105	–	–
Error	–	–	–	
Total	–	145		

(b) Test whether the four dosage levels have significantly different effects on the physiological measure y (Use $\alpha = 0.05$).

(c) Are there any assumptions that must be satisfied? State them and conduct the appropriate tests for the validity of these assumptions.

(d) The treatment SS has been partitioned into linear, quadratic, and cubic components using tables of orthogonal polynomial coefficients displayed below.

Linear	-3	-1	1	3
Quadratic	1	-1	-1	1
Cubic	-1	3	-3	1

How was the linear SS computed? Test your results and suggest the most suitable regression model for the data.

(e) What relationship, if any, is there between the single observed F value of part (a) and the three observed F values of part (d)?

5. An agronomist has conducted a four-level fertilizer response experiment in which he tested the following three planned comparisons:

	μ_1	μ_2	μ_3	μ_4
L_1	1	-1	0	0
L_2	1	1	-2	0
L_3	1	1	1	-3

The sum of squares (SS) for the three comparisons are 75, 175, and 125, respectively. The value of MSE equals 25, and there were plots for each level.

(a) Is it possible to perform the test of the omnibus null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ from the available information? If so, is the test significant? If it is not possible, explain why not.
(b) Find the observed F value for each of the planned comparisons tested by the agronomist. Which, if any, are statistically significant at $\alpha = 0.05$?
(c) What relationship, if any, is there between the single observed F value of part (a) and the three observed F values of part (b)?
(d) For each of the L_i, indicate the null hypothesis that is being tested. Find the observed F value for each of the planned comparisons tested by the agronomist. Which, if any, are statistically significant at $\alpha = 0.05$?

6. Three different methods were used to determine the dissolved oxygen content of lake water (in mg/kg). Each of the three methods was applied to a sample of water six times, with the following results.

Method 1	Method 2	Method 3
10.96	10.88	10.73
10.77	10.75	10.79
10.90	10.80	10.78
10.69	10.81	10.82
10.87	10.70	10.88
10.60	10.82	10.81

Test the hypotheses that the three methods yield equally variable results ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2$).

7. To evaluate the effects of high levels of copper in their feed, six chicks were fed a standard basal diet to which three levels of copper (0, 400, 800 ppm) were added. The following data show the feed efficiency ratio (g feed/g weight gain) at the end of 3 weeks.

Copper	Chicks					
level	1	2	3	4	5	6
0	1.57	1.54	1.65	1.57	1.59	1.58
400	1.91	1.71	1.55	1.67	1.64	1.67
800	1.88	1.62	1.75	1.97	1.78	2.20

Analyze the above data and test for significant differences of means at $\alpha = 0.05$ level. Fit a response model to the data.

8. The partially completed ANOVA table given here is for a one-way experiment

Source	d.f.	SS	MS	F
Treatments	–	126	–	–
Error	20	–	16	
Total	23	–		

   a. Complete the table.
   b. Give the number of levels for the treatments.
   c. How many observations were collected for each treatment level, that is, replication per treatment?
   d. Test to determine whether the treatment means differ. Use $\alpha = 0.05$.

Chapter 7
Regression Analysis

7.1 Introduction

Sometimes it is desired to build a mathematical or functional relationship between two or more related variables. In addition, it may be of interest to measure the strength of relationship between these variables. The latter topic is referred to as correlation analysis while the former which is used to examine and draw inferences about the functional relationship existing among these variables is called regression analysis.

An example of a relationship that may be of interest is the response of maize to varying amounts of fertilizers. In this example, the response of maize which may be yield per acre is the response or dependent variable while the varying amounts of fertilizers is the independent or explanatory variable.

The simplest form of relationship between two variables is the straight line, and is of the form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{7.1}$$

Here, y is the dependent or *response* variable, x is the independent or *explanatory* variable while β_0 and β_1 are called the *parameters* of the model and are to be estimated from the available data. The ε_i are the random error terms attributable to the observed value y_i, that is, the ith observation. The parameters β_0 and β_1 are also, respectively, the Y-intercept and the slope of the regression line. The error terms satisfy the following assumptions:

(i) $\mathrm{E}(\varepsilon_i) = 0$. That is, the error terms sum to zero.
(ii) $\mathrm{Var}(\varepsilon_i) = \sigma^2$. That is, each error term is distributed with a constant variance of σ^2.
(iii) The error terms are independently distributed normal. This, together with the assumptions in (i) and (ii) above implies that the $\varepsilon_i \sim N(0, \sigma^2)$.

The validity of these assumptions will be examined at a later section in this chapter.

The parameter β_0 in the above equation is interpreted to be the intercept while β_1 is the slope. The significance of β_0 in most experiments like the

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_7,

example above is that it is expected that even when the input of fertilizer is none, some yields are at least expected.

The implication of the above distribution of the error terms ε_i is that, each of the individual observation of the response variable y_i follows the following distribution:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

In practice, n pairs of observations on X and Y are normally available. These are usually denoted by $(x_i, y_i), i = 1, 2, \cdots, n$. With this notation it can be shown that the best line using the method of least squares is given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x} \tag{7.2}$$

Where as usual, $\bar{y} = \dfrac{\sum y_i}{n}$ and $\bar{x} = \dfrac{\sum x}{n}$ are the means of Y and X observations respectively, with,

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x \sum y}{n} \tag{7.3}$$

being the corrected sum of cross-products, and

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x)^2}{n} \tag{7.4}$$

is the corrected sum of squares of X.

7.2 Model Assumptions

For the simple linear model in (7.1), the following are the underlying model assumptions. These assumptions must be verified in the light of available data. Any model violations must be examined and corrections sought. We shall discuss the verification of these assumptions and further discuss measures for addressing any violations.

1. The random error term ε has a mean of 0. That is, $E(\varepsilon_i) = 0$, for all i and hence, $E(Y_i) = \beta_0 + \beta_1 x_i$. This is sometimes denoted as

$$\mu_{Y|X} = \beta_0 + \beta_1 X. \tag{7.5}$$

2. The random error term ε has a constant variance σ^2. That is, $\text{Var}(\varepsilon_i) = \sigma^2$, for all i.

This assumption implies that the variances do not depend on the values of X. In other words, the variance is constant from one observation to another. This homogeneity of variances is often referred to as *homoscedasticity* and therefore $\sigma^2_{Y|X} \equiv \sigma^2$, for all i. Here, $\sigma^2_{Y|X}$ is read as "the conditional variance of Y given X." Further,

$$\text{Var}(\beta_0 + \beta_1 x_1 + \varepsilon_i) = \text{Var}(\varepsilon_i) = \sigma^2. \tag{7.6}$$

3. The random error term ε is distributed normal. That is, combining assumptions (1) and (2), we have $\varepsilon \sim N(0, \sigma^2)$.
4. The error terms are assumed to be uncorrelated. That is,

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad \text{for all } i \neq j. \tag{7.7}$$

A consequence of this is that any two dependent variables Y_i and Y_j are uncorrelated, or simply stated are independently distributed.

7.3 Estimating the Parameters of the Simple Model

Several methods are used to estimate the parameters of the linear regression model, of which the model in (7.1) is a simple linear regression model. Our focus here is to obtain point estimates for the parameters of our model using available data from our random sample. By far the most used method is the *ordinary least squares* (OLS) method developed by Gauss. We list below two of the methods that have been used to estimate the parameters of a regression model:

(a) method of ordinary least squares (OLS)
(b) method of maximum likelihood (MLE)

We now discuss these methods in the following sections.

7.3.1 The Ordinary Least Squares (OLS) Method

For the simple regression equation,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{7.8}$$

the ordinary least squares (least squares for short) method minimizes the sum of squared deviations with respect to the parameters β_0 and β_1. That is, the method sought to minimize

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2, \tag{7.9}$$

with respect to β_0 and β_1. This implies that we need to obtain $\frac{\partial S}{\partial \beta_0}$ and $\frac{\partial S}{\partial \beta_1}$ and set them to zero. Hence,

$$\frac{\partial S}{\partial \beta_0} = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-1), \tag{7.10a}$$

$$\frac{\partial S}{\partial \beta_1} = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-x_i). \tag{7.10b}$$

Setting the equations in (7.10a) and (7.10b) to zero, we have

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0, \tag{7.11a}$$

$$[0.05\ in]\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0. \tag{7.11b}$$

Multiplying Eq. (7.11a) by $\sum x_i$ and Eq. (7.11b) by n, and dropping the subscripts for brevity, we have

$$\sum x_i \sum y_i - n\hat{\beta}_0 \sum x_i - \hat{\beta}_1 \left(\sum x_i\right)^2 = 0, \tag{7.12a}$$

$$n\sum x_i y_i - n\hat{\beta}_0 \sum x_i - n\hat{\beta}_1 \sum x_i^2 = 0. \tag{7.12b}$$

Subtracting Eq. (7.12b) from Eq. (7.12a), we have

$$\sum x_i \sum y_i - \hat{\beta}_1 \left(\sum x_i\right)^2 - n\sum x_i y_i + n\hat{\beta}_1 \sum x_i^2 = 0. \tag{7.13}$$

That is,

$$n\hat{\beta}_1 \left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right] = n\sum x_i y_i - \sum x_i \sum y_i.$$

Dividing through by n, we have,

$$\hat{\beta}_1 \left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right] = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i,$$

and hence,

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}. \tag{7.14}$$

That is,

$$\hat{\beta}_1 = \begin{cases} \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \\ \\ \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ \\ \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \end{cases} \tag{7.15}$$

Writing

$$S_{xy} = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

and

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2,$$

we can therefore write the parameter estimate of β_1 more succinctly as

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}. \tag{7.16}$$

From Eq. (7.8), we have $E(Y_i) = \beta_0 + \beta_1 E(X_i)$. Hence, the parameter estimates of β_0 can be obtained as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{7.17}$$

The estimated regression equation is therefore given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}). \tag{7.18}$$

Example 7.1.1

Data in Table 7.1 represent systolic blood pressure (SBP) and age readings on a sample of 24 individuals of a particular ethnic group. The ages range from 21 to 70 years.

Table 7.1 Observations on age and SBP for 24 individuals

Individual i	Age X	SBP Y	Individual i	Age X	SBP Y	Individual i	Age X	SBP Y
1	34	116	9	46	144	17	47	139
2	26	112	10	53	150	18	42	135
3	51	151	11	29	111	19	61	163
4	58	161	12	50	148	20	38	128
5	34	122	13	40	135	21	57	159
6	40	129	14	34	126	22	66	177
7	31	119	15	67	172	23	42	135
8	57	158	16	23	100	24	53	149

```
MTB > READ C1-C2
DATA> 34 116
DATA> 26 112
DATA> 51 151
.............
DATA> 66 177
DATA> 42 135
DATA> 53 149
DATA> end

MTB > PRINT C1-C2

Data Display


 Row      X      Y

   1     34    116
   2     26    112
   3     51    151
   4     58    161
   5     34    122
   6     40    129
   7     31    119
   8     57    158
   9     46    144
  10     53    150
  11     29    111
  12     50    148
  13     40    135
  14     34    126
  15     67    172
  16     23    100
  17     47    139
  18     42    135
  19     61    163
  20     38    128
  21     57    159
  22     66    177
  23     42    135
  24     53    149
MTB > GStd.
* NOTE  * Character graphs are obsolete.

MTB > Plot 'Y' 'X';
SUBC>   Symbol '+'.
```

```
Plot

      -
    175+                                                              +
      -                                                               +
 Y    -                                                    +
      -                                                2 +
      -
    150+                                          ++ 2
      -                                    +
      -                                     +
      -                           +  2
      -                          + +
    125+                 +
      -              +   +
      -                  +
      -     +  +
      -
    100+  +
       --+---------+---------+---------+---------+---------+----X
        24.0       32.0       40.0       48.0       56.0       64.0

MTB > GPro.
```

For the above data, the scatter plot is presented above and we have the following computations:

Individual	X	Y	X^2	Y^2	XY
1	34	116	1156	13,456	3944
2	26	112	676	12,544	2912
3	51	151	2601	22,801	7701
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
22	66	177	4356	31,329	11,682
23	42	135	1764	18,225	5670
24	53	149	2809	22,201	7897
Sum	1079.0	3339.0	52,119.0	474,053.0	155,921.0

Thus,

$$\sum x = 1079.0, \qquad \sum x^2 = 52119.0, \qquad n = 24$$
$$\sum y = 3339.0, \qquad \sum y^2 = 474053.0, \qquad \sum xy = 155921.0$$

where,

$$\sum x = 34 + 26 + \cdots + 53 = 1079.0$$

$$\sum x^2 = 34^2 + 26^2 + \cdots + 53^2 = 52,119.0.$$

Hence,

$$\bar{x} = \frac{1079.0}{24} = 44.9583, \tag{7.19}$$

and,

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 52,119.0 - \frac{1079.0^2}{24} = 3608.9583 \tag{7.20}$$

Similarly,

$$\sum y = 116 + 112 + \cdots + 149 = 3339.0$$
$$\sum y^2 = 116^2 + 112^2 + \cdots + 149^2 = 474,053.0.$$

Hence,

$$\bar{y} = \frac{3339.0}{24} = 139.1250, \tag{7.21}$$

and,

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 474,053.0 - \frac{3339.0^2}{24} = 9514.6250. \tag{7.22}$$

Also,

$$\sum xy = (34)(116) + (26)(112) + \cdots + (53)(149) = 155,921.0.$$

Hence,

$$S_{xy} = \sum xy - \frac{(\sum x)\,(\sum y)}{n} = 155,921.0 - \frac{(1079.0)(3339.0)}{24} = 5805.1250. \tag{7.23}$$

From the above expressions for the OLS estimates of the parameters, we have

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{5805.1250}{3608.9583} = 1.6085$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x} = 139.1250 - (1.6085 \times 44.9583) = 66.8096.$$

Therefore, the fitted regression equation is

$$\hat{y}_i = 66.8096 + 1.6085\, x_i. \tag{7.24}$$

7.3.2 Interpretations of Parameter Estimates

The estimate of the slope relating SBP with age is 1.6085, which means that for a unit increase in age, SBP increases by 1.6085 units. Similarly at age 0, corresponding to $x = 0$, the SBP would be 66.8096, but $x = 0$ was not within the range of values of x, namely, $23 \leq X \leq 67$ when we built our model. Hence it would be very unwise to predict what would happen when $x = 0$, since this value is not in our sample data. We shall discuss this further later in the text. Usually, we are not too interested in the interpretation of $\hat{\beta}_0$. We can implement the regression model in MINITAB with the following commands and partial output.

```
MTB > Regress 'Y' 1 'X';
SUBC>   Constant;
SUBC>   Brief 2.

Regression Analysis: Y versus X


The regression equation is
Y = 66.8 + 1.61 X

Predictor        Coef     SE Coef          T        P
Constant       66.808       2.200      30.37    0.000
X            1.60853     0.04720      34.08    0.000

S = 2.836       R-Sq = 98.1%     R-Sq(adj) = 98.1%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      9337.7      9337.7   1161.31    0.000
Residual Error    22       176.9         8.0
Total             23      9514.6

Unusual Observations
Obs          X          Y         Fit      SE Fit    Residual     St Resid
  1       34.0    116.000     121.498       0.776      -5.498       -2.02R

R denotes an observation with a large standardized residual
```

7.4 Inferences on Parameter Estimates

In order to make inferences about the estimated parameters, we need to carry out an analysis of variance as follows:

For a general simple linear model:

(i) The Total SS $= S_{yy}$ and is based on $(n-1)$ d.f.

(ii) The Reg SS $= \dfrac{S_{xy}^2}{S_{xx}}$ and is based on 1 d.f.

(iii) Error SS $= S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}$ and is based on $(n-1-1) = (n-2)$ d.f. (obtained by subtraction)

Hence, the ANOVA table becomes:

Source of variation	d.f.	SS	MS	F
Regression	1	$\dfrac{S_{xy}^2}{S_{xx}}$	$\dfrac{S_{xy}^2}{S_{xx}}$	RMS/EMS
Residual (error)	$n-2$	$S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}$	ESS$/(n-2) = S^2$	
Total	$n-1$	S_{yy}		

where ESS is the residual or Error SS, RMS is the regression mean square and S^2 is the error mean square.

For our data example therefore, $S_{yy} = 9514.625$ from previous calculation. The fitted (or explained or regression) sum of squares is defined as

$$\frac{S_{xy}^2}{S_{xx}} = \frac{(5805.1250)^2}{3608.9583} = 9337.7294$$

The residual (Error) SS = Total SS−Fitted SS $= S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}$. Hence, the analysis of variance table is displayed in Table 7.2.

Table 7.2 Analysis of Variance Table

Source of variation	d.f.	SS	MS	F
Regression	1	9337.7294	9337.7294	1161.31**
Residual (error)	22	176.8956	$8.0407 = S^2$	
Total	23	9514.6250		

**Significant at the 0.001 % point

where,

- d.f.: degree of freedom
- SS: sum of Squares
- MS: mean square = SS/corresponding d.f.
- F: The F ratio—(Regression MS)/(Residual MS) is F distributed with 1 and 22 $= (n-)$ degrees of freedom.

The degrees of freedom for F are those of the regression and residual lines in the analysis of variance table. The regression SS is based on 1 d.f. which equals the number of parameters being estimated in the model minus one. That is, Reg d.f. $=$ # number of parameters $- 1 = (2-1) = 1$ in this case,

since only two parameters β_0 and β_1 are being estimated in the example. Note that in general:

$$\text{Total SS} = \text{Regression SS} + \text{Error SS} \quad \text{and, therefore}$$
$$\text{Total d.f.} = \text{Reg d.f.} + \text{Error d.f.}$$

The corresponding MINITAB output for this from the above program is presented below. Our results agree with those obtained from MINITAB.

```
Analysis of Variance

Source            DF          SS          MS          F        P
Regression         1      9337.7      9337.7    1161.31    0.000
Residual Error    22       176.9         8.0
Total             23      9514.6

S = 2.836       R-Sq = 98.1%     R-Sq(adj) = 98.1%
```

We recognize that the residual sum of squares (RSS) agrees to two decimal places with the sum of squares deviations obtained earlier on. In general, these two are always the same.

The significance of the regression is tested by the computed value of F in the analysis of variance table. The hypothesis being tested here is:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0. \tag{7.25}$$

These hypotheses are tested with the computed F value or p value given in the output. The decision rule here is to reject H_0 if $F^* \geq F_{(1,22)}(.975) = 5.79$. Since, $1161.31 \ggg 5.79$, therefore, we would strongly reject H_0 and conclude that the the slope of the linear relationship is not zero. In other words, the explanatory variable X (age) is important in the model. We could also have conducted the above test using the p value. Here the p value is $0.0000 \lll 0.05$, which again leads to strongly rejecting H_0.

In the analysis of variance table, the residual mean square $= S^2$ provides an estimate of the population variance σ^2. The population variance σ^2, may or may not be equal to S^2, the variance about the regression. If the model is true, then $\sigma^2 = S^2$. If the model is not true, then $\sigma^2 < S^2$ and we say that the postulated model is incorrect or suffers from lack of fit.

7.4.1 Confidence Interval for β_0 and β_1

The estimates of β_0 and β_1 obtained from a sample of 24 subjects in Table 7.1 are not fixed, since if we were to take another sample of 24 subjects, we are more than likely to obtain different estimates for β_0 and β_1. Thus, it is very

important to build a level of confidence around our parameter estimates. To do this however, we need to obtain the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$.

The variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by:

$$\text{Var}(\hat{\beta}_0) = \left(\frac{1}{n}\frac{\sum x_i^2}{S_{xx}}\right)\sigma^2$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}.$$

Thus, estimates of these variances are given by,

$$\widehat{\text{Var}}(\hat{\beta}_0) = \left(\frac{1}{n}\frac{\sum x_i^2}{S_{xx}}\right)S^2$$

$$\widehat{\text{Var}}(\hat{\beta}_1) = \left(\frac{1}{S_{xx}}\right)S^2. \tag{7.26}$$

where S^2 equals the error mean square (EMS). Hence, from (7.26), the standard error (s.e.) of $\hat{\beta}_1$ is given by:

$$\text{s.e.}\,(\hat{\beta}_1) = \frac{S}{\sqrt{S_{xx}}} = \sqrt{\frac{8.0407}{3608.9583}} = 0.0472$$

A $100(1-\alpha)\,\%$ confidence interval limits for β_1 is therefore given by:

$$\hat{\beta}_1 \pm t_{\alpha/2}\ \text{s.e.}\ (\hat{\beta}_1) \tag{7.27}$$

where the t value is the 95 % percentage point of a Student's t distribution with $(n-2)$ degrees of freedom, i.e., the number of degrees of freedom on which the estimate S^2 was based (the error d.f. in the ANOVA table).

For the example above, with $\alpha = 0.05$, then $t_{0.975}(22\ \text{d.f.}) = 2.0739$. Hence, a 95 % confidence interval for the slope β_1 is

$$1.6085 \pm 2.0739 \times 0.0472 = 1.6085 \pm 0.098 = (1.511, 1.707). \tag{7.28}$$

Thus with $\alpha = 0.05$, a 95 % confidence interval for the slope β_1 is (1.511, 1.707).

In addition to estimating the slope β_1 from the data, one may also be interested in testing a hypothesis about the value of the slope. In particular, suppose we are interested in testing whether the slope β_1 is some hypothesized value (say, b_1), that is, an hypothesis of the form

$$H_0 : \beta_1 = b_1$$

$$H_1 : \beta_1 \neq b_1. \tag{7.29}$$

To test this hypothesis, we calculate

$$t = \frac{\hat{\beta}_1 - b_1}{\text{s.e.}\,(\hat{\beta}_1)}$$

and compare this with the t distribution on $(n-2)$ degrees of freedom. In particular, if $b_1 = 0$, then for our example above, we have:

$$t = \frac{1.6085}{0.0472} = 34.0784.$$

For this particular case, the test is equivalent to the F test discussed earlier and it is not too difficult to see that $34.0784^2 = 1161.34$ (very close to the earlier calculated F value in the analysis of variance table).

7.4.2 Confidence Interval Estimation for β_0

Similarly, from (7.26), the standard error (s.e.) of $\hat{\beta}_0$ is given by:

$$\text{s.e}\,(\hat{\beta}_0) = \sqrt{\left(\frac{\sum x_i^2}{nS_{xx}}\right) S^2} \qquad = \sqrt{\left(\frac{52119.0}{24 \times 3608.9583}\right) 8.0407} = 2.1996$$

A $100(1-\alpha\,\%)$ confidence interval limits for β_0 is also given by:

$$\hat{\beta}_0 \pm t_{\alpha/2}\ \text{s.e.}\ (\hat{\beta}_0) \tag{7.30}$$

where the t value is the 95 % percentage point of a Student's t distribution with $(n-2)$ degrees of freedom, i.e., the number of degrees of freedom on which the estimate S^2 was based.

From our previous result, if $\alpha = 0.05$, then $t_{0.975}(22\ \text{d.f.}) = 2.0739$. Hence a 95 % confidence interval for the slope β_0 is

$$66.810 \pm 2.0739 \times 2.1996 = 66.810 \pm 4.562 = (62.248, 71.372) \tag{7.31}$$

Thus with $\alpha = 0.05$, a 95 % confidence interval for the slope β_0 is $(62.248, 71.372)$.
Similarly, a hypothesis of the form

$$\begin{aligned} H_0 &: \beta_0 = b_0 \\ H_a &: \beta_0 \neq b_0 \end{aligned} \tag{7.32}$$

can be conducted with the test statistic,

$$t = \frac{\hat{\beta}_0 - b_0}{\text{s.e.}\ (\hat{\beta}_0)}$$

and compare this with the t distribution on $(n-2)$ degrees of freedom. In particular, if $b_0 = 0$, then for our example above, we have:

$$t = \frac{66.8096}{2.1996} = 30.3735.$$

Both of these tests for β_1 and β_0 are provided in the MINITAB output. Here the *SE Coef* refers to the standard error of the coefficient and the T relate to the calculated t statistics for testing each of the hypotheses in (7.29) and (7.32) for the cases when the hypothesized values are zeros. The p values obtained in both cases indicate that $\beta_1 \neq 0$ nor is $\beta_0 = 0$.

```
Predictor       Coef     SE Coef          T        P
Constant      66.808       2.200      30.37    0.000
X            1.60853     0.04720      34.08    0.000
```

7.5 Residuals

The residuals are given by $\hat{e}_i = y_i - \hat{y}_i$ which for the data in Table 7.1, we have for instance, when $x = 34$, $\hat{y} = 66.8081 + 1.6085(34) = 121.497$ and hence, the residual is equal to $116 - 121.497 = -5.497$. Other residuals are computed in the same way and the results are presented in Table 7.3.

Table 7.3 Observed, fitted, and residuals for the data in Table 7.1

Subject	x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$
1	34	116	121.498	−5.49817	30.2299
2	26	112	108.630	3.37009	11.3575
3	51	151	148.843	2.15679	4.6517
4	58	161	160.103	0.89706	0.8047
5	34	122	121.498	0.50183	0.2518
6	40	129	131.149	−2.14936	4.6198
7	31	119	116.673	2.32743	5.4169
8	57	158	158.494	−0.49441	0.2444
9	46	144	140.801	3.19945	10.2365
10	53	150	152.060	−2.06028	4.2447
11	29	111	113.456	−2.45551	6.0295
12	50	148	147.235	0.76532	0.5857
13	40	135	131.149	3.85064	14.8274
14	34	126	121.498	4.50183	20.2665
15	67	172	174.580	−2.57973	6.6550
16	23	100	103.804	−3.80432	14.4728
17	47	139	142.409	−3.40909	11.6219
18	42	135	134.366	0.63357	0.4014
19	61	163	164.929	−1.92853	3.7192
20	38	128	127.932	0.06770	0.0046
21	57	159	158.494	0.50559	0.2556
22	66	177	172.971	4.02881	16.2313
23	42	135	134.366	0.63357	0.4014
24	53	149	152.060	−3.06028	9.3653
$\sum$				0.0000	176.8956

We note here that the sum of squared deviations $\sum_{i=1}^{24}(y_i - \hat{y}_i)^2 = \sum \hat{e}_i^2 =$ 176.8956 and that the sum of deviations $\sum_{i=1}^{24}(y_i - \hat{y}_i) = \sum \hat{e}_i = 0.0000$. The former equals the Error sum of squares obtained in the analysis of variance table displayed earlier.

7.6 Prediction of Y from X

In many regression problems, the purpose is to predict Y from knowledge of the corresponding X. The predicted value of y for a given value of x say x_c is $\hat{y}_c = \hat{\beta}_0 + \hat{\beta}_1 x_c$.

However, it is important to draw a clear distinction between predicting for an individual value of x_c or predicting for a mean value of x_c. Let us explain this in detail. In our example, we have observations on age and SBP from a random sample of 24 subjects. Consider the case when age $= x_c = 34$. Do we want to predict the SBP for an individual in the population whose age is 34? Or do we want to predict for all individuals in the population whose ages are 34 (there are certainly several people with age 34 in any human population). The above questions inform on the type of analysis that we would have to employ. Confidence levels obtained with the former are often called *fiducial confidence intervals*, while predictions based on the latter are often referred to as mean predictions with corresponding mean prediction confidence intervals.

7.6.1 Mean Prediction

In our example, the predicted SBP for all subjects that are 34 years old would be $66.808 + 1.6085(34) = 121.497$. The standard error of the predicted value is given by

$$\text{s.e. } (\hat{y}_c) = \sqrt{S^2 \left\{ \frac{1}{n} + \frac{(x_c - \bar{x})^2}{S_{xx}} \right\}}. \tag{7.33}$$

From the expression in (7.33), relating SBP and age, the predicted mean SBP for a large population of subjects who are all 34 years old is 121.497 and the standard error of this predicted mean is

$$\sqrt{8.0407 \left\{ \frac{1}{24} + \frac{(34 - 44.9583)^2}{3608.9583} \right\}} = 0.7763.$$

The s.e. above applies to the predicted mean value of y for a given $x = x_c$.

7.6.2 Individual Prediction

The prediction of y for an individual new member of the population for which y has been measured has a standard error given by

$$\text{s.e. of (individual member)} = \sqrt{S^2\left\{1 + \frac{1}{n} + \frac{(x_c - \bar{x})^2}{S_{xx}}\right\}}. \tag{7.34}$$

This prediction is by far the most frequent in experimental data analysis. The SBP of a single individual who is 34 years old is still 121.497 but with a s.e. given by

$$\sqrt{8.0407\left\{1 + \frac{1}{24} + \frac{(34 - 44.9583)^2}{3608.9583}\right\}} = 2.9399.$$

A 95 % confidence interval for the mean predicted SBP is given by

$$\begin{aligned} 121.497 \pm t_{0.975} \times \text{s.e.} &= 121.497 \pm 2.0739 \times 0.7763 = 121.497 \pm 1.610 \\ &= (119.887, 123.107) \end{aligned}$$

while that for a predicted individual SBP is given by

$$121.497 \pm 2.0739 \times 2.9399 = 121.497 \pm 6.097 = (115.400, 127.594)$$

Both of these for $x_c = 34$ are implemented in MINITAB with the following, together with a partial output. Our manually calculated results are very close to those obtained from MINITAB.

```
MTB > Regress 'Y' 1 'X';
SUBC>   Constant;
SUBC>   Predict 34.

Predicted Values for New Observations

New Obs     Fit     SE Fit         95.0% CI              95.0% PI
1       121.498      0.776   ( 119.888, 123.108)   ( 115.401, 127.595)

Values of Predictors for New Observations

New Obs        X
1           34.0
```

We present in Fig. 7.1 the plot of the predicted regression line for the data in Table 7.1.

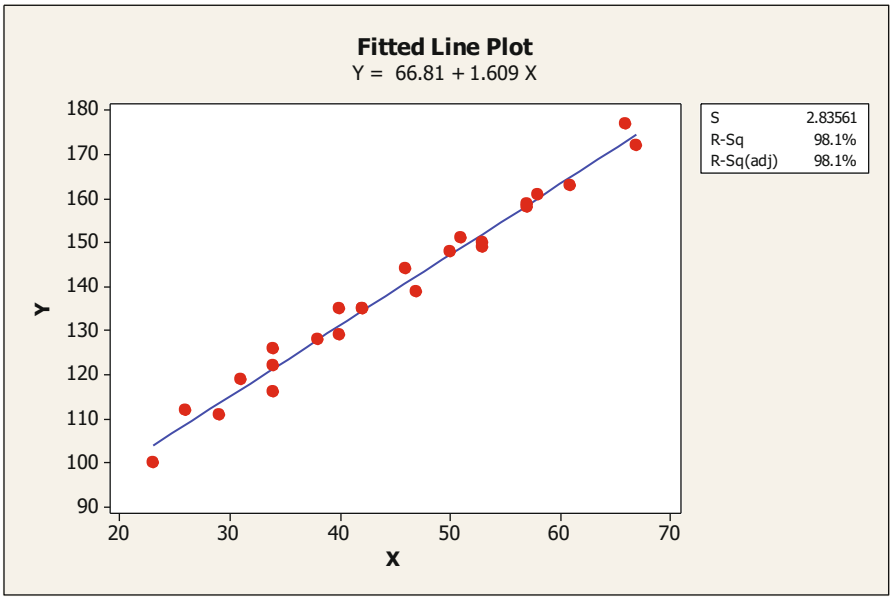


Fig. 7.1 Plot of predicted equation

A word of caution when predicting in regression analysis. we must not try to predict for values of x that are not in the sample data. For example, in our data example here, we have the range of X is $23 \leq x \leq 67$. What this means is that we can not predict values of Y for $x < 23$ or $x > 67$. Such a result will lead to interpolation as we do not know whether the form of the regression equation will still be as estimated outside these sample values.

7.6.3 Percentage Variation

The percentage variation is given by

$$R^2 = \frac{\text{SS due to regression}}{\text{Total SS}} \times 100\,\% = \frac{S_{xy}^2}{S_{xx}\,S_{yy}} \times 100\,\%. \tag{7.35}$$

That is, it is the proportion of the total variation about the mean $\bar{y}$ explained by the regression. For our data, this is equal to

$$\frac{9337.7294}{9514.625} \times 100 = 98.14\,\%.$$

We shall see in Sect. 7.8 how this is related to the sample correlation coefficient. The R^2 obtained in this example is very high indicating that our model might be very good. We will further examine the adequacy of this model in Sect. 7.7.

7.7 Adequacy of the Regression Model

The fitted regression line is calculated in the previous section on the assumption that the true relationship between y and x is of the form $y = \beta_0 + \beta_1 x$. This is an assumption we should not blindly accept but should tentatively entertain. Two procedures that are commonly used to examine the adequacy of a regression model will be discussed here.

The first of these is when genuine replicated observations are available. Consider the data in Table 7.1, we observe here that certain values of x are repeated, e.g., $x = 34$ (three times), 40 (two times) and so on. These repeated observations and their corresponding response values (y_i) are presented in Table 7.4. These are obtained after employing the following MINITAB commands to sort variable age and to carry along with it, the corresponding values of SBP.

```
MTB > sort c1 c3
MTB > sort c2 c4;
SUBC> by c3.
```

Table 7.4 Replication observations and their corresponding response values

x_i	y_i's
34	116, 122, 126
40	129, 135
42	135, 135
53	150, 149
57	158, 159

For the data in Table 7.1, therefore, it would be possible to conduct a test of lack of fit by breaking the Error SS into two components:

(a) Lack of Fit SS
(b) Pure Error SS

We proceed with the calculations to conduct a test of adequacy, by first calculating the *Pure Error SS* at these replicated points, viz.:

Calculation of Pure Error SS

$$\text{At } x = 34,\ \text{SS} = 116^2 + 122^2 + 126^2 - \frac{(116 + 122 + 126)^2}{3} = 50.6667 \quad \text{on 2 d.f.}$$

$$\text{At } x = 40,\ \text{SS} = 129^2 + 135^2 - \frac{(129 + 135)^2}{2} = 18.00 \quad \text{on 1 d.f.}$$

$$\text{At } x = 42,\ \text{SS} = 135^2 + 135^2 - \frac{(135 + 135)^2}{2} = 0.00 \quad \text{on 1 d.f.}$$

$$\text{At } x = 53,\ \text{SS} = 150^2 + 149^2 - \frac{(150 + 149)^2}{2} = 0.50 \quad \text{on 1 d.f.}$$

$$\text{At } x = 57,\ \text{SS} = 158^2 + 159^2 - \frac{(158 + 159)^2}{2} = 0.50 \quad \text{on 1 d.f.}$$

Note that the SS at $x = 40$ can also be computed as $\frac{(129-135)^2}{2} = 18.00$. Hence, Total Pure Error SS $= 50.6667 + 18.0 + 0.0 + 0.5 + 0.5 = 69.6667$ and is based on $2 + 1 + 1 + 1 + 1 = 6$ degrees of freedom. We therefore have the revised analysis of variance for these data in Table 7.5.

Table 7.5 Revised analysis of variance table

Source of variation	d.f.	SS	MS	F
Reg.	1	9337.7294	9337.7294	1161.31**
Residual (error)	22	176.8956	8.0407=S^2	
Lack of fit	16	107.2289	6.7018	0.577
Pure error	6	69.6667	11.6111=S_e^2	
Total	23	9514.6250		

Since the lack of fit is not significant, there is therefore no reason to doubt the adequacy of the model. We would therefore still use the error mean square S^2 to carry out the F test and to obtain confidence limits, etc. However with a significant lack of fit, it means that the proposed model is not adequate and we seek ways to improve the model by examining the residuals. Any test of significance carried out with this type of model will not be valid. We can implement the lack of fit test in MINITAB with the following statements and corresponding partial output.

```
MTB > Regress 'Y' 1 'X';
SUBC>   Constant;
SUBC>   Pure;
SUBC>   Brief 3.

Regression Analysis: Y versus X


Analysis of Variance

Source            DF          SS          MS          F        P
Regression         1      9337.7      9337.7    1161.31    0.000
Residual Error    22       176.9         8.0
  Lack of Fit     16       107.2         6.7       0.58    0.823
  Pure Error       6        69.7        11.6
Total             23      9514.6
```

The results we see are identical with those calculated manually.

7.7.1 Examinations of Residuals

The residuals are defined as the n differences $\hat{\varepsilon}_i = y_i - \hat{y}_i, i = 1, 2, \cdots, n$ where y_i is an observed value and $\hat{y}_i$ is the corresponding fitted value obtained from the fitted regression equation.

In performing the regression analysis, we have assumed the following:

(i) Constant variance from one observation to the other.
(ii) The errors (residuals) are independently distributed randomly with mean zero and constant variance σ^2.
(iii) The errors are individually distributed normal.

Any of these assumptions may be violated in addition to misspecification of the model. We now examine the various techniques for the examination of residuals. These techniques are all graphical. This is usually accomplished by plotting the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$ against the y_i, x_i or $\hat{y}_i$. The shape of the plot could assume any of the following:

The plot in Fig. 7.2 indicates that the residuals are randomly distributed about its mean of zero, except for one outlier whose plot is too far away from the horizontal reference line at zero. This observation would need to be examined further.

Several plots are displayed in Fig. 7.3. However, from the plot of the residuals versus fitted values, there does not seem to be any pattern to the distribution of the residuals about its mean of 0 except for the single observation whose plot is not consistent with those of the other 23 observations. The normal probability plot of the residuals in Fig. 7.3 indicates that we can reasonably assume that the residuals and hence the error terms follow the normal distribution. A formal test of this is presented in Fig. 7.4, which gives

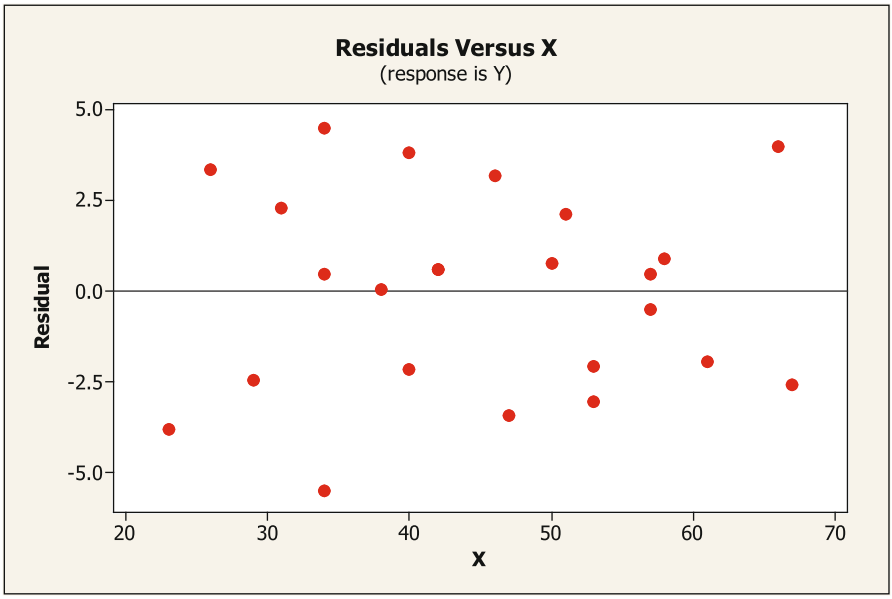


Fig. 7.2 Plot of residuals versus x_i

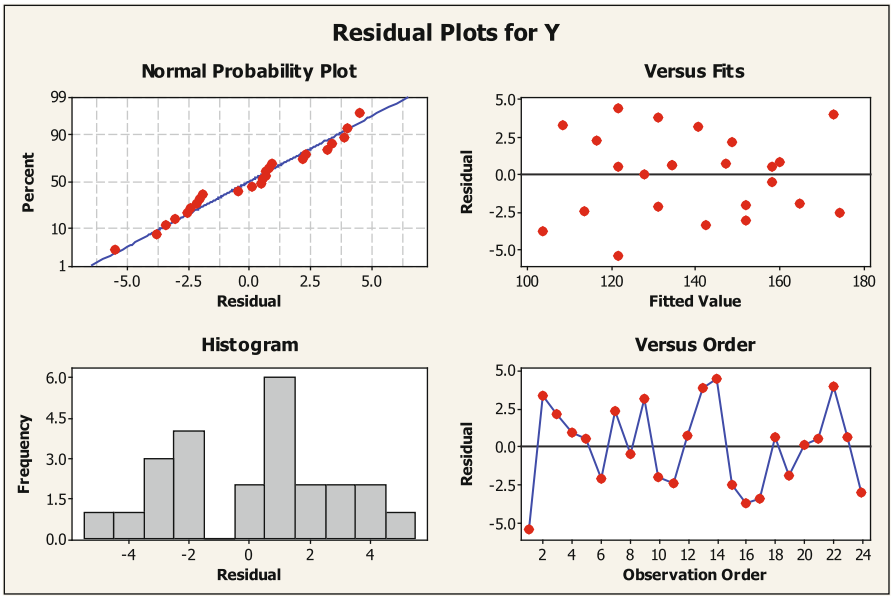


Fig. 7.3 Various residual plots for the data example

the Anderson–Darling test p value of 0.410, indicating that the residuals indeed can be assumed to follow the normal probability distribution. Thus, the normality assumption is not violated in this example.

In the graphical diagnostic plots displayed from Figs. 7.2 to 7.8, we have:

(i) The horizontal band displayed in Fig. 7.2 indicates no abnormality and our least squares analysis would appear to be invalidated.

(ii) If the variance is not constant as assumed; there is a need for weighted least squares or a transformation on the observed y_i before making a regression analysis. A plot of the residuals against either fitted values or the explanatory variable will indicate this as in Fig. 7.5.

(iii) If there is an error in the analysis, a linear term in time should have been included in the model, i.e., an β_0 term.

(iv) Linear and quadratic terms in time should have been included in the model. There is need for extra terms in the model (e.g., square or cross products terms) or need for a transformation on the observations y_i before analysis. Again, a plot of the residuals against the explanatory variable will indicate this as in Fig. 7.6.

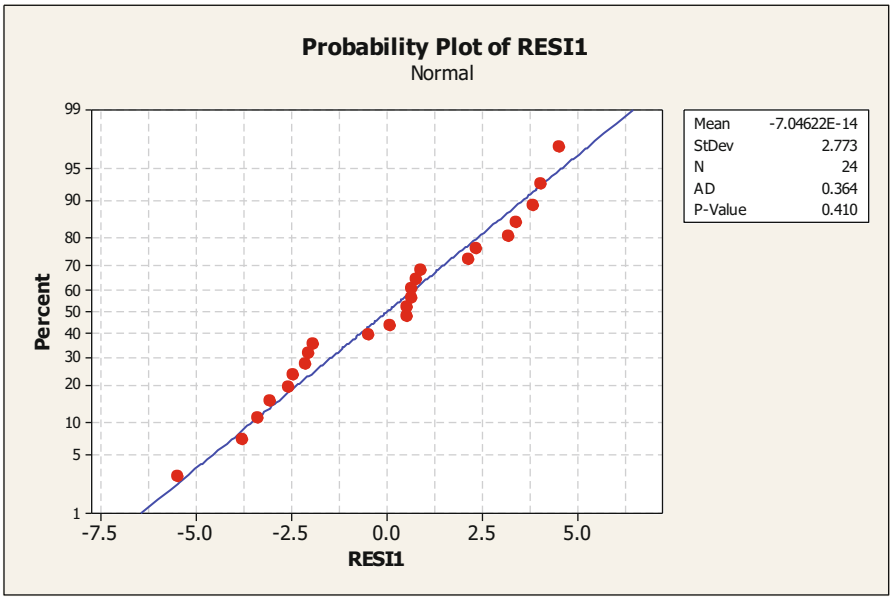


Fig. 7.4 Normal probability plot of the residuals

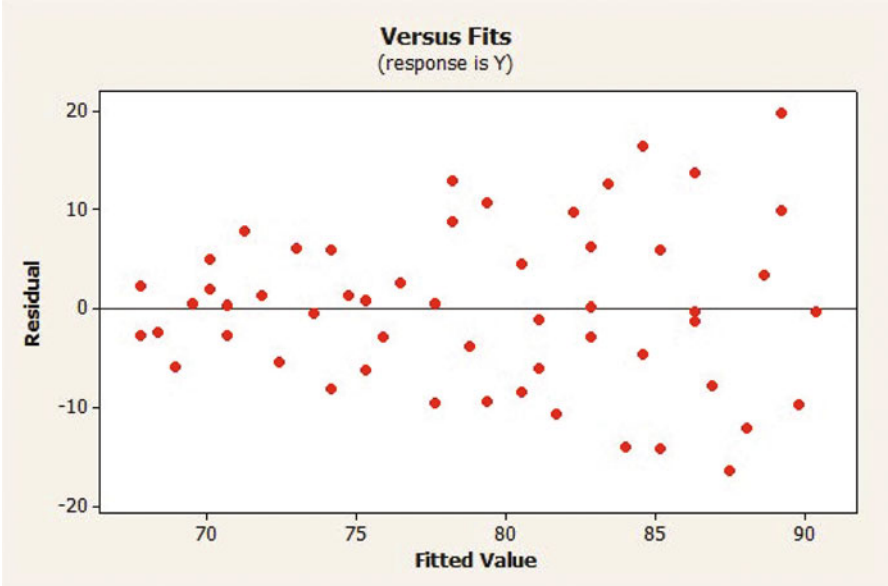


Fig. 7.5 Error variance increasing with $\hat{Y}$

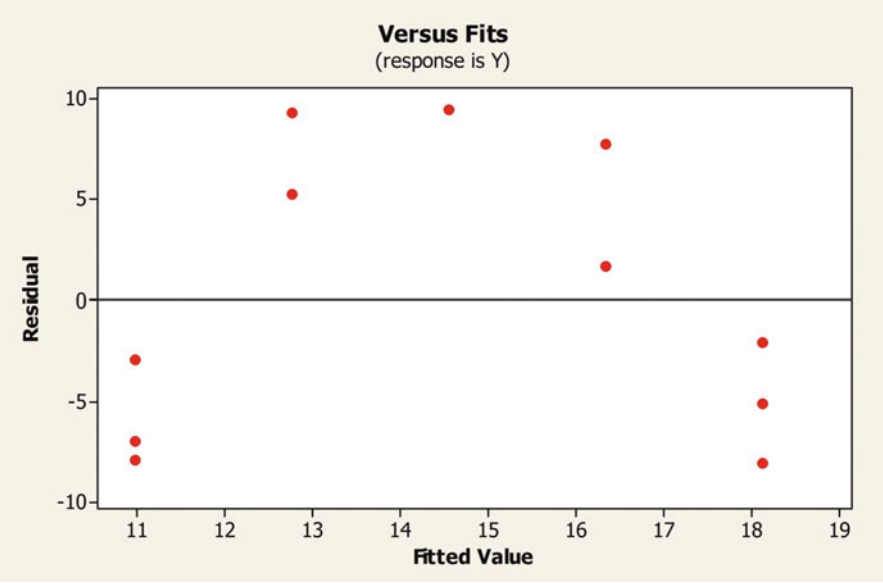


Fig. 7.6 Data departing from linearity

Another very important invalidation of a regression model is if the errors are serially correlated (i.e., not independent). In this case, the Durbin–Watson test statistic provides a good guide for this inadequacy.

In our example, we plotted the standardized residuals, first against the explanatory variable X in Fig. 7.7 and then against the predicted values in Fig. 7.8. Both indicate that there is a single observation with its standardized residual outside the interval $[-2, +2]$. This observation is numbered 1, and all such other observations are usually flagged by MINITAB. In our case, MINITAB warns us with the following statement:

```
Unusual Observations
Obs          X          Y         Fit      SE Fit    Residual     St Resid
  1       34.0    116.000     121.498       0.776      -5.498       -2.02R

R denotes an observation with a large standardized residual
```

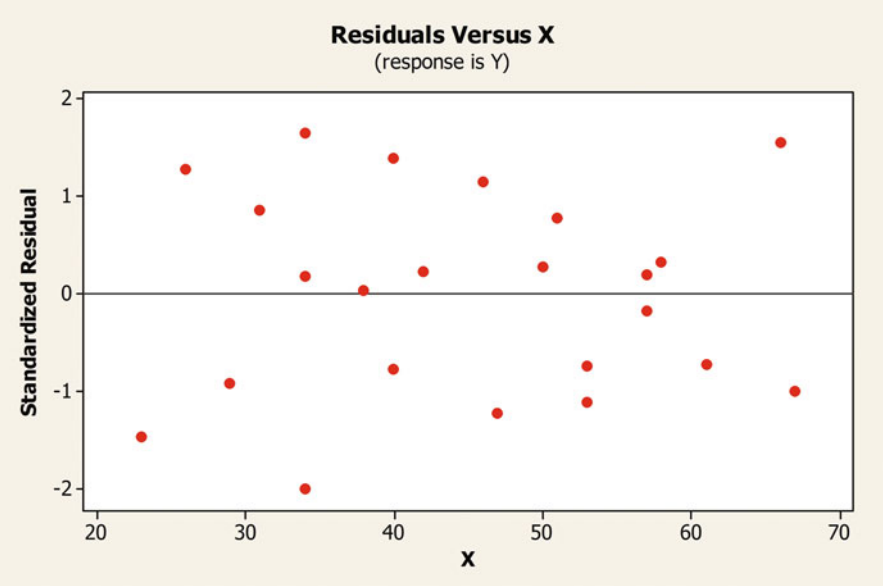


Fig. 7.7 Plot of studentized residuals versus x_i

It is often suggested that any observation with a studentized residual $|r_i| > 2$ should be considered a possible outlier. However, in this example the value of $|r_1| = 2.02$ indicates a violation not too serious. This value is not particularly high to dissuade us from possibly fitting a new model with this observation removed or from reverting to some transformation of the dependent variable.

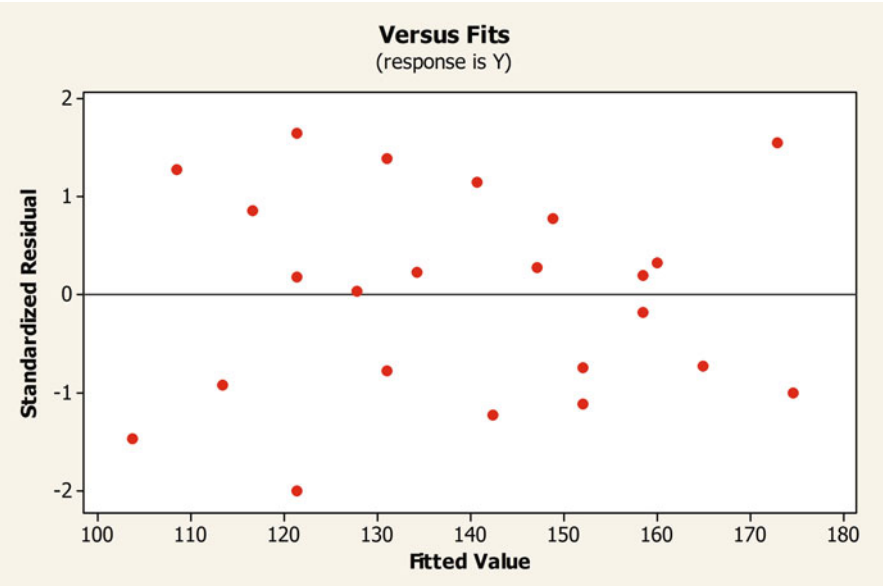


Fig. 7.8 Plot of studentized residuals versus $\hat{y}_i$

7.8 Correlation Coefficient

Correlation is a measure of the strength of linearity between two variables X and Y. The sample correlation coefficient is defined as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \tag{7.36}$$

From the above definition of r in (7.36), we have

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} \quad \text{that is,}$$

$$r^2 S_{yy} = \frac{S_{xy}^2}{S_{xx}}.$$

From our results in the previous sections, we showed that,

$$\text{Residual SS} = \text{Total SS} - \text{Fitted SS}$$

That is,

$$\text{Residual SS} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = S_{yy} - r^2 S_{yy} = S_{yy}(1 - r^2).$$

That is,

$$\frac{\text{Residual SS}}{\text{Total SS}} = 1 - r^2.$$

Thus r^2 is the estimated proportion of the variance of y that can be attributed to the linear regression while $(1 - r^2)$ is the proportion free from x. We see that the square of the correlation coefficient is equal to the R^2, the coefficient of determination computed earlier on.

7.8.1 Properties of r

(i) r must lie between -1 and $+1$, that is, $-1 \leq r \leq 1$ where -1 represents perfect negative linear association in the sample and $+1$ represents perfect positive linear association in the sample.
(ii) Its numerical strength measures the strength of the linear relationship and the sign of r indicates the direction of the relationship.
(iii) r^2 is the proportion of variation in y accounted for by the fitting of the straight line.

For our data,

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{(5805.1250)^2}{3608.9583 \times 9514.6250} = 0.9814$$

Hence, $r = \sqrt{0.9814} = 0.9907$. r^2 is sometimes referred to as the coefficient of determination while $\sqrt{(1-r^2)}$ is similarly called the coefficient of alienation.

The population correlation coefficient is denoted by ρ. To test the hypothesis

$$H_0 : \rho = 0 \quad \text{vs} \quad H_a : \rho \neq 0$$

we calculate the test statistic,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{7.37}$$

and reject H_0 if $|t| \geq t_{1-\alpha/2}$ where $t_{1-\alpha/2}$ is a Student's t distribution with $(n-2)$ d.f. at $\alpha = 0.05$.

In our example above, to test the hypothesis $H_0 : \rho = 0 \quad \text{vs.} \quad H_1 : \rho \neq 0$, we compute

$$t = \frac{0.9907\sqrt{24}}{\sqrt{1-0.9907^2}} = \frac{4.8534}{\sqrt{0.1361}} = 35.6605.$$

Since $t > t_{1-\alpha/2} = 2.0739$, i.e., we reject H_0 and conclude that $\rho \neq 0$. That is, there is a very strong linear relationship between the age and SBP.

7.8.2 General Hypotheses Concerning ρ

To test a more general hypothesis

$$H_0 : \rho = \rho_0 \quad \text{vs} \quad H_a : \rho \neq \rho_0$$

a large sample test is based on the fact that

$$\begin{aligned} Z &= \sqrt{n-3}\left\{\frac{1}{2}\log\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\log\left(\frac{1+\rho_0}{1-\rho_0}\right)\right\} \sim N(0,1) \\ &= \sqrt{n-3}\frac{1}{2}\log\left(\frac{1+r}{1-r} \times \frac{1-\rho_0}{1+\rho_0}\right). \end{aligned}$$

That is,

$$Z = \sqrt{n-3}\frac{1}{2}\log\left\{\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right\} \tag{7.38}$$

is distributed as a standard normal variate. That is, $Z \sim N(0,1)$. Given a two-sided alternative and $\alpha = 0.05$, H_0 is rejected if $|Z| > 1.96$.

For a null hypothesis for the data above, $H_0 : \rho = 0.90$ vs. $H_a : \rho \neq 0.90$ for instance, we have

$$Z = \sqrt{(24-3)}\frac{1}{2}\log\left\{\frac{(1+0.9907)(1-0.90)}{(1-0.9907)(1+0.90)}\right\} = 4.5826\left[\frac{1}{2}\log\left(\frac{0.1991}{0.0177}\right)\right]$$

$$= \left(\frac{4.5826}{2}\right)\log(11.2486) \qquad = 5.546$$

Since $|Z| > 1.96$, i.e., we would reject H_0 and conclude that r is significantly different from 0.90.

7.9 Multiple Regression

In the previous sections, we assumed a simple linear model involving only one independent variable. We did not consider models involving several independent variables. Situations in which we allow y to depend on more than one variable give rise to multiple regression, when we have p independent variables. The model can be written as

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon_i. \tag{7.39}$$

The parameters of this model are $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ while $x_1, x_2, \cdots, x_p$ are the factors supposedly influencing the effect of y (that is, explanatory variables) and the ε_i are the random error terms.

Of course each of the x's can take quadratic, cubic, quartic, ... , forms. For example, the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon_i$$

can be rewritten as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i \quad \text{where } x_2 = x_1^2.$$

This model we recognize as a quadratic model. We shall illustrate the problem of the multiple regression by the simplest multiple regression model, that is the model involving only two independent or explanatory variables. Suppose the model is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon. \tag{7.40}$$

Suppose n sets of observations are available on Y, X_1, and X_2. Then by the method of least squares,

$$b_1 = \hat{\beta}_1 = \frac{S_{x_1 y} S_{x_2} - S_{x_2 y} S_{x_1 x_2}}{S_{x_1} S_{x_2} - (S_{x_1 x_2})^2} \tag{7.41a}$$

$$b_2 = \hat{\beta}_2 = \frac{S_{x_2y}S_{x_1} - S_{x_1y}S_{x_1x_2}}{S_{x_1}S_{x_2} - (S_{x_1x_2})^2} \tag{7.41b}$$

$$b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 \tag{7.41c}$$

where

$$S_{x_1} = \sum x_1^2 - \frac{(\sum x_1)^2}{n}, \qquad S_{x_2} = \sum x_2^2 - \frac{(\sum x_2)^2}{n}$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}, \qquad S_{x_1y} = \sum x_1y - \frac{\sum x_1 \sum y}{n}$$

$$S_{x_2y} = \sum x_2y - \frac{\sum x_2 \sum y}{n}, \qquad S_{x_1x_2} = \sum x_1x_2 - \frac{\sum x_1 \sum x_2}{n}.$$

The estimated regression equation is given by

$$\hat{y}_i = b_0 + b_1x_{1i} + b_2x_{2i} \tag{7.42}$$

7.9.1 Example

An examination of corn plants on various soils have the concentration of inorganic (x_1) and organic (x_2) phosphorus in the soils and the phosphorus content y on corn grown in the soils were measured for 17 soils. Table 7.6 gives the results of this examination.

Table 7.6 Data for this example on multiple regression

Soil sample	x_1	x_2	y	Soil sample	x_1	x_2	y
1	0.4	53	64	9	11.6	29	93
2	0.4	23	60	10	12.6	58	51
3	3.1	19	71	11	10.9	37	76
4	0.6	34	61	12	23.1	46	96
5	4.7	24	54	13	23.1	50	77
6	1.7	65	77	14	21.6	44	93
7	9.4	44	81	15	23.1	56	95
8	10.1	31	93	16	1.9	36	54
				17	29.9	51	99

The summary statistics for the data in Table 7.6 are presented in the following with $n = 17$.

$$\sum x_1 = 188.20, \qquad \sum x_2 = 700.00, \qquad \sum y = 1292.00$$

$$\sum x_1^2 = 2341.1376, \qquad \sum x_2^2 = 31,712.0, \qquad \sum y^2 = 103,075$$

$$\sum x_1x_2 = 8585.112, \qquad \sum x_1y = 16,203.812, \qquad \sum x_2y = 54,081.029$$

$$\bar{x}_1 = 11.07, \qquad \bar{x}_2 = 41.18, \qquad \bar{y} = 76.18.$$

Hence,

$$S_{yy} = 4426.5, \qquad S_{x_1} = 1519.30, \qquad S_{x_2} = 2888.50$$
$$S_{x_1 y} = 1867.40, \qquad S_{x_2 y} = 757.50, \qquad S_{x_1 x_2} = 835.70$$

If we define D to be the denominator in expressions (7.41a) and (7.41b), then we have,

$$\begin{aligned} \mathrm{D} &= S_{x_1} S_{x_2} - (S_{x_1 x_2})^2 \\ &= (1519.30)(2888.50) - (835.70)^2 \\ &= 3690.104 \end{aligned}$$

Hence,

$$\begin{aligned} b_1 &= \frac{[(1867.4)(2888.5) - (757.5)(835.7)]}{D} = 1.2902 \\ b_2 &= \frac{[(757.5)(1519.3) - (1867.4)(835.7)]}{D} = -0.1110 \\ b_0 &= \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 \\ &= 76.18 - (1.2902)(11.07) + (0.1110)(41.18) = 66.47 \end{aligned}$$

The estimated regression equation is therefore given by:

$$\hat{y} = 66.47 + 1.2902 x_1 - 0.1110 x_2. \tag{7.43}$$

Thus for soil sample (1) where $X_1 = 0.4$, and $X_2 = 53$, we have,

$$\hat{y} = 66.47 + 1.2902(0.4) - (0.1110)(53) = 61.1.$$

The regression sum of squares is computed as:

$$\begin{aligned} \text{Reg. SS} &= b_1 S_{x_1 y} + b_2 S_{x_2 y} \\ &= 1.2902(1867.4) + (-0.1110)(757.5) \\ &= 2325.237 \\ \text{Residual SS} &= \text{Total SS} - \text{Regression SS} \\ &= S_{yy} - \text{Fitted SS} \\ &= 4426.5 - 2325.237 \\ &= 2101.3. \end{aligned}$$

The analysis of variance table is therefore given in Table 7.7.

Table 7.7 Regression analysis of variance table

Source	d.f.	SS	MS	F
Regression	2	2325.237	1162.62	7.75*
Residual	14	2101.3	$150.09 = S^2$	
Total	16	4426.5		

*Significant at $\alpha = 0.01$

Hence, we reject the null hypothesis that

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{vs} \quad H_a : \text{ at least one of the } \beta s \text{ is not zero.}$$

In other words, at least one of the parameters is not zero. Again,

$$R^2 = \frac{\text{Fitted SS}}{\text{Total SS}} = \frac{2325.237}{4426.50} = 0.525$$
$$R = \sqrt{R^2} = 0.725.$$

While R^2 is called the coefficient of multiple determination, R on the other hand is called the multiple correlation coefficient between Y and the X's. It can be shown that

$$F = \frac{(n-k)\, R^2}{(k-1)(1-R^2)}$$

where k is the number of parameters in the model. Thus F supplies a test of the significance of the multiple correlation coefficient. The adjusted R^2 can be simply computed with the following expression:

$$\text{Adj. } R^2 = 1 - \frac{\text{MSE}}{\text{Total MS}} = 1 - \frac{150.09}{4426.5/16} = 1 - 0.5425 = 0.4575$$

The multiple regression procedure is implemented in MINITAB with the following. Here, x_1, x_2, and y are read into columns C1, C2, and C3, respectively.

```
MTB > REGRESS C3 2 C1 C2;
SUBC>   Constant;
SUBC>   Brief 3.

Regression Analysis: Y versus X1, X2


The regression equation is
Y = 66.5 + 1.29 x1 - 0.111 x2

Predictor        Coef     SE Coef          T        P
Constant       66.465       9.850       6.75    0.000
x1             1.2902      0.3428       3.76    0.002
x2            -0.1110      0.2486      -0.45    0.662

S = 12.25       R-Sq = 52.5%     R-Sq(adj) = 45.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      2325.2      1162.6      7.75    0.005
Residual Error    14      2101.3       150.1
Total             16      4426.5

Source       DF      Seq SS
x1            1      2295.2
x2            1        29.9

Unusual Observations
Obs         x1          y         Fit      SE Fit    Residual     St Resid
 10       12.6      51.00       76.28        4.98      -25.28       -2.26R

R denotes an observation with a large standardized residual
```

The results we obtained from manual calculations agree with those obtained from MINITAB. Note that the regression degrees of freedom is 2 since we are now estimating three parameters. Hence, d.f. equals $(3 - 1) = 2$ in this case.

7.9.2 Partial F tests

Sometimes we may be interested in such questions as whether it was worth while or economically reasonable to include both dependent variables x_1 and x_2 in the last problem. Would it be reasonable to collect data on variable x_2 if for instance only variable x_1 is sufficient to predict y. In order to answer these and similar questions, we need to conduct what is known as partial F test.

For the problem above, we have

$$\text{Fitted SS} = 2325.237$$

$$\text{Total SS} = 4426.5.$$

If the model $y = \beta_{01} + \beta_1 x_1 + \varepsilon$ were fitted, the fitted SS due to fitting x_1 would be:

$$\frac{(S_{x_1 y})^2}{S_{x_1}} = \frac{1867.4^2}{1519.3} = 2295.256.$$

Similarly if only x_2 had been in the model, viz., $y = \beta_{02} + \beta_2 x_2 + \varepsilon$, then fitted SS due to β_2 would be:

$$\frac{(S_{x_2 y})^2}{S_{x_2}} = \frac{755.5^2}{2888.5} = 198.652$$

Revised analysis of variance table

Source		d.f.	SS	MS	F
(1)	Full regression	2	2325.237	150.9	
(2)	Regression of x_1	1	2295.256	2295.23	15.29**
(3)	Regression of x_2	1	198.652	198.65	1.32
(4)	Total variation	16	4426.5		
(1) − (2)	Extra SS due to fitting $x_2 \mid X_1$	1	30.001	30.00	0.20
(1) − (3)	Extra SS due to fitting $x_1 \mid X_2$	1	2126.611	2126.61	14.17**

We see from the above table that the inclusion of x_2 in the model given that x_1 is already in the equation is not at all significant (F value = 0.20, p value = 0.3384) indicating that once x_1 is already in the model, adding x_2 does not contribute much to the model. On the other hand, the reverse is not true. That is, if x_2 is first fitted, the additional SS due to x_1 given that x_2 is already in the model is highly significant. Then for the above data a simple model of the form $y = \beta_0 + \beta x_1 + \varepsilon$ will suffice, that is,

$$\hat{\beta} = \frac{S_{x_1 y}}{S_{x_1}} = 1.229$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x} = 76.18 - 1.229(11.07)$$
$$= 62.57.$$

That is, the estimated equation in this case would be:

$$\hat{y} = 62.57 + 1.229 x_1. \tag{7.44}$$

We may note here that MINITAB will generate the sequential SS due to each variables as they enter the regression model. In this example, MINITAB gives these sequential SS as:

```
Source       DF      Seq SS
x1            1      2295.2
x2            1        29.9
```

The above means that sequentially, SS($x_1 = 2295.2$) while SS($x_2|x_1$) = 29.9. We can similarly obtain the SS(x_2) and the SS($x_1|x_2$) if we reverse the order of entry of the variables in a new regression procedure. Some other programs avoid this by giving us what is called the type II sum of squares. Unfortunately, MINITAB does not automatically generate these.

7.10 Outliers and Influential Observations

A regression analysis may be inadequate if any of the following occur:

- The relationship is curvilinear
- Presence of outliers
- Presence of influential observations

A data point can be unusual in a regression analysis if the point is an outlier (is unusual in a vertical direction); one that is unusual in the horizontal direction is called a high leverage point. An observation can be both an outlier and have a high leverage point. When an outlier is present in a data set, then the error mean squares are usually inflated, thus reducing the correlation between the explanatory variable(s) and the dependent variable. It may also unduly influence the estimated regression line.

The leverage of an observation denoted by h_i is used to identify influential observations. A good rule of thumb identifies an observation to be influential if its leverage h_i is such that:

$$h_i > \frac{3(k+1)}{n} \tag{7.45}$$

where k is the number of explanatory variables in the model, and n is the number of observations in the data.

A more robust overall measure of influential observations is either Cook's distance measure or the *Dffits*. The latter gives the difference between the predicted value when all the observations are used in the regression, and when the ith observation is deleted. Both statistics are provided in a MINITAB regression analysis. The bound for Cook's distance is $\frac{1}{n}$ and a value of D greater than the 95th percentile of an F distribution with $(k+1)$ and $n-(k+1)$ degrees of freedom is considered an unusual observation. Consider a multiple regression model having three explanatory variables x_1, x_2 and x_3 with $n = 30$ observations on each variable and the response variable y. The results of analysis from the MINITAB application is presented below.

7.10.1 Example

```
 Regression Analysis: y versus x1, x2, x3

The regression equation is
y = 1.81 - 0.531 x1 - 0.440 x2 + 0.209 x3


Predictor      Coef  SE Coef      T      P    VIF
Constant     1.8110   0.2795   6.48  0.000
x1         -0.53146  0.06958  -7.64  0.000  1.046
x2         -0.43964  0.07304  -6.02  0.000  1.243
x3          0.20898  0.04064   5.14  0.000  1.199


S = 0.213468   R-Sq = 82.3%   R-Sq(adj) = 80.2%

PRESS = 1.51475   R-Sq(pred) = 77.36%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       3  5.5047  1.8349  40.27  0.000
Residual Error  26  1.1848  0.0456
Total           29  6.6895


Source  DF  Seq SS
x1       1  3.4460
x2       1  0.8538
x3       1  1.2049


Unusual Observations

Obs    x1       y     Fit  SE Fit  Residual   St Resid
  3  3.34  0.3800  0.7973  0.0544   -0.4173     -2.02R
 10  3.10  0.7700  1.1695  0.0783   -0.3995     -2.01R

R denotes an observation with a large standardized residual


MTB > print c1-c7

Data Display

Row    x1    x2    x3     y       HI1      COOK1      DFIT1
  1  3.05  1.45  5.67  0.34  0.097152  0.103328  -0.682864
  2  4.22  1.35  4.86  0.11  0.135938  0.014286   0.236056
  3  3.34  0.26  4.19  0.38  0.064896  0.070898  -0.568802
  4  3.77  0.23  4.42  0.68  0.105082  0.001799   0.083276
  5  3.52  1.10  3.17  0.18  0.149217  0.004185   0.127099
  6  3.54  0.76  2.76  0.00  0.167164  0.039271  -0.394624
  7  3.74  1.59  3.81  0.08  0.172851  0.035228   0.372962
  8  3.78  0.39  3.23  0.11  0.132582  0.037014  -0.384540
  9  2.92  0.39  5.44  1.53  0.100859  0.063853   0.518806
 10  3.10  0.64  6.16  0.77  0.134379  0.156992  -0.845616
```

11	2.86	0.82	5.48	1.17	0.076758	0.004387	0.130432
12	2.78	0.64	4.62	1.01	0.059251	0.000022	-0.009147
13	2.22	0.85	4.49	0.89	0.155650	0.112024	-0.689419
14	2.67	0.90	5.59	1.40	0.103917	0.039358	0.399648
15	3.12	0.92	5.86	1.05	0.084138	0.003260	0.112283
16	3.03	0.97	6.60	1.15	0.166074	0.000016	-0.007909
17	2.45	0.18	4.51	1.49	0.126445	0.012589	0.221531
18	4.12	0.62	5.31	0.51	0.140753	0.002771	0.103364
19	4.61	0.51	5.16	0.18	0.265126	0.003324	-0.113155
20	3.94	0.45	4.45	0.34	0.102589	0.008336	-0.180066
21	4.12	1.79	6.17	0.36	0.191406	0.089547	0.604725
22	2.93	0.25	3.38	0.89	0.095939	0.001015	0.062516
23	2.66	0.31	3.51	0.91	0.107166	0.005277	-0.142951
24	3.17	0.20	3.08	0.92	0.114972	0.045597	0.430562
25	2.79	0.24	3.98	1.35	0.081006	0.045952	0.438348
26	2.61	0.20	3.64	1.33	0.111527	0.042193	0.413680
27	3.74	2.27	6.50	0.23	0.260386	0.005582	0.146706
28	3.13	1.48	4.28	0.26	0.110855	0.013270	-0.227793
29	3.49	0.25	4.71	0.73	0.081181	0.005342	-0.144011
30	2.94	2.22	4.58	0.23	0.304739	0.000000	0.001169

In this example, $\bar{h} = \frac{3(k+1)}{n} = \frac{3\times 4}{30} = 0.40$. The column labeled "HI" in the MINITAB output indicates that there is no possible influential in the data since all the $h_i < 0.40$. Also, $\mathrm{F}(4,26,0.95) = 2.743$ and none of Cook's D values exceeds this value. Hence we can similarly assume that there are no unusual observations in the data. However, observations 3 and 10 have standardized residuals greater then $|2|$ and hence these two observations are possible outliers.

7.10.2 Multicollinearity

When several explanatory variables are involved in a multiple regression model, there is a need to ensure that the explanatory variables are not themselves highly correlated. When this occurs, it leads to what is called multicollinearity and the attendant inflation of standard errors of the estimated regression coefficients which may lead to wrongful hypotheses decision. In fact one should test for multicollinearity in a multiple regression involving more than two independent variables, more so, if some of the variables are derivatives (e.g, x_1^2, x_2^2, or x_1x_2) of the original variables. The multicollinearity is checked with the use of *variance inflation factors* defined as:

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2} \quad \text{for } j = 1, 2, \cdots, k.$$

A VIF value greater than ten is said to constitute a presence of multicollinearity in our data. In the above example, the VIF values are all less than ten, and we can therefore conclude that multicollinearity is not present in our explanatory variables.

7.11 Rank Correlation

Our earlier analysis on correlation and regression assumes that the data came from normally distributed populations. However, if this were not the case, then we need to employ *rank correlation analysis*. Two methods commonly in use are Spearman's (1904) and Kendall's (1938) rank correlation analyses.

Consider the following data (Blaisdell 2010) which relate to data collected by a certain professor who wishes to determine if there is an association between the final examination scores in her course and the times required to complete the exam. Exam scores and times (in minutes) for 15 students are presented in the table below.

Student	Exam score (Y)	Time (X)
1	86	108
2	94	100
3	73	115
4	78	113
5	54	118
6	93	99
7	69	110
8	79	109
9	84	111
10	82	117
11	41	120
12	67	116
13	98	89
14	74	112
15	71	110

```
MTB > Rank 'Y' c3.
MTB > Rank 'X' c4.
MTB > let c5=(c3-c4)**2
MTB > sum c5
Sum of d2 = 978

MTB > print c1-c5
Data Display
Row   Y    X   y1  x1   d2
  1  86  108  12   5   49
  2  94  100  14   3  121
  3  73  115   6  11   25
  4  78  113   8  10    4
  5  54  118   2  14  144
  6  93   99  13   2  121
  7  69  110   4   7    9
  8  79  109   9   6    9
  9  84  111  11   8    9
 10  82  117  10  13    9
 11  41  120   1  15  196
 12  67  116   3  12   81
 13  98   89  15   1  196
 14  74  112   7   9    4
 15  71  107   5   4    1
```

Columns labeled $y1$ and $x1$ contain respectively, the rankings of Y and X. The column labeled $d2$ has $d2 = (y1 - x1)^2$, with $\sum_{i=1}^{15} d_i = 978$. The appropriate formula for computing Spearman's rank correlation is given by:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(978)}{15(225-1)} = 0.7464.$$

```
MTB > XTABS 'Y'  'X';
SUBC>   Layout 1 1;
SUBC>   Counts;

SUBC>   DMissing 'Y'  'X';
SUBC>   Correlation.


Pearson's r     -0.746429
Spearman's rho  -0.746429
```

The MINITAB code for doing this can be simply:

```
MTB > xtabs c1 c2;
SUBC> correlation.
```

Alternatively, we can perform Pearson's correlation on the ranked variables, however, specifying that p values should not be computed. Thus,

```
MTB > Correlation 'y1' 'x1';
SUBC>   NoPValues.

Correlations: y1, x1

Pearson correlation of y1 and x1 = -0.746
```

7.12 Concordance Correlation

Most often, the researcher may want to see if measurements on one instrument can be reproduced on another instrument (which may be of the same or different make). In such situations, pairs of observations are collected on the same samples from two different methods or two different instruments. Consider the example below relating to avian plasma concentrations (in nanograms per milliliter) that were determined by two different assay methods immediately after the blood was collected (Zar 2010).

Blood sample	Method A	Method B
1	6.1	5.0
2	8.6	7.7
3	11.0	11.4
4	13.2	13.9
5	16.9	18.5
6	20.5	21.7
7	22.7	25.3
8	25.8	27.9
9	26.7	29.5
10	28.8	32.6
11	31.4	35.9
12	34.3	38.4

The concordance correlation coefficient, r_c, can be computed as:

$$r_c = \frac{2S_{xy}}{S_{xx} + S_{yy} + (n-1)(\bar{x} - \bar{y})^2}. \tag{7.46}$$

The summary statistics for the data are:

$$\sum x = 246.00 \qquad \sum y = 267.80 \qquad \sum xy = 6628$$

$$\sum x^2 = 5997.98 \qquad \sum y^2 = 7335.48 \qquad S_{xy} = 1138.10$$

$$S_{xx} = 954.98 \qquad S_{yy} = 1359.08$$

$$\bar{x} = 20.50 \qquad \bar{y} = 22.32 \qquad n = 12.$$

Hence, from (7.46), r_c is computed as:

$$\frac{2(1138.10)}{954.98 + 1359.08 + (12-1)(20.5 - 22.32)^2} = \frac{2276.20}{2350.4964} = 0.9684$$

7.13 Multiple and Partial Correlations

Suppose the the zero-order correlation coefficients between an independent variable y and two explanatory variables x_1 and x_2 are displayed as follows for the data in Table 7.6.

```
MTB > Correlation 'x1' 'x2' 'y'.

Correlations: x1, x2, y

       x1     x2
x2  0.399
    0.113

y   0.720  0.212
    0.001  0.414
```

Here, $r_{12} = 0.399$ ($r_{12}^2 = 0.1592$), $r_{y1} = 0.720$ ($r_{y1}^2 = 0.5184$), and $r_{y2} = 0.212$ ($r_{y2}^2 = 0.0449$).

The multiple correlation coefficient for the two independent variables (IVs) as a function of the zero-order correlation coefficients is given by:

$$R_{y.12} = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}} \tag{7.47}$$

Thus,

$$R_{y.12} = \sqrt{\frac{0.5184 + 0.0449 - 2(0.720)(0.212)(0.399)}{1 - 0.1592}} = \sqrt{\frac{0.4415}{0.8408}} = 0.7246.$$

Alternatively, we can compute $R_{y.12}$ from the expression:

$$R_{y.12} = \hat{\beta}_{y1.2}. \tag{7.48}$$

7.13.1 Partial Correlations

The partial correlations written as $r_{y1.2}$ or $r_{y2.1}$ can be computed, respectively, from the following expressions:

$$r_{y1.2} = \frac{r_{y1} - r_{y2}\, r_{12}}{\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}}$$

$$r_{y2.1} = \frac{r_{y2} - r_{y1}\, r_{12}}{\sqrt{(1 - r_{y1}^2)(1 - r_{12}^2)}}. \tag{7.49}$$

Again, for our example,

$$r_{y1.2} = \frac{0.720 - 0.212(0.399)}{\sqrt{(1 - 0.0449)(1 - 0.1592)}} = \frac{0.6354}{0.8961} = 0.7091$$

and $r_{y1.2}^2 = 0.7091^2 = 0.5028$. Similarly,

$$r_{y2.1} = \frac{0.212 - 0.720(0.399)}{\sqrt{(1 - 0.5184)(1 - 0.1592)}} = \frac{-0.0753}{0.6363} = -0.1183$$

and $r_{y2.1}^2 = -0.1183^2 = 0.0140$.

In general, partial correlation coefficients are computed from simple correlation coefficients as:

$$r_{ij.k} = \frac{r_{ij} - r_{ik}\, r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}. \tag{7.50}$$

While MINITAB does not directly compute the partial correlation coefficients, however, SAS software computes the squares of the partials and we

present a typical output from SAS that estimates these values. The signs on the partials are derived from the signs of the regression parameter estimates (either positive or negative).

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Number of Observations Read	17
Number of Observations Used	17

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2325.17948	1162.58974	7.75	0.0054
Error	14	2101.29111	150.09222		
Corrected Total	16	4426.47059			

Root MSE	12.25121	R-Square	0.5253
Dependent Mean	76.17647	Adj R-Sq	0.4575
Coeff Var	16.08267		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	66.46540	9.84961	6.75	<.0001	98649	6834.56761
x1	1	1.29019	0.34276	3.76	0.0021	2295.23440	2126.54090
x2	1	-0.11104	0.24859	-0.45	0.6619	29.94508	29.94508

Parameter Estimates

Variable	DF	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	.	.
x1	1	0.51852	0.50299
x2	1	0.01405	0.01405

Model: MODEL1

Test 1 Results for Dependent Variable Y

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	29.94508	0.20	0.6619
Denominator	14	150.09222		

The REG Procedure
Model: MODEL1

Test 2 Results for Dependent Variable Y

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	1191.81502	7.94	0.0137
Denominator	14	150.09222		

Predictor	Sequential SS	Partial SS
x_1	2295.2344	2126.5409
x_2	29.9451	29.9451

The square of the partial correlation of coefficient of y and x_2 adjusted for x_1 is computed as:

$$r^2_{y2.1} = \frac{R(\beta_2|\beta_0, \beta_1)}{\text{Total SS} - R(\beta_1|\beta_0)} = \frac{29.9451}{4426.4705 - 2295.2344} = 0.01405$$

7.14 Comparisons of Regressions

Suppose that an experimenter wishes to determine how the response y is influenced by the dosage x of each of two comparable treatments. Treatment 1 is administered to n_1 subjects in different dosages, and their response measurements are recorded. Similarly, Treatment 2 is administered to an independent group of n_2 subjects and their responses are recorded. If we use the suffix 1 to denote data from the first set of data (Treatment 1) and the suffix 2 for the second set of data (Treatment 2). If we assume that a linear relationship is appropriate for each treatment, we have

$$\begin{aligned} &\text{Treatment 1: } y_{1i} = \alpha_1 + \beta_1 x_{1i} + e_{1i} \qquad i = 1, 2, \cdots, n_1 \\ &\text{Treatment 2: } y_{2j} = \alpha_2 + \beta_2 x_{2j} + e_{2j} \qquad j = 1, 2, \cdots, n_2 \end{aligned} \quad .$$

It is often of practical interest to test the null hypothesis that the two regression lines have equal slope, that is

$$H_0 : \quad \beta_1 = \beta_2.$$

Graphically, this is equivalent to the hypothesis that these two lines are parallel.

From previous knowledge

$$b_1 = \frac{S_{x_1 y_1}}{S_{x_1}}, \quad a_1 = \bar{y}_1 - b_1 \bar{x}_1$$

$$\text{Error SS} = \text{SSE}(1) = S_{y_1} - \frac{(S_{x_1 y_1})^2}{S_{x_1}} \text{ on } (n_1 - 2) \text{ d.f.}$$

Similarly,

$$b_2 = \frac{S_{x_2 y_2}}{S_{x_2}}, \quad a_2 = \bar{y}_2 - b_2 \bar{x}_2$$

$$\text{Error SS} = \text{SSE}(2) = S_{y_2} - \frac{(S_{x_2 y_2})^2}{S_{x_2}} \text{ on } (n_2 - 2) \text{ d.f.}$$

If a single straight line were fitted,

$$b = \frac{S_{xy}}{S_x}, \quad \text{and SSE} = S_y - \frac{(S_{xy})^2}{S_x} \text{ on } (n_1 + n_2 - 2) \text{ d.f.}$$

where

$$\sum x = \sum x_1 + \sum x_2, \quad \sum x^2 = \sum x_1^2 + \sum x_2^2$$
$$\sum y = \sum y_1 + \sum y_2, \quad \sum y^2 = \sum y_1^2 + \sum y_2^2$$
$$\sum xy = \sum x_1 y_1 + \sum x_2 y_2, \quad \text{and}$$
$$S_x = \sum x^2 - \frac{(\sum x)^2}{n_1 + n_2}, \quad S_y = \sum y^2 - \frac{(\sum y)^2}{n_1 + n_2}$$
$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n_1 + n_2}.$$

An estimate of the pooled variance S^2 is given by

$$\frac{\text{SSE}(1) + \text{SSE}(2)}{n_1 + n_2 - 4}.$$

Hence the difference $= \text{SSE} - [\text{SSE}(1) + \text{SSE}(2)]$ will be based on $(n_1 + n_2 - 2) - (n_1 + n_2 - 4) = 2$ d.f.

and

$$F* = \frac{\text{Difference MS}}{S^2}$$
$$= \frac{(n_1 + n_2 - 4)[\text{SSE} - \text{SSE}(1) - \text{SSE}(2)]}{2[\text{SSE}(1) + \text{SSE}(2)]}$$

and the computed F will be distributed F with 2 and $(n_1 + n_2 - 4)$ d.f.

7.14.1 Example

Consider the following data relating two treatments A and B with regards to the reduction in blood sugar level, where the explanatory variable X refers to the dosage level for each of the treatments (Table 7.8).

Table 7.8 Blood sugar reduction levels for various dosage levels

Treatment A		Treatment B	
x_i	y_i	x_i	y_i
0.20	30	0.20	23
0.25	26	0.25	24
0.25	46	0.30	42
0.30	35	0.40	49
0.40	54	0.40	55
0.50	56	0.50	70
0.50	65		

For treatment A: $n_1 = 7$ and,

$$\sum x_1 = 2.4, \qquad \sum x_1^2 = 0.915, \qquad S_{x_1} = 0.0921$$
$$\sum y_1 = 312.0, \qquad \sum y_1^2 = 15194.0 \qquad S_{y_1} = 1287.7143$$
$$\sum x_1 y_1 = 116.60, \qquad S_{x_1 y_1} = 9.6286.$$

Hence,

$$\hat{\beta}_1 = \frac{S_{x_1 y_1}}{S_{x_1}} = 104.5451$$

and therefore,

$$\text{Residual SS} = 1287.7143 - \frac{9.6286^2}{0.0921}, \quad \text{that is,}$$
$$\text{SSE}(1) = 281.0917.$$

For treatment B: $n_2 = 6$ and,

$$\sum x_2 = 2.05, \qquad \sum x_2^2 = 0.7625, \qquad S_{x_2} = 0.0621$$
$$\sum y_2 = 263.0, \qquad \sum y_2^2 = 13195 \qquad S_{y_2} = 1666.8333$$
$$\sum x_2 y_2 = 99.8, \qquad S_{x_2 y_2} = 9.9417$$

Hence,

$$\hat{\beta}_2 = \frac{S_{x_2 y_2}}{S_{x_2}} = 160.1342$$

and again, therefore,

$$\text{Residual SS} = 1666.8333 - \frac{9.9417^2}{0.0621}, \quad \text{that is,}$$
$$\text{SSE}(2) = 74.8322.$$

If a single straight line were fitted, the estimate of the pooled variance S^2 is

$$S^2 = \frac{74.8322 + 281.0917}{7 + 6 - 4} = 39.5471 \quad \text{on 9 d.f.}$$

with $n_1 + n_2 = 7 + 6 = 13$, and,

$$\sum x = \sum x_1 + \sum x_2 = 2.4 + 2.05 = 4.45$$
$$\sum x^2 = \sum x_1^2 + \sum x_2^2 = 0.915 + 0.7625 = 1.6775$$

$$\sum y = \sum y_1 + \sum y_2 = 312 + 263 = 575$$
$$\sum y^2 = \sum y_1^2 + \sum y_2^2 = 15,194 + 13,195 = 28,389$$
$$\sum xy = \sum x_1 y_1 + \sum x_2 y_2 = 116.6 + 99.8 = 216.4.$$

Therefore,

$$S_x = 1.6775 - \frac{4.45^2}{13} = 0.1542$$
$$S_y = 28,389 - \frac{575^2}{13} = 2956.3077$$
$$S_{xy} = 216.4 - \frac{4.45 \times 575}{13} = 19.5731$$
$$\text{SSE} = 2956.3077 - \frac{(19.5731)^2}{0.1542} = 471.8314.$$

$$\begin{aligned}\text{Difference} &= \text{SSE} - \text{SSE}(1) - \text{SSE}(2)\\ &= 471.8314 - 281.0917 - 74.8322\\ &= 115.9075 \quad \text{on 2 d.f.}\end{aligned}$$

Therefore,

$$F = \frac{115.9075/2}{39.5471} = 1.47 \quad \text{with 2 and 9 d.f.}$$

Since F is not significant, i.e., we accept the null hypothesis that the regression lines are coincident. In other words, the treatment effect can be ignored and a single straight line of the form $Y = \beta_0 + \beta_1 x$ would be appropriate for the data in Table 7.8. The results above can succinctly be displayed, as in Table 7.9.

Table 7.9 Analysis of variance table for testing the hypothesis of parallelism

Source	df	SS	MS	F
(a) SSE(1)	5	281.0917		
(b) SSE(2)	4	74.8322		
(c) (a) + (b)	9	353.9239	$39.5471 = S^2$	
(d) SSE	11	471.8314		
(e) Difference (d) − (c)	2	115.9075	57.9538	1.47

F value in Table 7.9 is computed as Difference MS$/S^2$ = (e)/(c) = 1.47, which is clearly not significant at $\alpha = 0.05$.

7.14.2 Alternative Approach

An alternative test is to proceed as follows:

$$\text{Compute } S^2 = \frac{\text{SSE}(1) + \text{SSE}(2)}{n_1 + n_2 - 4}$$

$$\text{Var}(\hat{\beta}_1 - \hat{\beta}_2) = S^2 \left(\frac{1}{S_{x_1}} + \frac{1}{S_{x_2}} \right)$$

which for our example

$$= 39.5471 \left(\frac{1}{0.0921} + \frac{1}{0.0621} \right) = 1066.2224.$$

Then the test $H_0 : \beta_1 = \beta_2$ is given by

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\text{var}\,(\hat{\beta}_1 - \hat{\beta}_2)}} = \frac{104.5451 - 160.1342}{\sqrt{1066.2224}}$$

$$= \frac{-55.5891}{32.6531} = -1.7024$$

where t is tabulated as Student's t distribution with $(n_1 + n_2 - 4) = 9$ d.f.

Clearly at the 5 % point, the calculated value of $t_{0.975} = 2.2622$ is not significant, i.e., we accept H_0 which is clearly the earlier conclusion arrived at in the previous section.

To implement the above in MINITAB, we do the following:

(a) Read the data in as $x1$, y and TRT in columns C1, C2, and C3 respectively. Notice that C3 is alphanumeric.

(b) To be able to use column C3, we create a dummy variable x_2 such that

$$x_2 = \begin{cases} 1 & \text{if TRT A} \\ 0 & \text{if TRT B} \end{cases}$$

This is accomplished with the statement beginning "Indicator." MINITAB actually creates two dummy variables into columns C4 and C5 corresponding to the levels of variable TRT. One of these is, however, redundant and we have labeled the redundant column NA (not applicable).

(c) Create a variable x_3, which is the interaction of x_1 and x_2. That is, $x_3 = x_1 \times x_2$.

```
MTB > Indicator 'TRT' C4 C5.
MTB > LET C6=C1*C4
MTB > PRINT C1-C6

Data Display

 Row     x1      y   TRT    x2    NA     x3

   1   0.20     30     A     1     0   0.20
   2   0.25     26     A     1     0   0.25
   3   0.25     46     A     1     0   0.25
   4   0.30     35     A     1     0   0.30
   5   0.40     54     A     1     0   0.40
   6   0.50     56     A     1     0   0.50
   7   0.50     65     A     1     0   0.50
   8   0.20     23     B     0     1   0.00
   9   0.25     24     B     0     1   0.00
  10   0.30     42     B     0     1   0.00
  11   0.40     49     B     0     1   0.00
  12   0.40     55     B     0     1   0.00
  13   0.50     70     B     0     1   0.00
```

We wish to fit the multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{7.51}$$

where x_2 is as defined above and x_3 is the product of x_1 and x_2.

```
MTB > Regress 'y' 3 'x1' 'x2' 'x3';
SUBC>   Constant;
SUBC>   Brief 2.

Regression Analysis: y versus x1, x2, x3


The regression equation is
y = - 10.9 + 160 x1 + 19.6 x2 - 55.6 x3

Predictor        Coef     SE Coef          T        P
Constant      -10.879       9.003      -1.21    0.258
x1             160.13       25.26       6.34    0.000
x2              19.62       11.71       1.68    0.128
x3             -55.64       32.67      -1.70    0.123

S = 6.293       R-Sq = 87.9%     R-Sq(adj) = 83.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     2599.91      866.64     21.88    0.000
Residual Error     9      356.40       39.60
Total             12     2956.31

Source       DF      Seq SS
x1            1     2483.97
x2            1        1.11
x3            1      114.82
```

From (7.51), we see that when $x_2 = 0$, we have the regression equation in (7.52a) for TRT B, and when we substitute $x_2 = 1$ again in (7.51), we have the regression equation in (7.52b) for TRT A. That is,

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \tag{7.52a}$$

$$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1 + \varepsilon. \tag{7.52b}$$

From (7.52a) and (7.52b), the intercepts and slopes of the two regression models for treatments A and B are presented in Table 7.10.

Table 7.10 Intercepts and slopes for the two regression lines

Regression equation for	Intercept	Slope
TRT A	$(\beta_0 + \beta_2)$	$(\beta_1 + \beta_3)$
TRT B	β_0	β_1

For the two lines to be coincident, therefore, both the parameters β_2 and β_3 in the equation for TRT A must be zero. That is, we would need to test the hypotheses:

$$H_0 : \beta_2 = \beta_3 = 0 \tag{7.53}$$

$$H_a : \text{at least one of these} \neq 0 \tag{7.54}$$

To conduct these hypotheses, we note from the MINITAB output that the sequential SS are:

```
Source       DF      Seq SS
x1            1     2483.97
x2            1        1.11
x3            1      114.82
```

That is, $\text{SS}(x_1 = 2483.97)$, $\text{SS}(x_2|x_1 = 1.11)$, and $\text{SS}(x_3|(x_1, x_2) = 114.82)$. Hence, the sum of squares due to x_2 and x_3 given that x_1 is already in the model is $\text{SS}(x_2, x_3|x_1) = 1.11 + 114.82 = 115.93$ and is based on 2 d.f. Hence, the test statistic for testing the hypotheses in (7.53) is computed as:

$$F = \frac{115.93/2}{\text{EMS}} = \frac{57.965}{39.60} = 1.46 \quad \text{with 2 and 9 d.f.}$$

where EMS is the error mean square in the multiple regression output from MINITAB, which is based on 9 d.f. We see that the above result is comparable with our earlier result from manual calculations. Once again, we would fail to reject the null hypothesis in (7.53), indicating that both β_2 and β_3 could be zero, reducing the equation for both treatments to that in (7.52a). Thus, the appropriate regression model for the data in Table 7.8 is the model

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

which when implemented in MINITAB gives the following output:

```
MTB > Regress 'y' 1 'x1';
SUBC>   Constant;
SUBC>   Brief 3.

Regression Analysis: y versus x1


The regression equation is
y = 0.79 + 127 x1

Predictor        Coef     SE Coef          T        P
Constant        0.789       5.994       0.13    0.898
x1             126.91       16.69       7.61    0.000

S = 6.553       R-Sq = 84.0%     R-Sq(adj) = 82.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      2484.0      2484.0     57.85    0.000
Residual Error    11       472.3        42.9
Total             12      2956.3
```

and the estimated regression equation in Fig. 7.9 has the form:

$$\hat{y}_i = 0.789 + 126.91 x_{1i}.$$

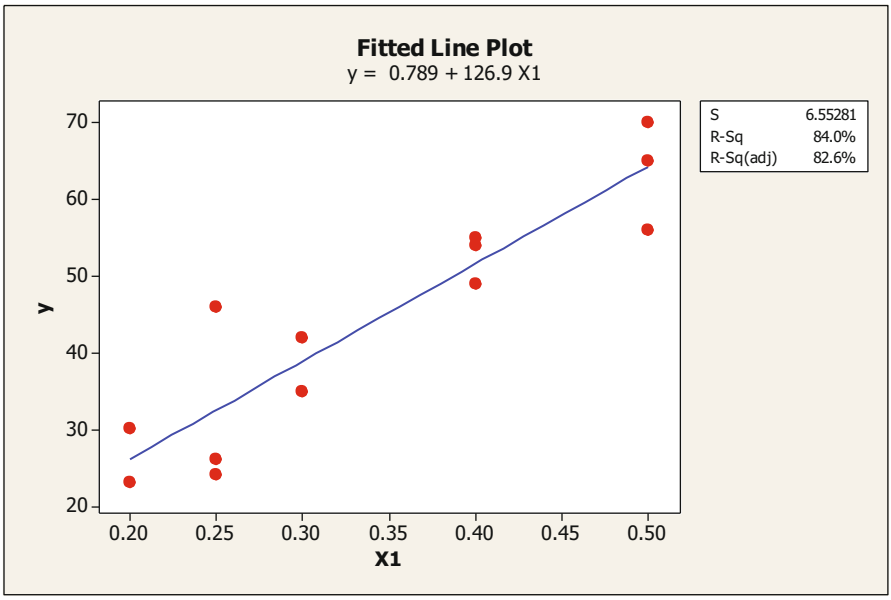


Fig. 7.9 Plot of the estimated regression equation

7.15 Fitting Parallel Lines

In the last example, our conclusion was that a single relationship holds for both sets of data, and of course there is no need to consider any other possibility. If however, we had found sufficient evidence that the relationships for the two sets are not identical, then one may be interested in testing whether the slopes of the two regressions are the same and that the two sets of data differ only in their intercepts. Thus, we would need to fit a model of the form:

$$\begin{aligned} \text{Set 1}: &\quad y_{1i} = \beta_{01} + \beta x_{1i}\, i = 1, 2, \cdots, n_1 \\ \text{Set 2}: &\quad y_{2j} = \beta_{02} + \beta x_{2j}\, j = 1, 2, \cdots, n_2. \end{aligned}$$

Here we have two different values of β_{0i}, $i = 1, 2$ but a common value of β, that is, the lines are parallel.
The estimate of the common slope is

$$b = \frac{S_{x_1 y_1} + S_{x_2 y_2}}{S_{x_1} + S_{x_2}} = \frac{S_{xy}}{S_x}$$

while the estimates of the intercepts are

$$\begin{aligned} b_{01} &= \bar{y}_1 - b\bar{x}_1 \\ b_{02} &= \bar{y}_2 - b\bar{x}_2 \quad \text{and} \end{aligned}$$

$$\text{Residual SS: RSS} = S_y - \frac{S_{xy}^2}{S_x} \quad \text{on } (n_1 + n_2 - 3) \text{ d.f.}$$

where

$$\begin{aligned} S_y &= S_{y_1} + S_{y_2} \\ S_{xy} &= S_{x_1 y_1} + S_{x_2 y_2} \quad \text{and} \\ S_x &= S_{x_1} + S_{x_2} \end{aligned}$$

Since there are three parameters estimated namely β_{01}, β_{02}, and β from the models, therefore, the difference degrees of freedom equals,

$$\begin{aligned} \text{Diff. d.f.} &= \text{RSS d.f.} - \text{ESS(1) d.f.} - \text{ESS(2) d.f.} \\ &= (n_1 + n_2 - 3) - (n_1 - 2) - (n_2 - 2) \\ &= (n_1 + n_2 - 3) - (n_1 + n_2 - 4) \\ &= 1 \quad \text{d.f.} \end{aligned}$$

and therefore,

$$F = \frac{\text{Diff. MS}}{S^2}$$

where S^2 is as given in the previous section, and F is distributed as F distribution with 1 and $(n_1 + n_2 - 4)$ degrees of freedom.

Example

For instance, for our example in Table 7.8

$$b = \frac{S_{xy}}{S_x} = \frac{9.6286 + 9.9417}{0.0921 + 0.0621} = \frac{19.5703}{0.1542} = 126.9150$$

and,

$$b_{01} = \bar{y}_1 - b\bar{x}_1 = 44.5714 - 126.9150(0.3429) = 1.0522$$
$$b_{02} = \bar{y}_2 - b\bar{x}_2 = 43.8333 - 126.9150(0.3417) = 0.4664.$$

Hence,

$$S_y = 1287.7143 + 1666.8333 = 2954.5476.$$

Therefore, the residual sum of squares RSS is computed as:

$$\text{RSS} = S_y - \frac{S_{xy}^2}{S_x} = 2954.5476 - \frac{19.5703^2}{0.1542} = 470.7821.$$

The difference is again computed as:

$$\text{Difference} = 470.7821 - \text{ESS}(1) - \text{ESS}(2) = 114.8582 \quad \text{on 1 d.f.}$$

Since ESS(1) = 281.0917 and ESS(2) = 74.8322 from our previous calculations. Hence,

$$F = \frac{114.8582}{39.5471} \quad = 2.90$$

F(1,9) = 7.21 at $\alpha = 0.05$. Thus, the calculated F value is not significant. That is, we can conclude that the two regression lines are parallel.

Alternatively, we could use the following approach, but first, let us display in the table below the summary statistics from the two data samples.

Sample	n_i	S_{yy}	S_{xx}	S_{xy}
1	7	1287.7143	0.0921	9.6286
2	6	1666.8333	0.0621	9.9417
Total	13	2954.5476	0.1542	19.5703

However, we need the following computations from the above table:

(a) Common slope which equals:

$$\text{b}_0 = \sum_{i=1}^{2} \left(\frac{S_{xy_i}}{S_{xx_i}} \right) = \frac{19.5703}{0.1542} = 126.9150.$$

(b) The Residual SS about separate regressions:

$$\begin{aligned}\text{RSS} &= \sum_{i=1}^{2} S_{yy_i} - \sum_{i=1}^{2} \left(\frac{S_{xy_i}^2}{S_{xx_i}} \right) \\ &= 2954.5476 - \left[\frac{(9.6286)^2}{0.0921} + \frac{(9.9417)^2}{0.0621} \right] \\ &= 2954.5476 - (1006.6226 + 1591.5845) \\ &= 2954.5476 - 2598.2071 \\ &= 356.3405.\end{aligned}$$

This will be based on $n_1 + n_2 - 2k$ d.f., where k is the number of regression lines.

(c) SS due to fitting common slope:

$$\text{FIT SS} = \frac{(\sum_i S_{x_i y_i})^2}{S_{xx_i}} = \frac{(19.5703)^2}{0.1542} = 2483.7655. \quad i = 1, 2.$$

This is based on 1 d.f.

(d) SS due difference in slope, that is between slopes SS (BSS):

$$\begin{aligned}\text{BSS} &= \sum_i S_{yy_i} - \frac{(\sum_i S_{xy_i})^2}{\sum_i S_{xx_i}} = 2598.2071 - 2483.7655 \\ &= 114.4416\end{aligned}$$

and is based on $k - 1$ degrees of freedom.

Table 7.11 Analysis of variance table for testing the hypothesis of parallelism

Source	d.f.	SS	MS	F
Common slope	1	2483.7655	2483.7655	
Between slopes	1	114.4416	114.4416	2.89
Separate residuals	9	356.3405	39.5934	
Total	11	2954.5476		

We can now summarize the above results succinctly in an ANOVA table as shown.

The F value in Table 7.11 is clearly not significant at $\alpha = 0.05$.

To implement parallelism in MINITAB, we notice again from Eqs. (7.52a) and (7.52b) that for the two regression equations to be parallel, the slopes must be equal. That is, β_3 must necessarily be zero. Hence, we would need to test the hypothesis that:

$$H_0 : \beta_3 = 0 \tag{7.55}$$

$$H_a : \beta_3 \neq 0. \tag{7.56}$$

To conduct these hypotheses, we note from the MINITAB output that the sequential SS are again given by:

```
Source      DF      Seq SS
x1           1     2483.97
x2           1        1.11
x3           1      114.82
```

Thus, $\text{SS}(x_3|(x_1, x_2) = 114.82)$ on 1 d.f. and the test statistic for testing the hypotheses in (7.55) is computed as:

$$F = \frac{114.82}{\text{EMS}} = \frac{114.82}{39.60} = 2.90 \quad \text{with 1 and 9 d.f.}$$

```
MTB > Regress 'y' 2 'x1' 'x2';
SUBC>   Constant;
SUBC>   Brief 3.

Regression Analysis: y versus X1, X2


The regression equation is
y = 0.48 + 127 X1 + 0.59 X2

Predictor        Coef     SE Coef          T        P
Constant        0.478       6.597       0.07    0.944
x1             126.89       17.48       7.26    0.000
x2              0.587       3.819       0.15    0.881

S = 6.865       R-Sq = 84.1%     R-Sq(adj) = 80.9%
```

The estimated regression equation when we set $\beta_3 = 0$ is:

$$\hat{y} = 0.478 + 126.89x_1 + 0.587x_2. \tag{7.57}$$

Consequently, if we set $x_2 = 1$ for treatment A and $x_2 = 0$ for treatment B in the above estimated equation, we have the estimated parallel regression lines for TRT A and TRT B respectively in expressions (7.58a) and (7.58b). These lines are plotted in Fig. 7.10. Notice that the lines are very close together, since based on our previous results, these lines are actually coincident.

$$\text{TRT A:} \quad \hat{y} = 1.065 + 126.89x_1 \tag{7.58a}$$

$$\text{TRT B:} \quad \hat{y} = 0.478 + 126.89x_1 \tag{7.58b}$$

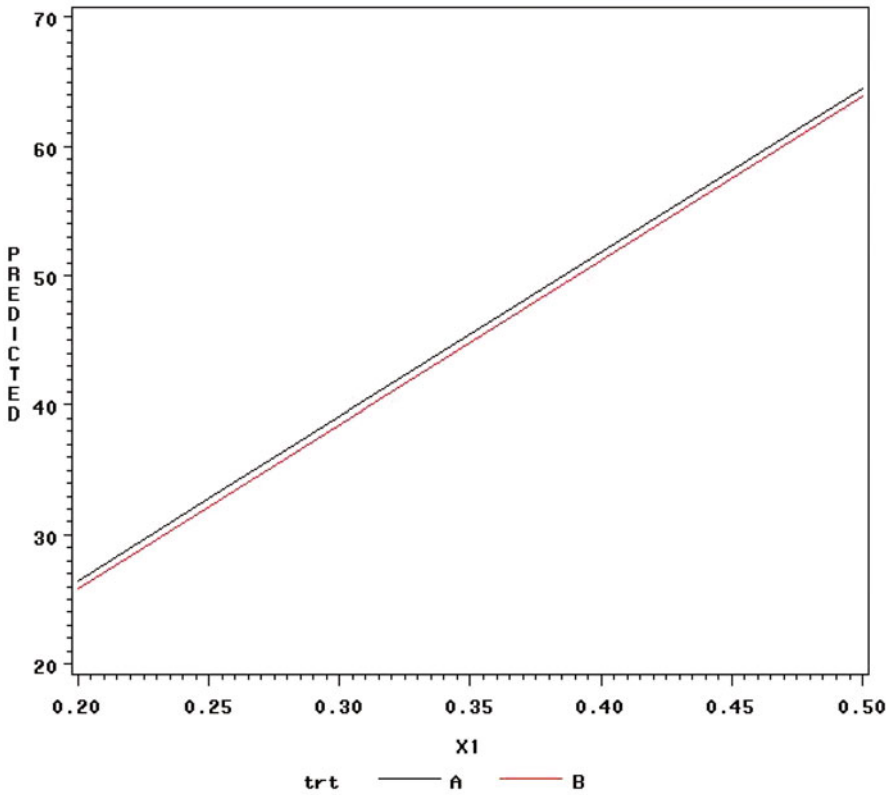


Fig. 7.10 Plot of the estimated parallel regression equations

7.16 Nonlinear Regression

In the earlier sections, we were fitting models of the form

$$Y = \beta_0 + \beta_1 x + \varepsilon$$
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \varepsilon.$$

These models are said to be linear with respect to their parameters $\beta_0, \beta_1, \beta_2$, etc. even though the second model indicates a quadratic relationship between y and x and both are often referred to as linear models. However, it has been established that for most biological situations, the linear model may be inadequate. Most nonlinear models applicable to biological or agricultural situations can be grouped into three kinds.

(a) Exponential growth curves

$$y = \beta_0 \beta_1^x$$

that is, an exponential growth curve, the parameters β_0 and β_1 are no longer linear. Another form of this curve is given by $Y = \beta_0 \beta_1^{-x}$, that is, an exponential decay law. In either case a simple transformation, e.g., logarithmic will make the model linear in their parameters. For example,

$$\text{for growth curve } \log\ y = \log \beta_0 + x \log \beta_1$$
$$\text{decay curve } \log\ y = \log \beta_0 - x \log \beta_1$$

which are now equivalent to the last models. Both models are sometimes written in the form $y = \beta_0 e^{\beta_1 x}$ and $y = \beta_0 e^{-\beta_1 x}$ respectively.
The graphs in Fig. 7.11 for instance give the exponential models for the cases when $\beta > 0$ and when $\beta < 0$. The former is the model $y = 6(1.0408)^x$, while the latter is $y = 400e^{-0.05x}$.

(b) The models in (a) are problematic because the rapid and unlimited increase as x increases and this has sometimes made it inappropriate for use with real biological data which often requires asymptotic convergence of y as x increases. To overcome this problem, the exponential model that includes limitation on growth has been suggested and is of the form:

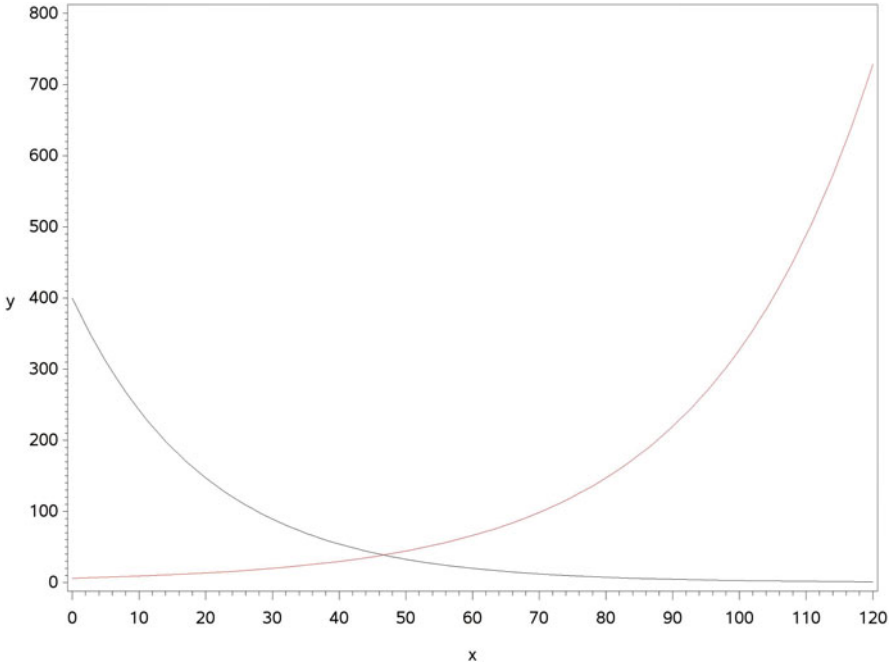


Fig. 7.11 Exponential growth and decay curves

$$y_i = a(1 - e^{-bx_i}) + \varepsilon_i. \tag{7.59}$$

The model in (7.59) is often referred to as the *negative exponential* model and we present in Fig. 7.12 the plots for $a = 20$ and two b values of -0.06 and -0.03, with the $b = -0.06$ curve being steeper.

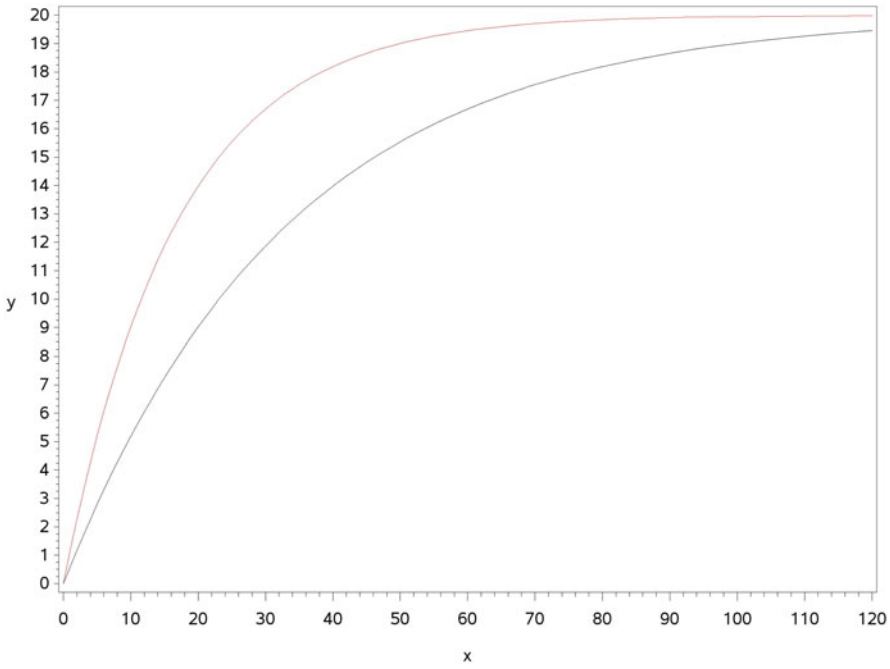


Fig. 7.12 Negative exponential growth curves

The model in (7.59) is a special case of growth models satisfying Mitcherlich Law. Specifically, this model has been referred to as the MacArthur–Wilson growth equation.

Other exponential curves include the *two-term exponential* curves which rise steeply from zero and then fall slowly to zero asymptotically. These curves are of the form

$$y_i = \frac{\gamma_1}{\gamma_1 - \gamma_2}(e^{-\gamma_2 x_i - \gamma_1 x_i}) + \varepsilon_i. \tag{7.60}$$

By using Eq. (7.60), Fig. 7.13 gives the two-term exponential curves with the higher curve having $\gamma_1 = 0.05$ and $\gamma_2 = 0.04$.

$$y_i = \frac{0.05}{0.05 - 0.04}(e^{-0.04x_i - 0.05x_i}) + \varepsilon_i,$$

$$y_i = \frac{0.05}{0.05 - 0.11}(e^{-0.11x_i - 0.05x_i}) + \varepsilon_i.$$

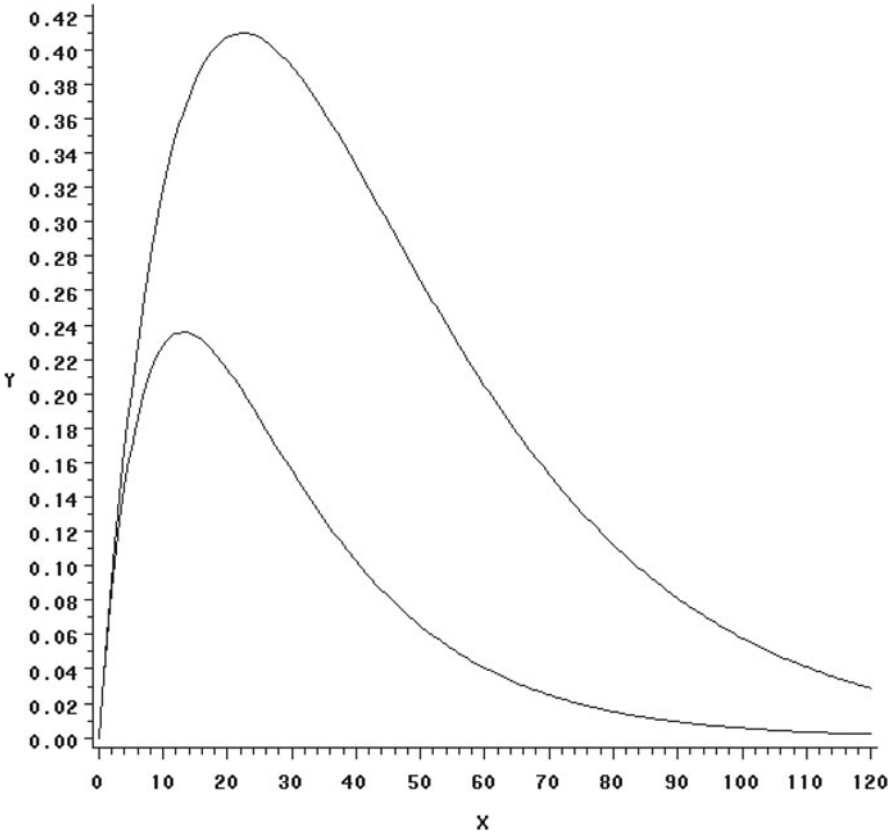


Fig. 7.13 Two-term exponential growth curves

(c) Also commonly used is the *Michaelis–Menten* (1913) model which has received wide usage for enzymatic and chemical kinetic reactions among others. The model is of the form:

$$y = \frac{ax}{b+x}. \tag{7.61}$$

Here as x increases, the function approaches an asymptote, a, and b can be assumed to be the value of x at which the function has reached half its asymptotic value and b is referred to as the Michaelis–Menten constant.

Example: The Puromycin Data

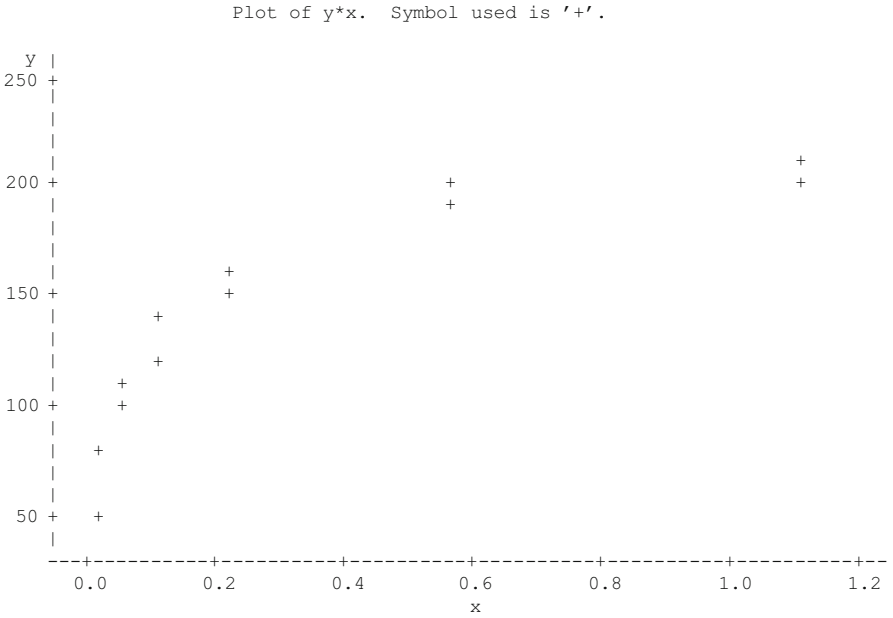
The following example from Bates and Watts (1988, p. 269) gives the relationship between the velocity of an enzymatic reaction (chemical kinetics), y,

and the substrate concentration, x. The data for this experiment is presented in Table 7.12.

Table 7.12 Reaction velocity and substrate concentration data

x	.02	.02	.06	.06	.11	.11	.22	.22	.56	.56	1.10	1.10
y	76	47	97	107	123	139	159	152	191	201	207	200

We present in the following figure scatter plot of reaction velocity (y) against substrate concentration (x).



To estimate the parameters of the the model given in (7.61), we make use of the *NLINEAR* procedure in MINITAB Regression module. We use $a = 100$ and $b = 0.10$ as initial values. We must also supply for MINITAB the expected functional form of the model which in this case is: $\frac{ax}{(b+x)}$. The MINITAB statements and partial output is presented below. We may note here that MINITAB uses the Gauss–Newton method in solving for the parameters of the model.

```
MTB > NLinear;
SUBC>   Response 'y';
SUBC>   Continuous 'x';
SUBC>   Parameter "a" 100;
SUBC>   Parameter "b" .1;
SUBC>   Expectation a*x/(b+x);
SUBC>   NoDefault;
SUBC>   GFCurve;
SUBC>   GHistogram;
SUBC>   TMethod;
SUBC>   TStarting;
SUBC>   TConstraints;
SUBC>   TEquation;
SUBC>   TParameters;
SUBC>   TSummary;
SUBC>   TPredictions.

Nonlinear Regression: y = a * x / (b + x)

Method

Algorithm       Gauss-Newton
Max iterations           200
Tolerance            0.00001


Starting Values for Parameters

Parameter  Value
a            100
b            0.1


Equation

y = 212.684 * x / (0.0641213 + x)


Parameter Estimates

Parameter  Estimate  SE Estimate
a           212.684      6.94716
b             0.064      0.00828

y = a * x / (b + x)

Correlation Matrix for Parameter Estimates
          a
b  0.765084
Lack of Fit

Source         DF       SS       MS     F      P
Error          10  1195.45  119.545
  Lack of Fit   4   497.95  124.487  1.07  0.447
  Pure Error    6   697.50  116.250

Summary

Iterations       11
Final SSE   1195.45
DFE              10
MSE          119.545
S            10.9337
```

Because the observations were replicated at all the values of x, a lack of fit test conducted gives a p value of 0.447 which is not significant, indicating that the model is quite adequate. The predicted observations as well as the residuals are presented below.

```
Data Display

Row     x     y    Resid1    Fits1
  1  0.02   76   25.4340   50.566
  2  0.02   47   -3.5660   50.566
  3  0.06   97   -5.8109  102.811
  4  0.06  107    4.1891  102.811
  5  0.11  123  -11.3616  134.362
  6  0.11  139    4.6384  134.362
  7  0.22  159   -5.6847  164.685
  8  0.22  152  -12.6847  164.685
  9  0.56  191    0.1671  190.833
 10  0.56  201   10.1671  190.833
 11  1.10  207    6.0311  200.969
 12  1.10  200   -0.9689  200.969
```

It seems that observation 1 has a very high residual, otherwise the others are well behaved. Based on the above, therefore, the estimated regression equation is:

$$\hat{y}_i = \frac{212.7x_i}{0.0641 + x_i}$$

with the corresponding estimated response displayed in Fig. 7.14.
We may observe here that MINITAB already had the Michaelis–Menten expectation model which can be readily invoked. It has used θ_1 and θ_2 respectively for a and b parameters in our model.

Transformational Approach

The MM model in (7.61) can be transformed into a linear model using the reciprocal transformation. That is,

$$y = \frac{ax}{b + x}$$

can be transformed into:

$$\frac{1}{y} = \frac{b + x}{ax} = \frac{b}{a}\left(\frac{1}{x}\right) + \frac{1}{a} \tag{7.62}$$

Thus a plot of $1/y$ against $1/x$ should give us a straight line and the regression of $1/y$ against $1/x$ should produce an intercept of $1/a$ and a slope of b/a. When this was implemented in MINITAB, we get the following partial out.

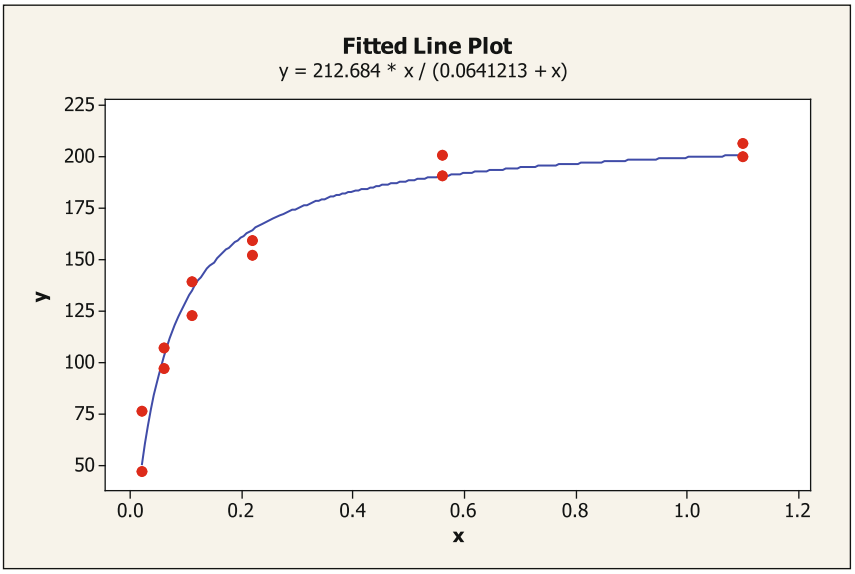


Fig. 7.14 Michealis–Menten predicted curve

```
Regression Equation

invy  =  0.00510718 + 0.000247221 invx


Coefficients

Term           Coef     SE Coef         T       P
Constant  0.0051072  0.0007040   7.25454   0.000
invx      0.0002472  0.0000321   7.70049   0.000


Summary of Model

S = 0.00189226       R-Sq = 85.57%       R-Sq(adj) = 84.13%
PRESS = 0.000116760  R-Sq(pred) = 52.94%


Analysis of Variance

Source          DF      Seq SS      Adj SS      Adj MS         F          P
Regression       1  0.0002123   0.0002123   0.0002123   59.2975   0.000016
  invx           1  0.0002123   0.0002123   0.0002123   59.2975   0.000016
Error           10  0.0000358   0.0000358   0.0000036
  Lack-of-Fit    4  0.0000019   0.0000019   0.0000005    0.0821   0.984946
  Pure Error     6  0.0000339   0.0000339   0.0000057
Total           11  0.0002481
```

Thus $1/a = 0.0051072$ gives $\hat{a} = 195.803$. Similarly $b/a = 0.0002472$ gives $\hat{b} = 0.0484$, leading to the model,

$$\hat{y}_i = \frac{195.803 x_i}{0.0484 + x_i}.$$

which when we realize that we could easily use manual calculations to obtain these values, then the use of transformation is well justified and the results are not too far from those obtained earlier on.

7.16.1 Example 7.16.1

The body weight W in kilograms and the sitting height h in centimeters of five people are given in the following table:

W	50	65	80	100	120
h	84.7	91.2	97.2	104.7	111.0

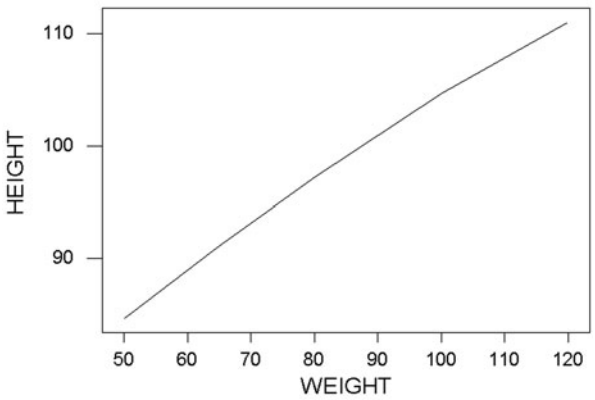


Fig. 7.15 Graph of h against weight W

A graph of the above data indicates that the shape is of the form $h = AW^{\alpha}$ (Fig. 7.15). Transforming here, we have

$$\log h = \log A + \alpha \log W$$
$$\Longrightarrow Y = \alpha + \beta x$$

where $Y = \log h$, $\alpha = \log A$, $\beta = \alpha$, and $X = \log W$, which is now very familiar to us. Hence,

$\log W$	3.91	4.17	4.38	4.61	4.79
$\log h$	4.44	4.51	4.58	4.65	4.71

and

$$\sum y = \sum \log h = 22.89, \qquad \sum y^2 = 104.8367, \qquad S_y = 0.0463$$
$$\sum x = \sum \log W = 21.86, \qquad \sum x^2 = 96.0576, \qquad S_x = 0.4857$$
$$\sum xy = 100.2249, \qquad S_{xy} = 0.1498.$$

Hence,

$$\hat{\beta} = \frac{S_{xy}}{S_x} = \frac{0.1498}{0.4857} = 0.3084$$

and,

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 3.2297$$

while,

$$\hat{A} = e^{\alpha} = e^{3.2297} = 25.272.$$

Hence the estimated nonlinear regression equation is:

$$\hat{h} = 25.272\ W^{0.3084} \tag{7.63}$$

$$R^2 = \frac{\text{Fitted SS}}{\text{Total SS}} = \frac{0.1498^2}{(0.48570 \times 0.0463)} = 0.9978$$

That is $r = 0.999$, a very good fit indeed.

```
MTB > Regress 'LH' 1 'LW';
SUBC>   Constant;
SUBC>   Brief 3.

Regression Analysis: LH versus LW


The regression equation is
LH = 3.22 + 0.311 LW

Predictor        Coef     SE Coef          T        P
Constant      3.21975     0.02689     119.73    0.000
LW           0.310634    0.006135      50.63    0.000

S = 0.004240    R-Sq = 99.9%     R-Sq(adj) = 99.8%

Analysis of Variance

Source           DF          SS          MS         F        P
Regression        1    0.046098    0.046098   2563.66    0.000
Residual Error    3    0.000054    0.000018
Total             4    0.046152
```

Obs	LW	LH	Fit	SE Fit	Residual	St Resid
1	3.91	4.43912	4.43496	0.00340	0.00415	1.64
2	4.17	4.51305	4.51646	0.00225	-0.00341	-0.95
3	4.38	4.57677	4.58096	0.00190	-0.00419	-1.10
4	4.61	4.65110	4.65028	0.00237	0.00082	0.23
5	4.79	4.70953	4.70691	0.00318	0.00262	0.93

Row	w	h	yhat
1	50	84.7	84.347
2	65	91.2	91.509
3	80	97.2	97.606
4	100	104.7	104.612
5	120	111.0	110.707

Alternatively, we could use the nonlinear procedure in MINITAB without having to transform the data, once we know the functional form of the model, which is $h = AW^b$. The following MINITAB results are generated:

```
Fitted Line: logh versus w

MTB > Name C5 "Fits1".
MTB > NLinear;
SUBC>   Response 'h';
SUBC>   Continuous 'w';
SUBC>   Parameter "a" 0.5;
SUBC>   Parameter "b" 0.5;
SUBC>   Expectation a *w**(b);
SUBC>   NoDefault;
SUBC>   GFCurve;
SUBC>   TMethod;
SUBC>   TStarting;
SUBC>   TConstraints;
SUBC>   TEquation;
SUBC>   TParameters;
SUBC>   TSummary;
SUBC>   TPredictions;
SUBC>   Fits 'Fits1'.

Nonlinear Regression: h = a * w ** b

Method

Algorithm       Gauss-Newton
Max iterations           200
Tolerance            0.00001


Starting Values for Parameters

Parameter  Value
a            0.5
b            0.5


Equation

h = 24.8484 * w ** 0.312216
```

```
Parameter Estimates

Parameter  Estimate  SE Estimate
a           24.8484     0.658102
b            0.3122     0.005964

h = a * w ** b


Lack of Fit

There are no replicates.
Minitab cannot do the lack of fit test based on pure error.
```

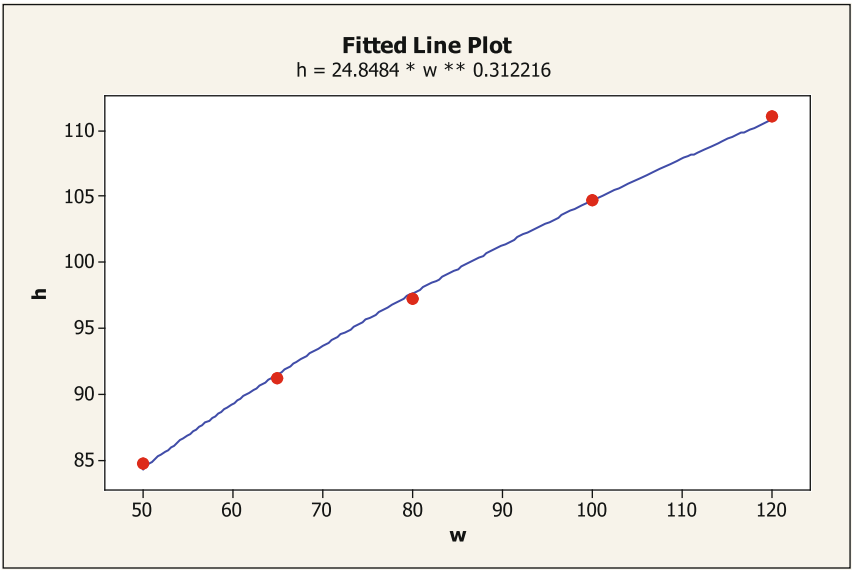


Fig. 7.16 Predicted sitting height h vs. weight W

From the above, the estimated nonlinear regression equation is presented below with corresponding fitted nonlinear regression displayed in Fig. 7.16

$$h = 24.8484\, W^{0.31222}$$

The predicted values as well as the associated residuals are presented in the display below.

```
Data Display

Row    w      h      yhat      Resid1    Fits1
  1   50   84.7    84.347    0.415996   84.284
  2   65   91.2    91.509   -0.278729   91.479
  3   80   97.2    97.606   -0.405598   97.606
  4  100  104.7   104.612    0.051839  104.648
  5  120  111.0   110.707    0.222063  110.778
```

Example 7.16.2

The following example is from Kutner et al. (2005). A hospital administrator wishes to develop a regression model for predicting the degree of long-term recovery after discharge from the hospital for severely injured patients. He intends to use as an explanatory variable, the number of days hospitalized (X) and a response variable being the prognostic index of long-term recovery (Y), with large values of the index reflecting a good prognosis. Data were collected on 15 patients and these are presented in Table 7.13.

Table 7.13 Data on severely injured patients

Patient i	Days hospitalized X_i	Prognostic index Y_i
1	2	54
2	5	50
3	7	45
4	10	37
5	14	35
6	19	25
7	26	20
8	31	16
9	34	18
10	38	13
11	45	8
12	52	11
13	53	8
14	60	4
15	65	6

A scatter plot of Y against X is displayed in Fig. 7.17.

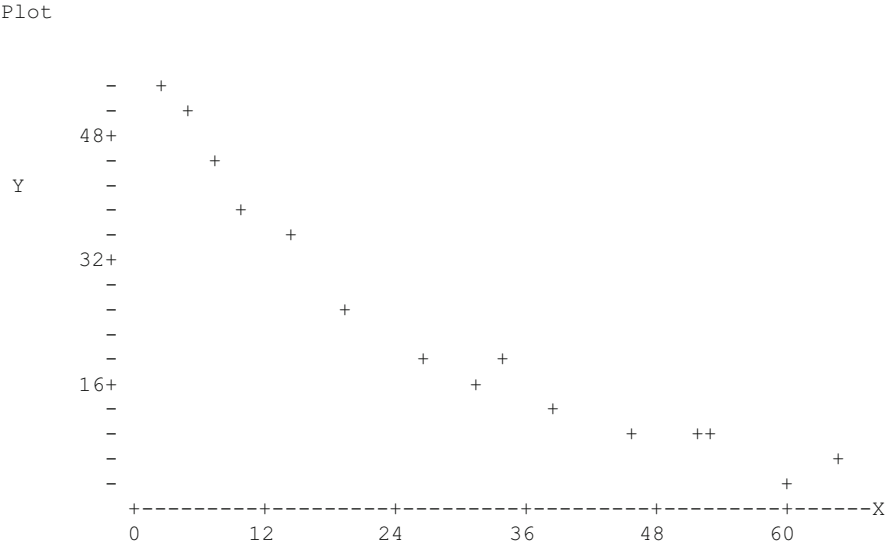


Fig. 7.17 Scatter plot of Y against X

Two possible exponential models are being considered, namely,

$$Y_i = \beta_0 {X_i}^{\beta_1} + e_i \tag{7.64a}$$

$$Y_i = \beta_0 e^{\beta_1 X_i} + e_i \tag{7.64b}$$

Taking logarithms of both sides of the two models, we have for the two models respectively,

$$\ln Y_i = \ln \beta_0 + \beta_1 \ln X_i + \epsilon_i \tag{7.65a}$$

$$\ln Y_i = \ln \beta_0 + \beta_1 X_i + \epsilon_i \tag{7.65b}$$

We decide to implement these models in MINITAB since the procedure involved is similar to that developed in the previous section. To implement the model (7.65a), we first transform Y and X to their logs and then run the regression on the transformed variables. The parameter estimates for this model are:

$$\log \hat{\beta}_0 = 5.0747, \qquad \hat{\beta}_1 = -0.7191.$$

Hence, $\hat{\beta}_0 = e^{5.0747} = 159.9242$. Consequently, the predicted regression equation is:

$$\hat{y}_i = 159.9242\, X_i^{-0.7191}. \tag{7.66}$$

For instance, when $x_i = x_1 = 2$, then $\hat{y}_1 = 159.9242(2)^{-0.7191} = 97.1502$. Other predicted values are similarly obtained and these are plotted against X in Fig. 7.18 with the observed values superimposed. The observed values are denoted with the "+" symbol. Clearly, this model does not fit the data well from the graph.

Similarly, to implement model (7.65b), we again transform Y to its log and then run the regression on the transformed variable against X. The parameter estimates under this model from the MINITAB output are:

$$\log \hat{\beta}_0 = 4.03716, \qquad \hat{\beta}_1 = -0.03797.$$

Hence, $\hat{\beta}_0 = e^{4.03716} = 56.6652$. Consequently, the predicted regression equation is:

$$\hat{y}_i = 56.6652\, e^{(-0.03797\, X_i)}. \tag{7.67}$$

For instance, when $x_i = x_1 = 2$, then $\hat{y}_1 = 56.6652 e^{2(-0.03797)} = 52.5214$. Other predicted values are similarly obtained and these are plotted against X in Fig. 7.19. The observed values are again superimposed on the same graph and are again denoted with the "+" symbol.

The MINITAB nonlinear regression modules generate the following results and accompanying fitted and residuals for both models. Clearly, the second model fits better.

The MINITAB outputs are presented briefly below for both models.

```
Equation

Y = 88.7885 * X ** -0.466221


Parameter Estimates

Parameter  Estimate  SE Estimate
a           88.7885      11.2726
b           -0.4662       0.0578

Y = a * X ** b


Lack of Fit

There are no replicates.
Minitab cannot do the lack of fit test based on pure error.


Summary

Iterations       11
Final SSE   625.715
DFE              13
MSE         48.1320
S           6.93772
--------------------------
Equation

Y = 58.6066 * EXP(-0.0395865 * X)

Parameter Estimates

Parameter  Estimate  SE Estimate
a           58.6066      1.47216
b           -0.0396      0.00171

Y = a * EXP(b * X)


Lack of Fit

There are no replicates.
Minitab cannot do the lack of fit test based on pure error.


Summary

Iterations       13
Final SSE   49.4593
DFE              13
MSE         3.80456
S           1.95053
```

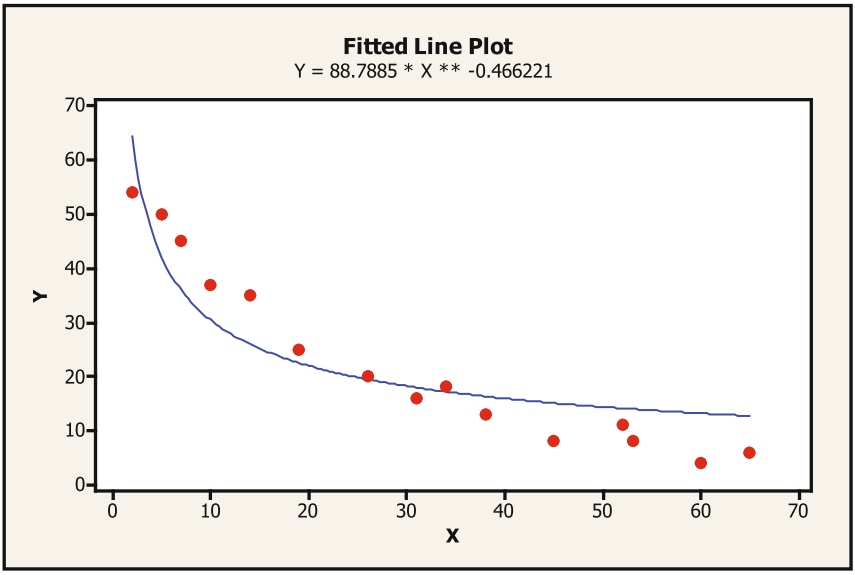


Fig. 7.18 Plot of estimated regressions for model (a)

Data Display

Row	X	Y	Resid1	Fits1	Resid2	Fits2
1	2	54	-10.2703	64.2703	-0.14544	54.1454
2	5	50	8.0741	41.9259	1.91768	48.0823
3	7	45	9.1611	35.8389	0.57770	44.4223
4	10	37	6.6516	30.3484	-2.44796	39.4480
5	14	35	9.0578	25.9422	1.32902	33.6710
6	19	25	2.5004	22.4996	-2.62453	27.6245
7	26	20	0.5614	19.4386	-0.93870	20.9387
8	31	16	-1.9082	17.9082	-1.17864	17.1786
9	34	18	0.8467	17.1533	2.74500	15.2550
10	38	13	-3.2865	16.2865	-0.02098	13.0210
11	45	8	-7.0520	15.0520	-1.86957	9.8696
12	52	11	-3.0708	14.0708	3.51911	7.4809
13	53	8	-5.9464	13.9464	0.80947	7.1905
14	60	4	-9.1627	13.1627	-1.45024	5.4502
15	65	6	-6.6805	12.6805	1.52848	4.4715

7.16.2 Example 7.16.3: Exponential Response Model

For fertilizer response trials, the Mitscherlich response function model:

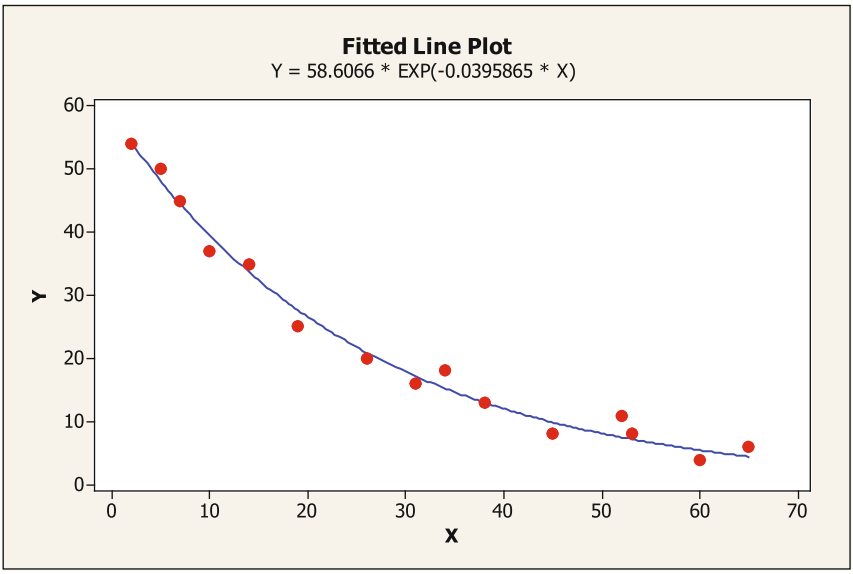


Fig. 7.19 Plot of estimated regressions for model (b)

$$y_i = a[1 - e^{-b(x+c)}] \tag{7.68}$$

has been suggested to be appropriate. We apply the above model to the data presented in Mead and Curnow (1983). The data relate to an experiment investigating the effect of nitrogen fertilization on sugarcane yields. Five nitrogen levels (0, 50, 100, 150, 200) kg/ha in four randomized blocks of five plots each. The yields are presented in the following table.

	Blocks				
Nitrogen	I	II	III	IV	Total
0	60	73	77	72	282
50	125	144	145	116	530
100	152	154	160	141	607
150	182	167	181	185	715
200	198	188	189	182	757
Total	717	726	752	696	2891

Our initial analysis of the above data in MINITAB gives the following results:

```
MTB > ANOVA 'Y' = Block N.

ANOVA: Y versus Block, N

Factor   Type  Levels  Values
Blocks  fixed       4  1, 2, 3, 4
N       fixed       5    0,  50, 100, 150, 200


Analysis of Variance for Y

Source  DF       SS      MS       F      P
Blocks   3    322.9   107.6    1.34  0.309
N        4  35392.7  8848.2  109.77  0.000
Error   12    967.3    80.6
Total   19  36683.0


S = 8.97821   R-Sq = 97.36%   R-Sq(adj) = 95.82%
```

Clearly, the nitrogen effects are very highly significant, while there is very little variation among the blocks. We now consider two alternatives for modeling the response of nitrogen.

(a) Fit a polynomial model using orthogonal polynomials since the levels of nitrogen are equally spaced. Since there are five levels, we can fit a fourth degree polynomial (a quartic model) to the data.
(b) Fit the exponential Mitscherlich response function model in (7.68).

7.16.3 Polynomial Model

The fourth degree polynomial model is implemented in MINITAB, first, by coding the levels of nitrogen into the four components using the table of orthogonal polynomial coefficients and then running the two-way ANOVA model. The results are displayed below.

```
MTB > Code (0) -2 (50) -1 (100) 0 (150) 1 (200) 2 'N' c4
MTB >  Code (0) 2 (50) -1 (100) -2 (150) -1 (200) 2 'N' c5
MTB >  Code (0) -1 (50) 2 (100) 0 (150) -2 (200) 1 'N' c6
MTB >  Code (0) 1 (50) -4 (100) 6 (150) -4 (200) 1 'N' C7

Data Display

Row    N  Block    Y  LN  QN  CN  QQN
  1    0      1   60  -2   2  -1    1
  2   50      1  125  -1  -1   2   -4
  3  100      1  152   0  -2   0    6
  4  150      1  182   1  -1  -2   -4
  5  200      1  198   2   2   1    1
  6    0      2   73  -2   2  -1    1
  7   50      2  144  -1  -1   2   -4
  8  100      2  154   0  -2   0    6
  9  150      2  167   1  -1  -2   -4
 10  200      2  188   2   2   1    1
 11    0      3   77  -2   2  -1    1
 12   50      3  145  -1  -1   2   -4
```

```
 13  100      3  160   0  -2   0    6
 14  150      3  181   1  -1  -2   -4
 15  200      3  189   2   2   1    1
 16    0      4   72  -2   2  -1    1
 17   50      4  116  -1  -1   2   -4
 18  100      4  141   0  -2   0    6
 19  150      4  185   1  -1  -2   -4
 20  200      4  182   2   2   1    1

MTB > GLM 'Y' = Block LN QN CN QQN;
SUBC>   Covariates 'LN' 'QN' 'CN' 'QQN';
SUBC>   Brief 2 .

General Linear Model: Y versus Block

Factor  Type   Levels  Values
Block   fixed       4  1, 2, 3, 4


Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF   Seq SS   Adj SS   Adj MS       F      P
Block       3    323.0    323.0    107.7    1.34  0.309
Linear      1  32205.6  32205.6  32205.6  399.53  0.000
Quadratic   1   2592.2   2592.2   2592.2   32.16  0.000
Cubic       1    275.6    275.6    275.6    3.42  0.089
Quartic     1    319.3    319.3    319.3    3.96  0.070
Error      12    967.3    967.3     80.6
Total      19  36682.9


S = 8.97821   R-Sq = 97.36%   R-Sq(adj) = 95.82%
```

The analysis indicates that a second degree polynomial will be suitable for the response model. This and the model in (7.68) are implemented both on the average yields of each nitrogen level. Both the quadratic regression analysis and the nonlinear model analysis are presented below.

```
Polynomial Regression Analysis: y versus x

The regression equation is
y = 74.19 + 1.112 x - 0.002721 x**2
S = 8.62347   R-Sq = 98.3%   R-Sq(adj) = 96.6%

Analysis of Variance

Source       DF       SS       MS      F      P
Regression    2  8699.45  4349.72  58.49  0.017
Error         2   148.73    74.36
Total         4  8848.17


Sequential Analysis of Variance

Source     DF       SS      F      P
Linear      1  8051.41  30.32  0.012
Quadratic   1   648.04   8.71  0.098
-------------------------------------------------
NONLINEAR MODEL

MTB > NLinear;
SUBC>   Response 'Y';
SUBC>   Continuous 'x';
SUBC>   Parameter "a" 100;
```

```
SUBC>   Parameter "b" .01;
SUBC>   Parameter "c" 30;
SUBC>   Expectation a * ( 1-EXP(-b*( x +c)) );
SUBC>   NoDefault;
SUBC>   GFCurve;
SUBC>   TMethod;
SUBC>   TStarting;
SUBC>   TConstraints;
SUBC>   TEquation;
SUBC>   TParameters;
SUBC>   TSummary;
SUBC>   TPredictions.

Nonlinear Regression: y = a * (1 - EXP(-b * (x + c)))

Method

Algorithm       Gauss-Newton
Max iterations           200
Tolerance            0.00001


Starting Values for Parameters

Parameter  Value
a            100
b           0.01
c              30
Equation

y = 202.988 * (1 - EXP(-0.0108615 * (x + 40.2389)))
Parameter Estimates

Parameter  Estimate  SE Estimate
a           202.988      14.4450
b             0.011       0.0029
c            40.239      10.3793

y = a * (1 - EXP(-b * (x + c)))

Summary

Iterations       7
Final SSE   86.5870
DFE              2
MSE         43.2935
S           6.57978
```

Based on both analyses, the estimated regression models are, respectively:

$$\hat{y} = 74.19 + 1.112\ N - 0.002721\ N^2$$

$$\hat{y} = 202.988\left[1 - e^{-0.011(N+40.239)}\right]$$

Data Display

Row	N	y	RESI2	FITS2	Resid2	Fit2
1	0	70.50	-3.69286	74.193	-1.37082	71.871
2	50	132.50	9.52143	122.979	5.68564	126.814
3	100	151.75	-6.40714	158.157	-6.98433	158.734
4	150	178.75	-0.97857	179.729	1.47147	177.279
5	200	189.25	1.55714	187.693	1.19804	188.052

The residuals and fitted values for both models are displayed below. Clearly, model 2 fits better. Thus, we see in this example that the Mitscherlich response model fits much better. Its Error SS is 86.5870 as against 148.73 for the quadratic. The plots of estimated regression models in both cases are presented in Figs. 7.20 and 7.21 respectively.

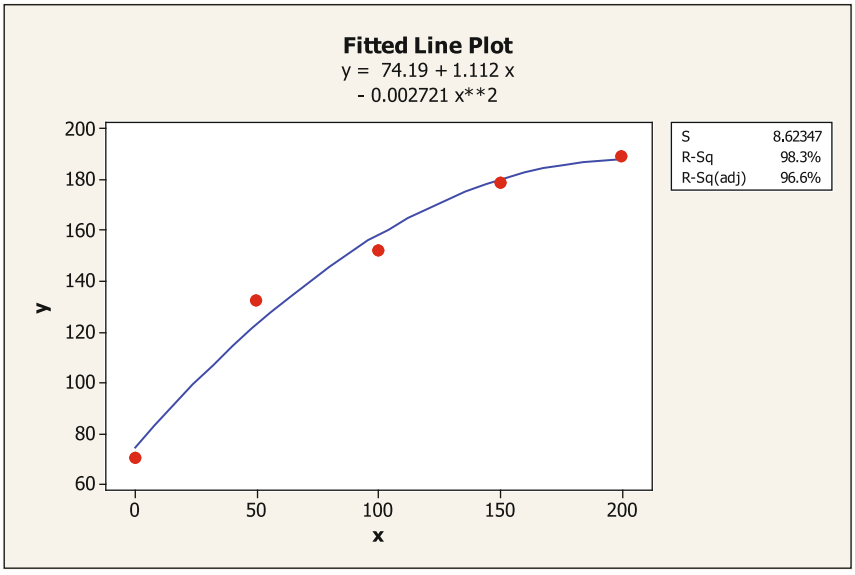


Fig. 7.20 Plot of estimated quadratic model

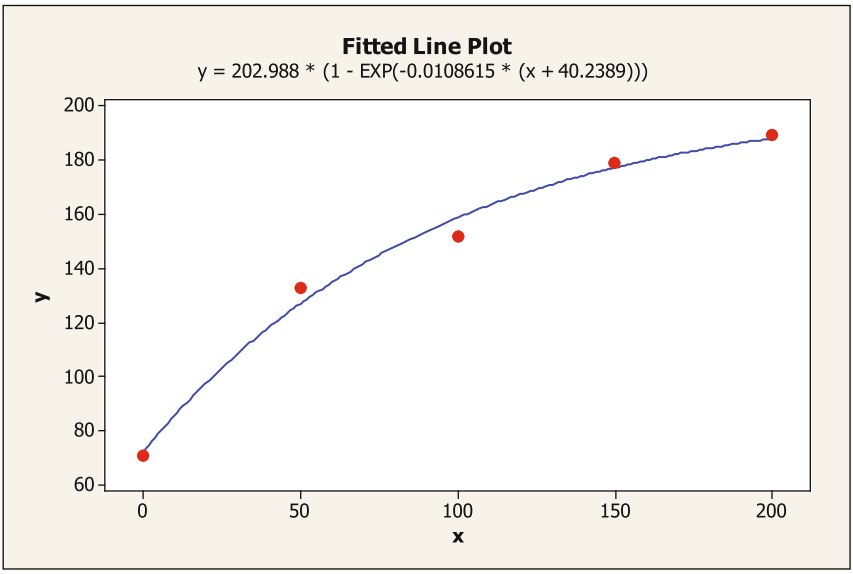


Fig. 7.21 Plot of estimated Mitscherlich response model

Other nonlinear growth curves include the logistic curve that will be discussed in Chap. 17. Others are:

- Drug Responsiveness Model: These are often used in pharmacological settings and the model describes the effects of varying dose levels of drugs. The model has the form

$$y_i = \beta_0 - \frac{\beta_0}{1 + \left(\frac{x_i}{\beta_2}\right)^{\beta_1}} + \varepsilon_i. \tag{7.69}$$

  In the above model, x_i is the dosage at level i and y is the response variable as a percentage of maximum possible responsiveness. β_0 describes the expected response at the dose saturation, β_2 is the concentration that produces a half-maxima response; and the β_1 parameter determines the slope of the function.

Example 7.16.4 (Drug Responsiveness)

The following example is taken from Kutner et al. (2005, p. 550) (reproduced with permission of The McGraw-Hill Companies) and relates to a pharmacologist modeling the responsiveness to a drug using the nonlinear regression model discussed in (7.69), which is reproduced below.

$$y_i = \beta_0 - \frac{\beta_0}{1 + \left(\frac{x_i}{\beta_2}\right)^{\beta_1}} + \varepsilon_i.$$

The data for 19 cases at 13 dose levels are presented in Table 7.14. We see that the observations are replicated at $x = 3.5, 4.0, 4.5, 5.0, 5.5$, and 6.0. The plot of the observed responsiveness (y) against dose levels is presented in the MINITAB program with partial output.

Table 7.14 Drug responsiveness data

Dose	Responsiveness	
x	y	
1.0	0.5	
2.0	2.3	
3.0	3.4	
3.5	11.5	10.9
4.0	24.0	25.3
4.5	39.6	37.9
5.0	54.7	56.8
5.5	70.8	68.4
6.0	82.1	80.6
6.5	89.2	
7.0	94.8	
8.0	96.2	
9.0	96.4	

To fit the model in (7.69), we realize that we do not have initial values. Therefore, we use starting values $\beta_0 = 0, \beta_1 = 0$, and $\beta_2 = 0.1$. The choice of $\beta_2 = 0.1$ is motivated by the fact that we do not want the denominator in the expression involving x to be zero. The results of this analysis together with computed predicted values and residuals are in the following SAS output.

```
MTB > Name C3 "Resid1" C4 "Fits1".
MTB > NLinear;
SUBC>   Response 'y';
SUBC>   Continuous 'x';
SUBC>   Parameter "a" 100;
SUBC>   Parameter "c" 4;
SUBC>   Parameter "b" 5;
SUBC>   Expectation a-(a / ( 1 + ( x / c ) ** b ));
SUBC>   NoDefault;
SUBC>   GFCurve;
SUBC>   GFourPack;
SUBC>   TMethod;
SUBC>   TStarting;
SUBC>   TConstraints;
SUBC>   TEquation;
SUBC>   TParameters;
SUBC>   TCorrelation;
SUBC>   TSummary;
SUBC>   TPredictions;
SUBC>   Residuals 'Resid1';
SUBC>   Fits 'Fits1'.

Nonlinear Regression: y = a - a / (1 + (x / c) ** b)

Method

Algorithm       Gauss-Newton
Max iterations           200
Tolerance            0.00001

Starting Values for Parameters
```

```
Parameter  Value
a            100
c              4
b              5
```

Equation

```
y = 100.34 - 100.34 / (1 + (x / 4.81554) ** 6.48024)
```

Parameter Estimates

```
Parameter  Estimate  SE Estimate
a           100.340      1.17407
c             4.816      0.02801
b             6.480      0.19431
```

```
y = a - a / (1 + (x / c) ** b)
```

Correlation Matrix for Parameter Estimates

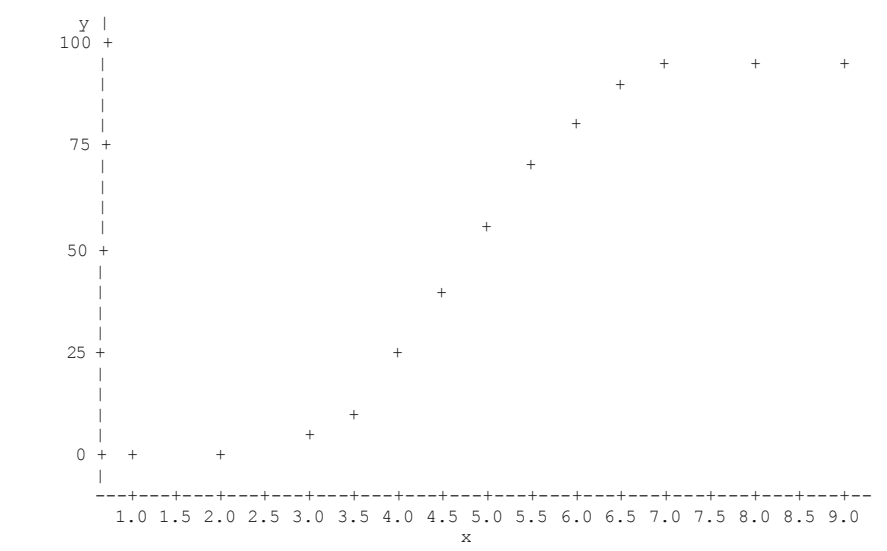
```
           a          c
c   0.817768
b  -0.696059  -0.531430
```

Lack of Fit

```
Source         DF       SS       MS     F      P
Error          16  35.7149  2.23218
  Lack of Fit  10  27.0349  2.70349  1.87  0.229
  Pure Error    6   8.6800  1.44667
```

Summary

```
Iterations       8
Final SSE  35.7149
DFE             16
MSE        2.23218
S          1.49405
```



The estimated nonlinear regression equation is therefore

$$\hat{y}_i = 100.340 - \frac{100.340}{1 + \left(\frac{x_i}{4.816}\right)^{6.480}}.$$

The estimated curve is plotted in Fig. 7.22 with the observed values superimposed over it. Clearly this model fits the data very well.

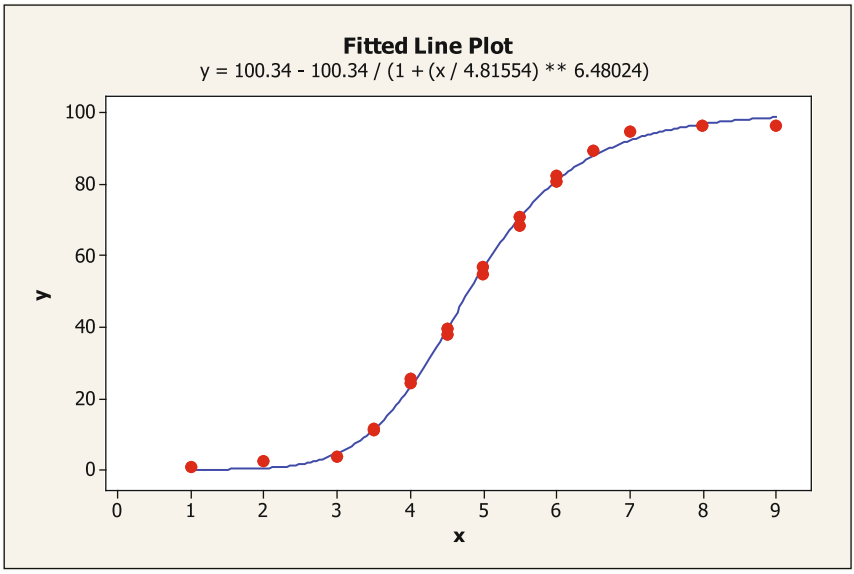


Fig. 7.22 Estimated drug responsiveness curve

Since the observations are replicated, the lack of fit test provided above gives a p value of 0.229 which clearly indicates that the null hypothesis of the adequacy of the model holds. Also the following predicted values and residuals indicate a very good fit of this model to the data.

Data Display

Row	x	y	Resid1	Fits1
1	1.0	0.5	0.49622	0.0038
2	2.0	2.3	1.96344	0.3366
3	3.0	3.4	-1.06541	4.4654
4	3.5	11.5	0.23467	11.2653
5	3.5	10.9	-0.36533	11.2653
6	4.0	24.0	0.81711	23.1829
7	4.0	25.3	2.11711	23.1829
8	4.5	39.6	0.27276	39.3272
9	4.5	37.9	-1.42724	39.3272
10	5.0	54.7	-1.55059	56.2506
11	5.0	56.8	0.54941	56.2506
12	5.5	70.8	0.26922	70.5308
13	5.5	68.4	-2.13078	70.5308
14	6.0	82.1	1.21243	80.8876
15	6.0	80.6	-0.28757	80.8876
16	6.5	89.2	1.42581	87.7742
17	7.0	94.8	2.62351	92.1765
18	8.0	96.2	-0.53400	96.7340
19	9.0	96.4	-2.22627	98.6263

We also present the diagnostics test for the residuals from the model in Fig. 7.23. The pattern of the residuals suggests randomness and the normality plot indicates that we can reasonably assume normality of the error term.

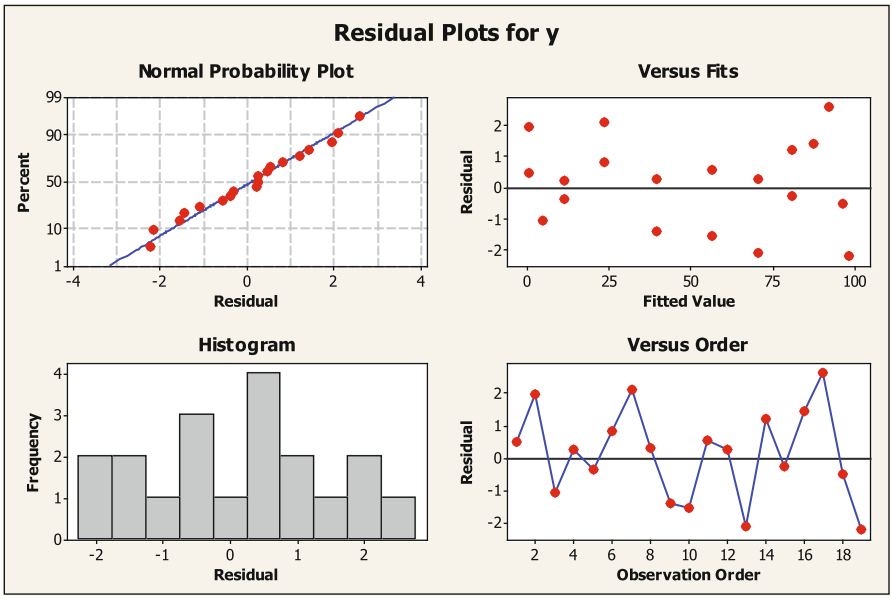


Fig. 7.23 Various residual diagnostics plot arising from the fitted model

While we have considered the above models in this section, there are several other nonlinear models to describe the relationship between an

explanatory variable x and a response variable y. For instance, the relationship between crop yield y and the spacing between rows of plants, concentration Y of a drug in the bloodstream, and time x after the drug is injected, and the rate Y of a chemical reaction and the amount x of catalyst used have the following relationships, for instance (Graybill and Iyer 1994) are:

$$y_i = \frac{1}{(b_1 + b_2 x_i)^{b_3}} \tag{7.70a}$$

$$y_i = \frac{1}{b_1 + b_2 x_i + b_3 x_i^2} \tag{7.70b}$$

$$y_i = \frac{1}{b_1 + b_2 x_i^{b_3}}. \tag{7.70c}$$

We display below the graphs of (7.70b) and (7.70c) for the given parameters. In Fig. 7.24,the upper graph has $b_1 = 8.94, b_2 = -22.4, b_3 = 16$, while the dotted curve has $b_1 = 8, b_2 = -8, b_3 = 1$. Similarly, the curves in Fig. 7.25, the upper graph has $b_1 = 1, b_2 = 6, b_3 = 3$, while the dotted curve has $b_1 = 1.2, b_2 = 9, b_3 = 0.9$.

7.17 Other Special Nonlinear Models

Growth models play important roles in biological applications. We list below some growth models:

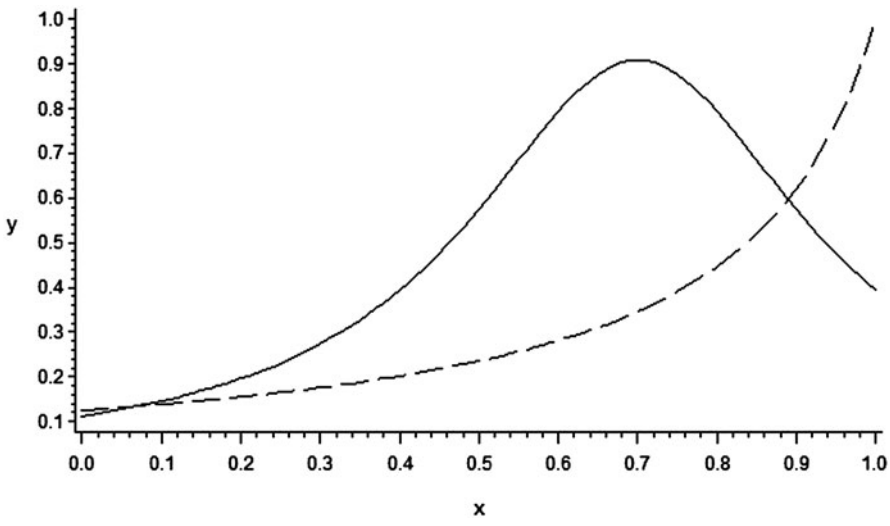


Fig. 7.24 Plot of typical family of curves in (7.70b)

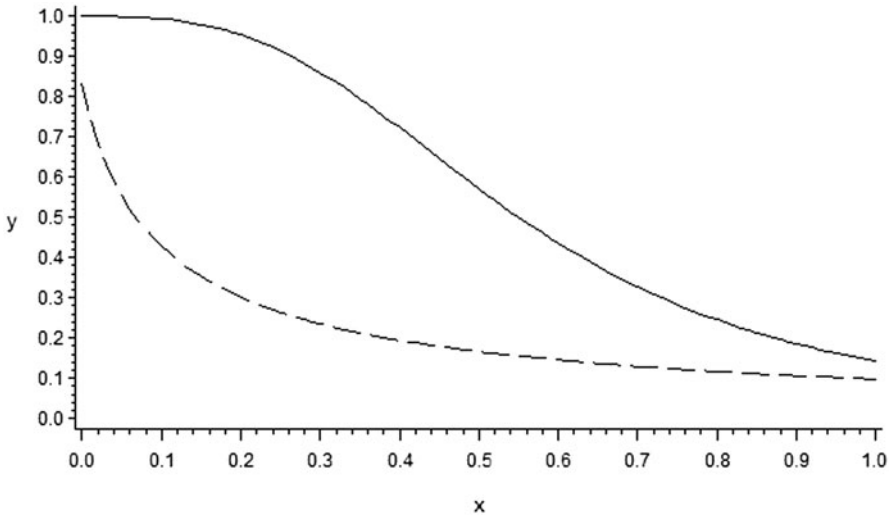


Fig. 7.25 Plot of typical family of curves in (7.70c)

7.17.1 The Logistic Growth Model

The logistic growth model has a continuous observation y being related to an explanatory variable x by the relationship:

$$y_i = \frac{a}{1 + be^{-kx_i}} + \varepsilon_i. \tag{7.71}$$

The model in (7.71) is similar to the logistic model used for binary response outcome in Chap. 17, except that the values of b and k are positive in this case and a is defined as the *limiting growth*, that is, a value that y approaches as the x becomes larger. We may note that the kx above may be replaced by $k\mathbf{x}$, where $\mathbf{x}$ has several explanatory variables. A typical graph of this model is displayed in Fig. 7.26 for the parameters $a = 10$, $b = 5$, and $k = 4, 2, 1, 0.25$ respectively from the upper curve.

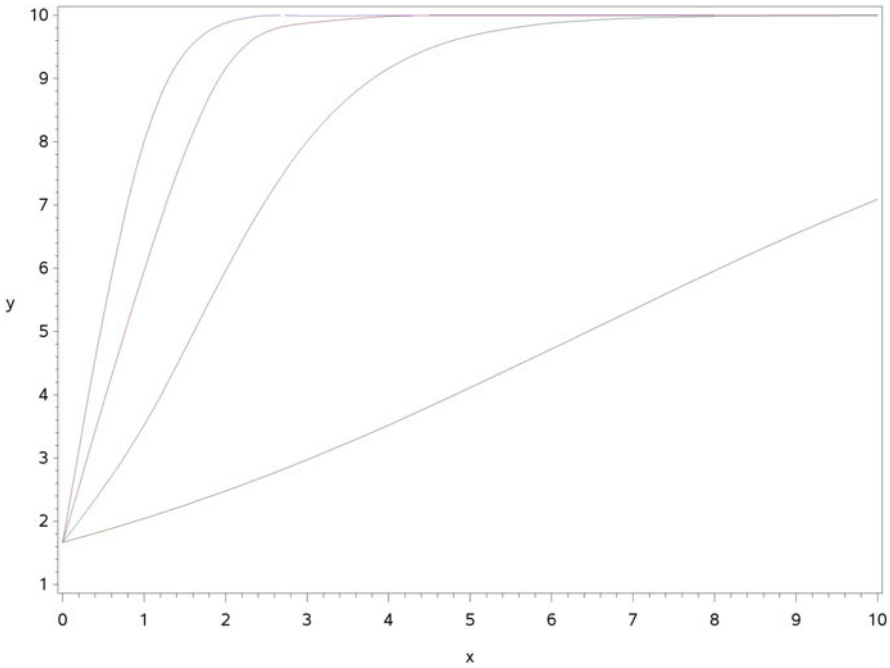


Fig. 7.26 Plot of typical family of curves in (7.71)

7.17.2 The Gompertz Growth Model

This model is defined as:

$$y_i = ae^{[-b\,e^{-kx_i}]} + \varepsilon_i. \tag{7.72}$$

This is similar to the double exponential model we discussed earlier and a again is the limiting growth, while for $x = 0$, $y = a\,e^{-b}$. Other important growth curves are the:

- Richards growth model having the form:

$$y_i = \frac{a}{[1 + b\,e^{-kx_i}]^{1/\delta}} + \varepsilon_i \tag{7.73}$$

- Weibull growth model with the form:

$$y_i = a - b\,e^{-\gamma x_i^{\delta}} + \varepsilon_i \tag{7.74}$$

The above are just some of the growth curves that have proved very useful in biological applications and there are certainly a richer list of growth models not covered in this text.

7.18 Polynomial Regressions

Sometimes the true relationships between the dependent variable Y and independent variable X may not be of the form

$$Y = \beta_0 + \beta_1 x \quad \text{that is, a straight line}$$

It may be curvilinear, that is, of the form

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \quad \text{, etc.}$$

We can reduce this model to a simple multiple regression problem of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad \text{, etc.}$$

where $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, etc. We give as an example, the following data:

Example 7.18.1

Consider the hypothetical data below. It is assumed that the model $y = \alpha + \beta x$ is appropriate and we wish to examine the adequacy of this model.

Y	4	3	8	18	22	24	24	18	13	10	16
$X = x_1$	2	2	2	3	3	4	5	5	6	6	6
$X^2 = x_2$	4	4	4	9	9	16	25	25	36	36	36

To fit the model $Y = \beta_0 + \beta_1 x + e$, we would have,

$$S_{x_1} = 28 \quad S_{x_1 y} = 50 \quad S_{yy} = 570.727$$

Hence, $\hat{\beta}_0 = 7.403$ and $\hat{\beta}_1 = 1.786$, with R^2 computed as:

$$R^2 = \frac{\text{Fitted SS}}{\text{Total SS}} = \frac{S_{x_1 y}^2}{S_{x_1}\, S_{yy}} = \frac{50^2}{(28 \times 570.727)}$$
$$= 0.1564$$

It is obvious from this low value of R^2 that this model is grossly inadequate.

A plot of residuals against fitted values in Fig. 7.27 reveals that a quadratic term will be necessary to model the relationship between Y and X. The model of interest is therefore

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e. \tag{7.75}$$

To fit these models we compute and obtain the following:

$$S_{x_2} = 1820.727, \qquad S_{x_1 x_2} = 224, \qquad S_{x_2 y} = 290.727.$$

Hence,

$$\hat{\beta}_1 = 32.216, \quad \hat{\beta}_2 = -3.804 \quad \text{and} \quad \hat{\beta}_0 = -50.94$$

and the estimated regression equation is

$$\hat{Y} = -43.7758 + 32.2161x - 3.8038x^2 \tag{7.76}$$

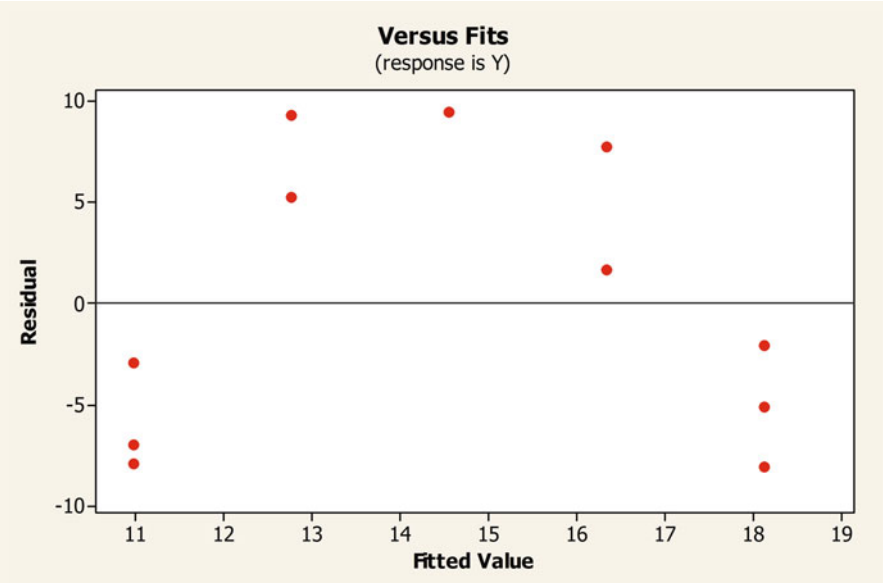


Fig. 7.27 Plot of the estimated regression equation

$$\begin{aligned}\text{Fitted SS} &= \hat{\beta}_1 S_{x_1 y} + \hat{\beta}_2 S_{x_2 y} \\ &= 504.87\end{aligned}$$

$$R^2 = \frac{504.87}{570.727} = 0.885$$

which is a much improved fit than the earlier model. We can, in fact, test that the inclusion of β_2 into the model is very significant by using the partial F test discussed earlier in this chapter which is equivalent to the result presented in MINITAB below for testing $H_0 : \beta_2 = 0$ vs $H_a : \beta_2 \neq 0$. This test provides a p value of 0.0000. We would therefore reject H_0 at $\alpha = .001$ level of significance.

```
Analysis of Variance

Source            DF         SS         MS         F       P
Regression         2    504.937    252.469   30.6998   0.000
Error              8     65.790      8.224
Total             10    570.727


Source      DF     Seq SS          F       P
Linear       1     89.286     1.6691   0.229
Quadratic    1    415.651    50.5426   0.000
```

A graph of the estimated regression equation is presented in Fig. 7.28. It is important to sound a word of warning regarding fitting polynomial regression. The explanatory variables are correlated and this usually lead to the problem of *multicollinearity* which is not being discussed in this text.

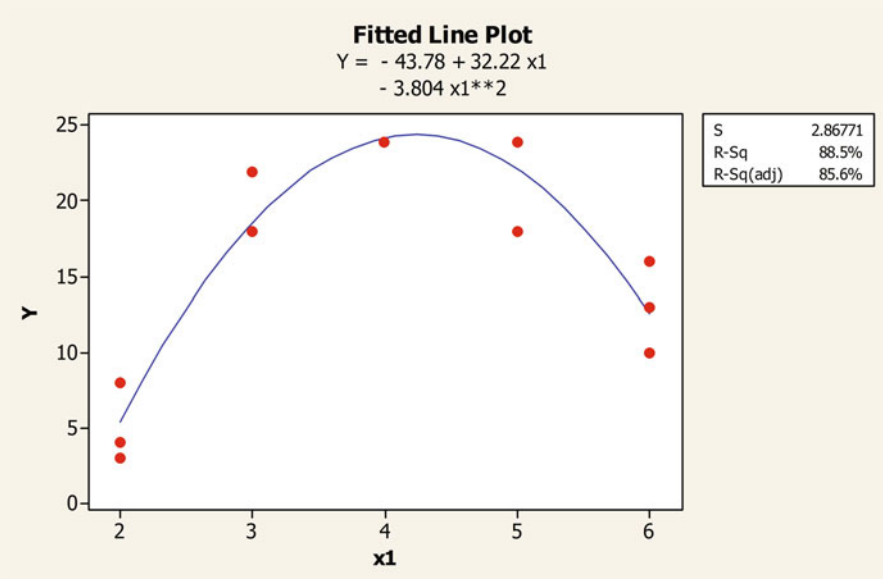


Fig. 7.28 Plot of the estimated quadratic model

7.19 Exercises

1. The data in this example relate to the marks obtained in a 100 level examination by seven students in mathematics (X) and physics (Y). It is envisaged that a simple linear relationship exists between marks obtained in physics and mathematics.

X	38	51	19	53	39	38	66
Y	50	72	36	64	52	56	80

   (a) Find the summary statistics S_{yy}, S_{xx}, and S_{xy}.
   (b) Obtain the Pearson's sample correlation r.
   (c) Fit a simple regression to the above data and generate the relevant analysis of variance table.
   (d) How would you rate your regression line?

2. Twenty plots, each of 10.4 m, were randomly chosen in a large field of corn. For each plot, the plant density and the mean cob weight were observed. The results are given in the table below:

Plant density X	Cob weight Y	Plant density X	Cob weight Y
137	212	173	194
107	241	124	241
132	215	157	196
137	208	112	230
135	225	184	193
115	250	112	224
173	185	124	248
103	241	80	257
102	237	165	200
65	282	160	190
132	220	85	244
149	206	157	208
85	246	119	224

Data adapted From Samuel et al. (2005)

(a) Use MINITAB to fit a simple linear regression to the data.
(b) Conduct model adequacy.
(c) Predict Y for values of $X = 58$, 120, and 134.
(d) What is the error mean square?

3. **Galileo's Experiment** The following is part of the data submitted by Dickey, D.A. to the Journal of Statistics Education data archive (see also Dickey and Arnold 1995). The data are from an experiment in which Galileo rolled a ball down an inclined plane above a floor. The data contain the height (y) of the ball rolled down and the horizontal distance (x) traveled before landing.

x	573	534	495	451	395	337	253
y	1000	800	600	450	300	200	100

To model the height (y) as a function of the distance (x), the following regression model is suggested: $y = \left(\beta_0 x^2\right)/(1 + \beta_1 x) + \varepsilon$.

(a) Fit the nonlinear regression model to the data.
(b) Construct an approximate 95 % confidence interval for the parameters.
(c) Obtain the fitted values and the residuals from the nonlinear regression models. Do the residuals follow a normal distribution? Comment on your findings.

4. **Galileo's Experiment** The following is part of the data submitted by Dickey, D.A. to the Journal of Statistics Education data archive (see also Dickey and Arnold 1995). The data are from an experiment in which Galileo rolled a ball down an inclined plane above a floor and the ball crossed a horizontal shelf built into the end of the ramp. The data contains the height (y) of the ball rolled down and the horizontal distance (x) traveled before landing.

x	1500	1340	1328	1172	800
y	1000	828	800	600	300

To model the height (y) as a function of the distance (x), the following regression model is suggested: $y = \left(\beta_0 x^2\right)/(1 + \beta_1 x) + \varepsilon$.

(a) Fit the nonlinear regression model to the data.
(b) Construct an approximate 95 % confidence interval for the parameters.
(c) Obtain the fitted values and the residuals from the nonlinear regression models. Do the residuals follow a normal distribution? Comment on your findings.

5. The following artificial data are reproduced from Mead and Curnow by permission. The data relate to change of activity of microorganism with time.

$y =$ activity	10	13	22	24	29	35	32	36
$x =$ time	2.5	5	10	20	30	40	60	80

Fit the Michaelis–Menten model to the above data.

6. A study is run to develop an equation by which the concentration of estrone in saliva can be used to predict the concentration of this steroid in free plasma. These data are obtained on 14 healthy males.

Individual	(Concentration of estrone in saliva, pg/ml) X	(Concentration of estrone in free plasma, pg/ml) Y
1	7.4	30.0
2	7.5	25.0
3	8.5	31.5
4	9.0	27.5
5	9.0	39.5
6	11.0	38.0
7	13.0	43.0
8	14.0	49.0
9	14.5	55.0
10	16.0	48.5
11	17.0	51.0
12	18.0	64.5
13	20.0	63.0
14	23.0	68.0

A model of the form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

is to be used to predict estrone in free plasma from estrone in saliva. Use MINITAB for this question to answer the following questions:

(a) What is the estimated value of β_1? What is the interpretation of this value? Find the 95 % confidence interval for β_1.
(b) Test the null hypothesis $H_0 : \beta_1 = 0$ against the two-sided alternative that $H_1 : \beta_1 \neq 0$, using $\alpha = 0.05$ and state the p value. What is the interpretation of the results of this test?
(c) Use the estimated regression equation to predict the estrone level in free plasma in two males whose saliva estrone levels are 17.5 and 24, respectively.
(d) Explain what R^2 tells us and how it is computed?
(e) Estimate the variance of ε.

7. In an experiment to study the growth behavior of protozoa colonization in a particular lake, 15 sponges were placed in a lake and 3 were gathered at a time. The number of protozoa were counted at 1, 3, 6, 15, and 21 days. The equation of the following form is suggested for the growth

$$Y = S_{eq}(1 - e^{kt})$$

where:

- Y: total protozoa on the sponge
- S_{eq}: species equilibrium constant
- k: parameter that measures how quickly growth rises
- t: time, number of days

The data are displayed below where Y is total number of protozoa counted

Obs.	Day	Y	Obs.	Day	Y	Obs.	Day	Y
1	1	17	6	3	25	11	15	33
2	1	21	7	6	33	12	15	33
3	1	16	8	6	31	13	21	39
4	3	30	9	6	32	14	21	35
5	5	25	10	15	34	15	21	36

(i) Estimate S_{eq} and k.
(ii) Give estimated standard errors of the parameters.

8. An experimenter wished to determine the relationship between temperature and heartbeat rate in the common grass frog, *Rana pipiens*. The temperature was manipulated in 2 ° increments ranging from 2 to 18 °C, with heartbeat rates recorded at each interval.

Recording number	Temperature (°C)	Heartbeat (/min)
1	2	5
2	4	11
3	6	11
4	8	14
5	10	22
6	12	23
7	14	32
8	16	29
9	18	32

A model of the form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

is to be used to predict heartbeat rate from body temperature. Use MINITAB to answer the accompanying questions.

(a) What are the estimated values of β_0 and β_1? What is the interpretation of this estimate of β_1? Give 95 % confidence limits for these estimates.
(b) Test the null hypothesis $H_0 : \beta_1 = 0$ against the two-sided alternative that $H_1 : \beta_1 \neq 0$, using $\alpha = 0.05$ and state the p value. What is the interpretation of the results of this test?
(c) Use the estimated regression equation to predict the mean heartbeat per minute for temperatures 9 and 20 °C respectively.
(d) Obtain the upper and lower 95 % confidence limits for the heart rate expected in *Rana pipiens* at 10 °C.
(e) Explain what R^2 tells us and how it is computed.
(f) Obtain the sample correlation coefficient r and interpret it.

9. In a multiple regression problem involving several explanatory variables, the following model is proposed:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i$$

$i = 1, 2, \cdots, 16$.
The analysis yielded the following partially completed ANOVA table.

Source	d.f.	SS	MS	F
Regression			250	
error				
Total		1500		

(a) Complete the above ANOVA table.
(b) State the hypotheses being tested by the F value above.
(c) Compute the coefficient of determination.

10. Are a person's brain size and body size predictive of his/her intelligence? Data on Y based on the performance IQ (PIQ) scores from Wechsler adult intelligence scale (revised), brain size (X_1) based on the count from MRI scans (given as count/10,000), and body size measured in height (X_2) in inches and weight (X_3) in pounds on 38 college students are displayed below.

Y	X1	X2	X3
124	81.69	64.5	118
150	103.84	73.3	143
128	96.54	68.8	172
134	95.15	65.0	147
110	92.88	69.0	146
131	99.13	64.5	138
98	85.43	66.0	175
84	90.49	66.3	134
147	95.55	68.8	172
124	83.39	64.5	118
128	107.95	70.0	151
124	92.41	69.0	155
147	85.65	70.5	155
90	87.89	66.0	146
96	86.54	68.0	135
120	85.22	68.5	127
102	94.51	73.5	178
84	80.80	66.3	136
86	88.91	70.0	180
84	90.59	76.5	186
134	79.06	62.0	122
128	95.50	68.0	132
102	83.18	63.0	114
131	93.55	72.0	171
84	79.86	68.0	140
110	106.25	77.0	187
72	79.35	63.0	106
124	86.67	66.5	159
132	85.78	62.5	127
137	94.96	67.0	191
110	99.79	75.5	192
86	88.00	69.0	181
81	83.43	66.5	143
128	94.81	66.5	153
124	94.94	70.5	144
94	89.40	64.5	139
74	93.00	74.0	148
89	93.59	75.5	179

A model of the form below is found to be appropriate:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

Use MINITAB to fit the above model and conduct the necessary diagnostic tests. From your fitted model,

(a) Write down the estimated regression parameters.
Give 90 % confidence intervals for these parameters.
(b) Test whether the explanatory variables X_1, X_2, and X_3 are important in the model. Use $\alpha = 0.05$.
(c) How was R^2 computed?
(c) Interpret $\hat{\beta}_1$ and $\hat{\beta}_3$.
(d) Test the hypotheses at $\alpha = 0.10$ from your confidence interval results in (a).

$$H_0 : \beta_3 = 0$$
$$H_a : \beta_3 \neq 0$$

(e) In your opinion, which variables are the most important in predicting Y, a person's IQ?
(f) Obtain the residuals for observations numbered 8 and 32. How were they obtained from the available output?

Chapter 8
Categorical Data Analysis

8.1 Introduction

Categorical variables may have categories which are naturally ordered called ordinal variables or those that have no natural order called nominal variables. For example, the variable "weight" with categories "small," "medium," and "big" is an ordinal variable, so also is the attitudinal variable with categories "agree," "neutral," and "disagree." On the other hand, variables such as "sex" and "color" of flowers which have no natural order are examples of nominal variables. In this chapter, emphasis will be placed on testing the agreement of frequency arising from data from experiments with known distribution (e.g., Poisson and Binomial) and analysis of two-way contingency tables.

8.1.1 Tests of Goodness of Fit

Consider the one-way classification in Table 8.1 in which a sample of size n is randomly distributed in each of the k classes.

Table 8.1 Table of observed frequencies and underlying probability distribution

Classes					
1	2	3	$\cdots$	k	Total
x_1	x_2	x_3	$\cdots$	x_k	n
p_1	p_2	p_3	$\cdots$	p_k	1

Let p_i $i = 1, 2, \cdots, k$ be the probability of an individual falling into the i-th class. A problem that often arises in research is the testing of the compatibility of a set of observed and theoretical frequencies. Let the observed frequency in class i be denoted by x_i such that $\sum x_i = n$, then to test the hypothesis that the observed frequencies are distributed according to specified probability p_i^0, we use the classical Pearson's X^2 test statistic given by

$$X^2 = \frac{\sum (O_i - E_i)^2}{E_i} \tag{8.1}$$

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_8,

where $O_i = x_i$ is the observed frequency in the ith class and $E_i = np_i^0$ is the corresponding expected frequency under the null hypothesis of specified probabilities in the ith class.

Alternatively, we often use the likelihood ratio test statistic:

$$G^2 = 2\sum O_i \log(O_i/E_i) \tag{8.2}$$

where np_i^0 is the expected frequency under the null hypothesis. Under the null hypothesis, X^2 or G^2 has asymptotically an χ^2 distribution with $k-1$ degrees of freedom, provided the expected values are not too small. Thus, the test is accomplished by comparing the observed value of X^2 with the tabulated χ^2 distribution with $(k-1)$ d.f.

Example 8.1.1

In a particular genetic experiment, the observations were classified as follows:

	Classes			
	A	B	C	D
x_i	99	33	24	4

The genetic theory calls for a 9 : 3 : 3 : 1 ratio. Hence the underlying distribution calls for,

$$p_1 = \frac{9}{16}, \quad p_2 = \frac{3}{16}, \quad p_3 = \frac{3}{16}, \quad \text{and} \quad p_4 = \frac{1}{16}$$

such that $\sum p_i = 1$. In this example, $n = \sum x_i = 99 + 33 + 24 + 4 = 160$.

Under H_0, the expected frequencies are respectively $160 \times \frac{9}{16} = 90, 30, 30,$ and 10 corresponding respectively to observed values, $O_1 = 99, O_2 = 33, O_3 = 24$, and $O_4 = 4$. Thus,

$$\begin{aligned} X^2 &= \frac{\sum (O_i - E_i)^2}{E_i} \\ &= \frac{(99-90)^2}{90} + \frac{(33-30)^2}{30} + \frac{(24-30)^2}{30} + \frac{(4-10)^2}{10} \\ &= 6.0. \end{aligned}$$

Since $6.0 < \chi^2_3 = 7.81$, we would fail to reject H_0 and conclude that the results of the experiment confirm the genetic theory at $\alpha = .05$ level of significance.

Similarly, the corresponding likelihood ratio test statistic is computed as:

$$\begin{aligned} G^2 &= 2[99\log(99/90) + 33\log(33/30) + 24\log(24/30) + 4\log(4/10) \\ &= 2[9.436 + 3.145 - 5.355 - 3.665] \\ &= 7.14. \end{aligned}$$

The result from the G^2 also indicates that we would fail to reject the null hypothesis at $\alpha = 0.05$ level of significance.

8.2 The 2×2 Contingency Table

Categorical data where people or things are classified simultaneously by two or more attributes are discussed in this section. The results of such a cross-classification can be conveniently arranged as a table of counts known as a *Contingency Table*. When just two classificatory variables are considered, the table is called a two-way (2D) contingency table. The simplest two-way table is the 2×2 contingency table.

The four fold 2×2 table with variables A and B has been and probably is still the most frequently employed means of presenting statistical data. Consider the following table of counts relating to two classificatory variables A and B each with two categories. Here, a sample of n subjects is jointly classified by two categories, A and B. We assume here that only the sample size n is known in advance, the resulting table of counts in Table 8.2 would be said to have arisen from a multinomial sampling scheme. Other sampling schemes include the case when both the row marginal totals, n_{1+} and n_{2+}, as well as the column marginal totals, n_{+1} and n_{+2}, are known in advance. See Lawal (2003) for a more detailed discussion of the conceptual implications of the various sampling schemes that could give rise to a 2×2 contingency table.

Let the corresponding underlying probability distribution be as presented in Table 8.3 under the multinomial sampling scheme. Then, $\sum_i \sum_j p_{ij} = 1$ since only the sample size n is assumed fixed here.

Table 8.2 Observed frequency count in a 2×2 table

	B		
A	1	2	Total
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Table 8.3 Observed frequency count in a 2×2 table

	B		
A	1	2	Total
1	p_{11}	p_{12}	p_{1+}
2	p_{21}	p_{22}	p_{2+}
Total	p_{+1}	p_{+2}	1

Test of Independence

We are concerned here with the null hypothesis of no association between A and B, which we will be formally stated as:

$$H_0 : \text{A and B are independent}$$

$$H_a : \text{A and B are not independent}$$

This can formally be stated mathematically as

$$H_0 : p_{ij} = p_{i+}\, p_{+j} \tag{8.3}$$

$$H_a : p_{ij} \neq p_{i+}\, p_{+j} \tag{8.4}$$

for $i = 1, 2;\ j = 1, 2$. It can be shown that under the null hypothesis in (8.3), the estimate of p_{ij} is given by,

$$\hat{p}_{ij} = \frac{n_{i+}\, n_{+j}}{n^2}.$$

Consequently, the expected value, $E_{ij} = n\hat{p}_{ij}$ equals

$$E_{ij} = n\, \frac{n_{i+}\, n_{+j}}{n^2} = \frac{n_{i+}\, n_{+j}}{n} \tag{8.5}$$

that is,

$$E_{ij} = \frac{n_{i+}\, n_{+j}}{n} = \frac{\begin{pmatrix} \text{marginal} \\ \text{row total} \end{pmatrix} \begin{pmatrix} \text{marginal} \\ \text{column total} \end{pmatrix}}{\text{sample size}}.$$

We present the calculations of the expected frequencies in the following example.

Example 8.2.1

In an investigation into the frequency of side effects, say nausea, with a particular drug, 50 subjects may be given the drug, 50 subjects a placebo, and the number of subjects suffering from nausea assessed in each sample. The table below (Table 8.4) shows a possible outcome:

Table 8.4 The classification of 100 subjects in this study

Treatment	Side effect (Nausea)		Total
	Present	Absent	
Drug given	15	35	50
Placebo given	4	46	50
	19	81	100

The expected values are computed as follows:

$$E_{11} = \frac{n_{1+}\, n_{+1}}{n} = \frac{50 \times 19}{100} = 9.5$$

$$E_{11} = \frac{n_{1+}\, n_{+2}}{n} = \frac{50 \times 81}{100} = 40.5$$

$$E_{21} = \frac{n_{2+}\, n_{+1}}{n} = \frac{50 \times 19}{100} = 9.5$$

$$E_{22} = \frac{n_{1+}\, n_{+1}}{n} = \frac{50 \times 81}{100} = 40.5.$$

The test statistic is the Pearson's X^2 defined in this case as,

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{8.6}$$

and is distributed χ^2 with $(2-1) \times (2-1) = 1$ degree of freedom. For this example, we have therefore,

$$\begin{aligned} X^2 &= \frac{\sum (O_i - E_i)^2}{E_i} \\ &= \frac{(15-9.5)^2}{9.5} + \frac{(35-40.5)^2}{40.5} + \frac{(4-9.5)^2}{9.5} + \frac{(46-40.5)^2}{40.5} \\ &= 7.862 \end{aligned}$$

Now, $\chi^2_1 = 3.841$ at the 5 % level. Since the calculated $X^2 = 7.862 > 3.841$, we would reject the null hypothesis and we are led to support the truth of our hypothesis that occurrence of side effects is dependent on the treatments involved. The above calculations can be obtained in MINITAB with the following:

```
MTB > PRINT C1 C2

 Row      A      B
   1     15     35
   2      4     46

MTB > ChiSquare 'A' 'B'.

Chi-Square Test: A, B

Expected counts are printed below observed counts

              A        B    Total
    1        15       35       50
           9.50    40.50

    2         4       46       50
           9.50    40.50

Total        19       81      100

Chi-Sq =  3.184 +  0.747 +
          3.184 +  0.747 = 7.862
DF = 1,  P-Value = 0.005
```

Result from MINITAB indicates that the p value for this test is 0.005 which is less than 0.05. This again leads us to rejecting the null hypothesis that the two variables are independent.

An alternative calculation of Pearson's test statistic X^2 for the 2×2 table is given by:

$$X^2 = \frac{n(n_{11}\, n_{22} - n_{12}\, n_{21})^2}{n_{1+}\, n_{2+}\, n_{+1}\, n_{+2}}. \tag{8.7}$$

Hence, for our example, we have,

$$X^2 = \frac{100(15 \times 46 - 35 \times 4)^2}{50 \times 50 \times 19 \times 81} = 7.862.$$

Yates (1934) has suggested an improvement to Pearson's X^2 test statistic when expected values are small. His statistic is given by

$$T^2 = \frac{n(|n_{11}\, n_{22} - n_{12}\, n_{21}| - \frac{n}{2})^2}{n_{1+}\, n_{2+}\, n_{+1}\, n_{+2}} \tag{8.8}$$

and is distributed χ^2 with 1 d.f.

In the above example, we notice that the number of subjects subjected to the treatment is prefixed, i.e., the marginal totals (50,50) are already fixed. This experiment is an example of the well-known comparative trial. Other possible sampling schemes are:

(i) The case of selecting 100 subjects and then administering the treatments to them at random. In this case, only the total sample size is fixed—this is, the Double Dichotomy Scheme. Other names for this scheme are "naturalistic," "cross-sectional," or "multinomial."

(ii) Suppose before the experiment was conducted, the marginals—(50,50) and (19,81)—were already predetermined. In this case, only the configuration inside the table can change. This sampling scheme is popularly known as the 2×2 independent trial and it is by far the most well-known. The appropriate test statistic is the well-known Fisher's exact test.

8.3 Fisher's Exact Test

Sometimes for the test of association in a 2×2 table conducted earlier, the expected values $\hat{E}_{ij}$ may be small (less than five), in this case, the underlying assumption that X^2 will follow an χ^2 distribution may be violated. In such a circumstance, the Fisher's exact test provides an alternative. The test is most useful therefore for small samples. With Fisher's exact test, consider again the 2×2 table (Table 8.5). Here, n_{11} will be described as the pivot cell. The Fisher's exact test expect that we

Table 8.5 Observed frequency count in a 2×2 table

	B		
A	1	2	Total
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Let us consider a simple example. The data in Table 8.6 refer to 35 aggressive interactions of territorial pairs of convict cichlids (*Chiclasoma nigrofasciatum*), either member of the resident pair is likely to attack a conspecific intruder into the territory. It is suspected that the resident male is more likely to attack an intruding male, and that the resident female is more likely to attack an intruding female.

Table 8.6 Interaction results of 35 aggressive cichlids

	Intruder		
Resident	Male	Female	Total
Male	12	5	17
Female	3	15	18
Total	15	20	35

Fisher's exact test consists first of generating all tables that are consistent with the given marginal totals $M_1 = \{17, 18\}$ and $M_2 = \{15, 20\}$. Then for each such table, we calculate the sums of the individual probabilities associated with tables which are as extreme or more extreme than the observed table. An extreme table is one in which

$$P(n'_{11}) \leq P(n_{11}) \tag{8.9}$$

where n'_{11} is any other table satisfying the marginal totals M_1, M_2 above and having the pivot cell n'_{11}.

The probability of the observed table is given by the hypergeometric distribution, $\mathbf{n}' = \{n_{11}, n_{12}, n_{21}, n_{22}\}$, given by the probability model:

$$P[n_{11}] = \frac{n_{1+}!n_{2+}!n_{+1}!n_{+2}!}{N!n_{11}!n_{12}!n_{21}!n_{22}!}. \tag{8.10}$$

For our data, the probability of the observed table is therefore given by:

$$\mathrm{P}(12, 5, 3, 15) = \frac{17!18!15!20!}{35!12!5!3!15!} = 0.0016.$$

The collection of all possible tables consistent with the marginal totals is displayed below.

0	17
15	3

1	16
14	4

2	15
13	5

3	14
12	6

4	13
11	7

5	12
10	8

6	11
9	9

7	10
8	10

8	9
7	11

9	8
6	12

10	7
5	13

11	6
4	14

12	5
3	15

13	4
2	16

14	3
1	17

15	2
0	18

Here the range of the pivot cell is determined by $n_{11} = \{\max(0, n_{1+} - n_{+2})$ to $\min(n_{1+}, n_{+1})\}$. Thus in our case, n_{11} varies from 0 to 15.

The hypergeometric probabilities of each consistent table are displayed below:

n'_{11}	$P(n'_{11})$
0	0.0000
1	0.0000
2	0.0004
3	0.0039
4	0.0233
5	0.0834
6	0.1853
7	0.2620
8	0.2382
9	0.1389
10	0.0513
11	0.0117
12	0.0016
13	0.0001
14	0.0000
15	0.0000

Tables that are extreme or more extreme than the observed table are therefore those having pivot cells, $0, 1, 2, 12, 13, 14,$ and 15. Consequently, the Fisher's exact two-sided test is the sum of these probabilities. That is, the p value is given by: $0.0016 + 0.0001 + 0.0004 = 0.0021$. Thus, we would reject the null hypothesis that intruder's attack and resident's attack are not associated. The MINITAB for implementing Fisher's test is presented below:

```
MTB > XTABS 'R' 'C';
SUBC>   Layout 1 1;
SUBC>   Frequencies 'COUNT';
SUBC>   Counts;
SUBC>   DMissing 'R' 'C';
SUBC>   ChiSquare;
SUBC>   Expected;
SUBC>   Fisher.

Tabulated statistics: R, C

Using frequencies in COUNT


Rows: R   Columns: C

           1      2    All

1         12      5     17
        7.29   9.71  17.00

2          3     15     18
        7.71  10.29  18.00

All       15     20     35
       15.00  20.00  35.00

Cell Contents:      Count
                    Expected count


Fisher's exact test: P-Value =  0.0020456
```

8.4 Combining Several 2 × 2 Tables

In many studies, a number of 2×2 tables, all bearing on the same questions, may be available, and we may wish to combine these in some way to make an overall test of association between the row and column factors. For instance, in the survival of infants to amount of prenatal care received, data may be obtained from several clinics and for each clinic, the data might be arranged in a 2×2 table. In general, we are interested in combining the information from each of the 2×2 tables across the levels of the subpopulation (clinics). However, there is always a danger in collapsing the tables across the subpopulations because conclusions drawn from such a collapsed table may not reflect the truth within the subpopulation interactions.

This may be illustrated with the data in the example in Table 8.7 relating to the survival of infants according to the amount of prenatal care received by the mothers. The amount of care is classified as "more" or "less" and the mothers attended one of two clinics denoted by A and B.

Table 8.7 Amount of care received from two clinics. (Source: Bishop *et al.*)

Clinics	Amount of prenatal care	Status		Total
		Survived	Died	
A	More	176	3	179
	Less	293	4	297
	Total	469	7	476
B	More	197	17	214
	Less	23	2	25
	Total	220	19	239
Total		689	26	715

The X^2 values for the test of independence for each of the clinics are given by,

$$\text{Clinic A: } X^2 = \frac{476(176 \times 4 - 293 \times 3)^2}{179 \times 297 \times 46 \times 7} = 0.083$$

$$\text{Clinic B: } X^2 = \frac{239(197 \times 2 - 17 \times 23)^2}{214 \times 25 \times 220 \times 19} = 0.000.$$

Both are not significant and both therefore indicate that we would fail to reject H_0, that is, the survival of an infant is independent of the amount of prenatal care. However, if we combine the information in both clinics by collapsing across the two clinics, we shall have the following 2×2 in Table 8.8.

Table 8.8 The 2×2 from the collapsing of the clinics

Care	Survived	Died	Total
More	373	20	393
Less	316	6	322
Total	689	26	715

The computed test statistic from the combined Table (Table 8.8) is,

$$X^2 = \frac{715(373 \times 6 - 20 \times 316)^2}{393 \times 322 \times 689 \times 26} = 5.25$$

which surprisingly is now significant, indicating that the survival of infants is dependent on the amount of prenatal care after ignoring the clinic.

These results indicate that within each of the clinics, survival does not seem related to the amount of prenatal care. There are apparent significant variations between the survival rates for the two clinics. What this means is that while the within clinics analysis suggests that survival is not related to the amount of prenatal care, this conclusion we see is certainly negated when the tables were collapsed across the clinics. This apparent contradiction indicates that an overall test of the hypothesis relating the amount of prenatal care and infant survival must account for potential differences among the clinics. This contradiction is known as *Simpson's paradox*.

A general statistic that combines information from each of these 2×2 tables (indeed from many such tables) across the levels of clinics is the Mantel–Haenzel test statistic,

$$Q_{MH} = \frac{(n_{+11} - m_{+11})^2}{V_{+11}} \tag{8.11}$$

where n_{+11}, m_{+11}, and V_{+11} are the corresponding sums of observed frequency, expected frequency, and variance of the pivot cell across two subtables (clinics).

For clinic A: Expected value of the pivot cell = 176.3676
For clinic B: Expected value of the pivot cell = 196.9874

The variance of each of the pivot cells is given by

$$V_{i11} = \frac{n_{i1+}n_{i2+}n_{i+1}n_{i+2}}{N^2(N-1)}$$

Hence, we can compute the variances for the pivot cells in both clinics as follows:

$$\text{Clinic A: } \frac{179 \times 297 \times 469 \times 7}{476^2(475)} = 1.6217$$

$$\text{Clinic B: } \frac{214 \times 25 \times 220 \times 19}{239^2(238)} = 1.6450.$$

For the above data,

$$\begin{aligned} n_{+11} &= 176 + 197 &&= 373 \\ m_{+11} &= 176.74 + 196.98 &&= 373.3550 \\ V_{+11} &= 1.622 + 1.645 &&= 3.2667. \end{aligned}$$

Hence, substituting this in the expression in 8.11, we have,

$$Q_{MH} = \frac{(373 - 373.355)^2}{3.2667} = 0.0386 \quad \text{on 1 d.f.}$$

Clearly, the Q_{MH} computed above is not significant. This further shows that for our data, survival does not depend on the amount of prenatal care after adjusting for the effects of the clinics.

The MINITAB implementation of the above is presented below. Note that the clinic is the layer variable.

```
Data Display

Row  Clinic  Care  Status  count
  1       1     1       1    176
  2       1     1       2      3
  3       1     2       1    293
  4       1     2       2      4
  5       2     1       1    197
  6       2     1       2     17
  7       2     2       1     23
  8       2     2       2      2


MTB > XTABS 'Care' 'Status' 'Clinic';
SUBC>   Layout 1 1;
SUBC>   Frequencies 'count';
SUBC>   Counts;
SUBC>   DMissing 'Care' 'Status' 'Clinic';
SUBC>   ChiSquare;
SUBC>   MHCTest.

Tabulated statistics: Care, Status, Clinic

Using frequencies in count


Results for Clinic = 1

Rows: Care   Columns: Status

         1  2  All

1      176  3  179
2      293  4  297
All    469  7  476

Cell Contents:      Count


Pearson Chi-Square = 0.084, DF = 1, P-Value = 0.773
Likelihood Ratio Chi-Square = 0.082, DF = 1, P-Value = 0.774

* NOTE * 2 cells with expected counts less than 5


Results for Clinic = 2

Rows: Care   Columns: Status

          1   2  All

1       197  17  214
2        23   2   25
All     220  19  239

Cell Contents:      Count
```

```
Pearson Chi-Square = 0.000, DF = 1, P-Value = 0.992
Likelihood Ratio Chi-Square = 0.000, DF = 1, P-Value = 0.992

* NOTE * 1 cells with expected counts less than 5


Results for all 2x2 tables

Common odds ratio  0.898038

MHCstatistic  DF   P-Value
    0.0064278   1  0.936099
```

The MINITAB computed $Q_{MH} = 0.0064$. This value is not correct as the true value should have been 0.0386. Of course, both results lead to the same conclusion that survival does not depend on the amount of prenatal care after adjusting for the effects of the clinics. The estimated common odds ratio is $\hat{\theta} = 0.8980$. This common ratio is based on the assumption that the strength of association (or direction of association) is the same across the clinics. The Breslow–Day test for the homogeneity of odds ratios gives $X^2 = 0.0440$ on 1 d.f. corresponding to a p value of 0.8338. Clearly, there is very strong evidence in this case that the odds-ratios are homogeneous across the two tables. However, if this were not the case, then we would have believed that there is an *interaction* or *effect modification* between clinic and survival. The clinic factor is often referred to as *effect modifier*.

8.5 The General $R \times C$ Contingency Table

Suppose a sample of n objects is jointly classified according to two different and independent classifications A and B with r and c classes respectively. Let n_{ij} be the observed frequency in cell (i, j) with $i = 1, 2, \cdots, r$ and $j = 1, 2, \cdots, c$. The observation can be displayed as in Table 8.9.

Two commonly experimental settings are often encountered in contingency tables analysis. These are:

1. Only the sample size n is fixed and the marginal totals are allowed to vary at random subject to the constraints $\sum_i n_{1+} = n$ and $\sum_j n_{+1} = n$. This setting leads to the usual test of independence.

Table 8.9 Observed $a \times c$ contingency table

	B				
A	1	2	$\cdots$	c	Total
1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+c}	n

2. The second setting assumes that one of the marginal totals is fixed. That is, either the row marginals $n_{i+}, i = 1, 2, \cdots, r$ or the column marginals $n_{+j}, j = 1, 2, \cdots, c$ are fixed. This setting leads to the *prospective* and *retrospective* studies respectively.

We are concerned here with the test of homogeneity. Asymptotically (that is, when n is large), the two tests lead to the same results. In the independence model, our hypotheses are of the form

$$H_0 : \text{A and B are independent}$$
$$H_a : \text{A and B are not independent}$$

which again mathematically can be formulated for $i = 1, 2, \cdots, r; j = 1, 2, \cdots, c$ as

$$H_0 : \ p_{ij} = p_{i+}\, p_{+j} \tag{8.12}$$
$$H_a : \ p_{ij} \neq p_{i+}\, p_{+j}. \tag{8.13}$$

In the homogeneity model, if we assume for instance that the row marginals n_{i+} are fixed, then we are interested in the hypotheses,

$$H_0 : \text{the proportional split among categories of variable A}$$
$$\text{is the same across each level of variable B}$$
$$H_a : \text{the proportional split among categories of variable A}$$
$$\text{is not the same across each level of variable B}$$

which mathematically can be formulated for $i = 1, 2, \cdots, r; j = 1, 2, \cdots, c$ as,

$$H_0 : = \begin{cases} p_{11} = p_{21} = \cdots = p_{r1} & \text{proportion in column 1 of B is the same} \\ p_{12} = p_{22} = \cdots = p_{r2} & \text{proportion in column 2 of B is the same} \\ \vdots \quad = \vdots \quad = \cdots = \vdots & \qquad\qquad \vdots \\ p_{1c} = p_{2c} = \cdots = p_{rc} & \text{proportion in column c of B is the same} \end{cases}$$

$$H_a : = p_{11} \neq p_{21} \qquad \text{at least any two of the } p\text{'s are not equal.}$$

Under both models, the expected frequencies are again given by,

$$E_{ij} = \frac{n_{i+}\, n_{+j}}{n} \tag{8.14}$$

That is, for a general $r \times c$ two-way contingency table, the expected value for a particular row and column is given by

$$\frac{\text{(Particular column total)} \times \text{(Particular row total)}}{\text{Grand total}}.$$

And again, the test statistic is the Pearson's X^2 defined in this case as,

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - E_{ij})^2}{E_{ij}},\ i = 1, 2, \cdots, 5; j = 1, 2, \cdots, c \tag{8.15}$$

and is distributed χ^2 with $(r-1) \times (c-1)$ degrees of freedom.

Example 8.5.1

Records of the number of lambs born to the ewes of various flocks are presented in Table 8.10. The records are given for two breeds of ewe on each of two farms. Do the proportions of ewes giving birth to 0, 1, 2, 3 or more lambs vary significantly for different farms and different breeds?

Table 8.10 Number of lambs for different breeds at two farms

		Number of lambs per birth				
Farm	Breed	0	1	2	3+	Total
1	A	10	21	96	23	150
	B	4	6	28	8	46
2	A	22	95	103	4	224
	B	18	49	62	0	129

First consider all four flocks simultaneously in a 4×4 contingency table (Table 8.11).

Table 8.11 Number of lambs cross-classified by farm/breed combinations (where the figures in brackets are the expected values)

	Number of lambs				
Farm/breed	0	1	2	3+	Total
1A	10(14.8)	21(46.7)	96(79.0)	23(9.61)	150
1B	4(4.5)	6(14.3)	28(24.21)	8(2.9)	46
2A	22(22.0)	95(69.8)	103(117.9)	4(14.3)	224
2B	18(12.7)	49(40.2)	62(67.9)	0(8.2)	129
Total	54	171	289	35	549

The expected values of 14.8 and 69.8 for instance are computed respectively from Table 8.11 as:

$$14.8 = \frac{150 \times 54}{549} \quad \text{and} \quad 69.8 = \frac{224 \times 171}{549}.$$

Overall $X^2 = 83.8$ on $(4-1)(4-1) = 9$ d.f. which is clearly very significant.

Because of the "factorial" nature of the four flocks, there are various ways in which we continue the analysis. One method is to look at each "main effect." Consider first the two-way table formed from farms and number of lambs leading to a 2×4 contingency table (Table 8.12).

For the data in Table 8.12, the computed $X^2 = 80.9$ on 3 d.f., so that there are clear overall differences between farms as they relate to the number of lambs born to the ewes.

Next, we create the two-way table, this time formed from breeds and number of lambs leading to a 2×4 contingency table (Table 8.13).

Table 8.12 The two-way table of farms and number of lambs

	Number of lambs				
Farm	0	1	2	3+	Total
1	14(19.3)	27(61.0)	124(103.2)	31(12.5)	196
2	40(34.7)	144(110.0)	165(185.8)	4(22.5)	353
Total	54	171	289	35	549

Table 8.13 The two-way table of breeds and number of lambs

	Number of lambs				
Breed	0	1	2	3+	Total
A	32(36.8)	117(116.5)	199.(196.9)	27(23.9)	374
B	22(17.2)	55(54.5)	90(92.1)	8(11.1)	175
Total	54	171	289	35	549

Again, for the data in Table 8.13, the computed $X^2 = 3.34$ on 3 d.f., which is not significant. Hence, the major difference between flocks is that farm 1 produces greater proportions of twin and triplet births than farm 2. Overall differences between breeds appear negligible but to check this, we examine differences between breeds for farms 1 and 2 respectively in Tables 8.14 and 8.15.

Table 8.14 Two-way table of breeds and number of lambs at farm 1

	Number of lambs (*Farm 1*)				
	0	1	2	3+	Total
Breed A	10(10.7)	21(20.7)	96(94.9)	23(23.7)	150
Breed B	4(3.3)	6(6.3)	28(29.1)	8(7.3)	46
Total	14	27	124	31	196

The computed statistic for the 2×4 contingency table (Table 8.14) is $X^2 = 0.3$ on 3 d.f. This value is clearly not significant at $\alpha = .05$ level.

Table 8.15 Two-way table of breeds and number of lambs at farm 2

Breed	Number of lambs (*Farm 2*) 0	1	2	3+	Total
Breed A	22(25.4)	95(91.4)	103(104.7)	4(2.5)	224
Breed B	18(14.6)	49(52.6)	62(60.3)	0(1.5)	129
Total	40	144	165	4	353

Similarly, the test statistic obtained for the data in Table 8.15 is $X^2 = 4.12$ on 3 d.f., again indicating that there is no significance at the 5 % level of significance.

Conclusion

There is clearly a considerable difference between the two farms in the proportions of 0, 1, 2, and 3 lambs per birth but no important differences between breeds.

Example 8.5.2

In a study of lung cancer, researchers were interested in the relationship between income (economic level) and smoking exposure. Two hundred and seventy-two males (age ≥ 55) were classified by current smoking patterns and five economic levels. The data are presented in Table 8.16.

Table 8.16 Cross-classification of smoking habits with economic level. (Source: brown et al. 1975)

Smoking	Economic level 1 (low)	2	3	4	5 (high)	Total
Never smoked	14	37	15	22	11	99
Past smoker	7	22	12	7	3	51
$\leq$1 pack/day	3	25	6	9	3	46
>1 pack/day	13	26	18	15	4	76
Total	37	110	51	53	21	272

We wish to test the hypothesis that smoking habits are independent of economic level at ages 55 and over. For these data, we employ the MINITAB to do the analysis. First we read the data into three columns named smoking, level, and count respectively (see the data displayed below). Next we use the MINITAB command to analyze the data again; see the commands used in the output below.

```
MTB > print c1-c3

Data Display

 Row  Smoking  Level  count

   1        1      1     14
   2        1      2     37
   3        1      3     15
   4        1      4     22
   5        1      5     11
   6        2      1      7
   7        2      2     22
   8        2      3     12
   9        2      4      7
  10        2      5      3
  11        3      1      3
  12        3      2     25
  13        3      3      6
  14        3      4      9
  15        3      5      3
  16        4      1     13
  17        4      2     26
  18        4      3     18
  19        4      4     15
  20        4      5      4

MTB > Table 'Smoking' 'Level';
SUBC>   Frequencies 'count';
SUBC>   ChiSquare 3;
SUBC>   Layout 1 1.

Tabulated Statistics: Smoking, Level
 Rows: Smoking     Columns: Level

          1       2       3       4       5       All

1        14      37      15      22      11        99
       13.47   40.04   18.56   19.29    7.64     99.00
        0.15   -0.48   -0.83    0.62    1.21       --

2         7      22      12       7       3        51
        6.94   20.63    9.56    9.94    3.94     51.00
        0.02    0.30    0.79   -0.93   -0.47       --

3         3      25       6       9       3        46
        6.26   18.60    8.63    8.96    3.55     46.00
       -1.30    1.48   -0.89    0.01   -0.29       --

4        13      26      18      15       4        76
       10.34   30.74   14.25   14.81    5.87     76.00
        0.83   -0.85    0.99    0.05   -0.77       --
```

```
All      37      110       51       53       21      272
      37.00   110.00    51.00    53.00    21.00   272.00
        --      --        --       --       --      --

Chi-Square = 12.374, DF = 12, P-Value = 0.416
2 cells with expected counts less than 5.0

  Cell Contents --
                  Count
                  Exp Freq
                  St Resid
```

We notice that MINITAB warns us that there are two cells with expected counts $E_{ij} < 5$. These are cells (2,5) and (3,5) with 3.94 and 3.55 expected values respectively. Previous studies have advocated that the test of independence is only valid if none of the expected cell frequencies is less than five. However recent studies by Yarnold (1970), Lawal and Upton (1984, 1989) have advocated that expected cell frequencies can indeed be lower than five without violating the validity of the approximation. Lawal (1980) has advocated that minimum expected frequencies can be tolerated for expected values satisfying:

$$F_{ij} \geq sd^{-3/2} \tag{8.16}$$

where F_{ij} is the smallest expected value, s is the number of cells having expected values less than three and $d = (r-1)(c-1)$. In the output above, none of the expected values is less than 3, therefore, the χ^2 would be suitable for this analysis.

The results indicate that $X^2 = 12.374$ on $(4-1)(5-1) = 12$ d.f. The corresponding p value is 0.416 which indicates that we would fail to reject H_0, that is, there is no strong evidence to suggest that smoking habit is associated with economic level in this example. In the above output, the first row gives the observed counts, the second row gives the expected counts, while the third row gives the standardized residuals

$$z_{ij} = \frac{(n_{ij} - F_{ij})}{\sqrt{F_{ij}}}$$

Thus, for z_{11} and z_{45} for example, these are computed as:

$$z_{11} = \frac{(14 - 13.47)}{\sqrt{13.47}} = 0.15$$

$$z_{45} = \frac{(4 - 5.87)}{\sqrt{5.87}} = -0.77.$$

We may notice here that $\sum_i \sum_j z_{ij}^2 = 0.15^2 + \cdots + (-0.77)^2 = X^2 = 12.374$.

The z_{ij} help us to detect cells that would otherwise make our independence model untenable. Values of $|z_{ij}| > 2$ are considered to be aberrant cells (see, Lawal 2003).

8.6 Fitting the Poisson Distribution

The Poisson density function is given by

$$f(x) = \frac{e^{\mu}\mu^{x}}{x!}, \quad x = 0, 1, 2, \cdots \tag{8.17}$$

and we showed in Chap. 4 that

$$f(x+1) = \frac{\mu}{X+1} f(x).$$

We shall now consider again the data in Chap. 4 on noxious weeds which we assume can be fitted by a Poisson distribution. The calculations in the chapter yield the following summary:

No. of noxious weeds x	Frequency	Expected frequency	$\frac{(O-E)^2}{E}$	Yarnold's	Lawal's
0	3	4.781	0.663		
1	17	14.440	0.454		
2	26	21.807	0.806		
3	16	21.955	1.615		
4	18	16.578	0.122		
5	9	10.015	0.103		
6	3	5.042	0.827		
7	5	2.176			
8	0	0.870			
9	1	0.274	2.01	2.01	1.204
10	0	0.083			
11 or more	0	0.030			
Total				6.60	8.29

The expected frequency for the Poisson distribution of counts of 11 or more is obtained by subtracting the sum of the expected frequencies for counts up to ten from the total number of noxious weeds. Since some of the expected values are less than five, we can employ Yarnold's rule. With Yarnold's rule, the minimum tolerable expected frequency is given by $\frac{5r}{k}$ where r is the number of cells with expectation less than five and k is the total number of cells. Here $k = 12$, $r = 6$, hence minimum frequency $= \frac{5\times 6}{12} = 2.5$. We would therefore need to combine the last five cells, we now have an expected value of 3.433. And by Yarnold's rule $k = 8$, $r = 2$ and hence, the minimum expectation $= \frac{5\times 2}{8} = 1.25$, and the new Pearson's $X^2 = 6.60$ on $8-2 = 6$ d.f. which when compared with $\chi^2_6(0.05) = 12.59$ is not significant, indicating that the model fits the data well.

Lawal (1980) has also advocated an improvement to Yarnold's rule above. Lawal suggested that the minimum expected frequency is given by $3r/k$ where

r is the number of expectations less then three. So all that we need collapse by this rule are the last four cells and it will give an $X^2 = 8.29$ with 9−2=7 d.f., and corresponding p value of 0.308 which is still not significant and of course this test is more powerful than the previous test because it is based on more degrees of freedom. When we employed MINITAB to fit the Poisson model to this data, we have the following partial output.

```
Results for: poisson.MTW

MTB > PGoodness 'x';
SUBC>   Frequencies 'f';
SUBC>   RTable.

Goodness-of-Fit Test for Poisson Distribution

Data column: x
Frequency column: f

Poisson mean for x = 3.02041

                     Poisson              Contribution
x     Observed  Probability    Expected     to Chi-Sq
0            3     0.048781      4.7806       0.66319
1           17     0.147339     14.4393       0.45413
2           26     0.222513     21.8062       0.80654
3           16     0.224026     21.9546       1.61502
4           18     0.169163     16.5779       0.12198
5            9     0.102188     10.0144       0.10276
6            3     0.051442      5.0413       0.82654
>=7          6     0.034548      3.3857       2.01869


 N  N*  DF   Chi-Sq  P-Value
98   0   6  6.60885    0.359


2 cell(s) (25.00%) with expected value(s) less than 5.
```

We see that MINITAB is employing Yarnol's rule and the computed $X^2 = 6.60885$ with corresponding p value of 0.359 indicating that the model fits the data.

8.7 Fitting the Binomial Distribution

We recall from Chap. 4 that the binomial distribution is given by

$$f(x) = \binom{n}{x} p^x \, q^{n-x} \tag{8.18}$$

where p is the probability of success in n independent trials and $q = 1 - p$.

It can be shown that the recurrence relation for this distribution is

$$f(x+1) = \frac{n-x}{X+1}\ \frac{p}{1-p}\ f(x). \tag{8.19}$$

8.7.1 Example

The number of males was recorded for each of 160 litters of five pigs with the following results:

Number of males	0	1	2	3	4	5
Frequency of litters	10	15	40	55	30	10

Assuming that any particular birth in a particular litter is equally likely to be male or female, test the hypothesis that the above data can be fitted by a binomial distribution.

We recognize that a birth can either result in a male or female, so we have a binomial case in our hand. Since each birth is equally likely, i.e. $p = \frac{1}{2} = q$. That is, $H_0 : p = \frac{1}{2}$ vs $H_1 : p \neq \frac{1}{2}$. Hence

$$\begin{aligned} f(x+1) &= \left(\frac{n-x}{x+1}\right).\left(\frac{p}{1-p}\right) f(x) \\ &= \frac{n-x}{x+1} f(x) \quad \text{since } p = q = \frac{1}{2} \end{aligned}$$

Further $n = 5$, thus

$$\begin{aligned} f(x+1) &= \frac{5-x}{x+1} f(x) \\ f(0) &= (1-p)^5 && = 1/32 \\ f(1) &= 5/1 \times 1/32 && = 5/32, \\ f(2) &= 4/2\ f(1) && = 10/32 \\ f(3) &= 3/3\ f(2) && = 10/32, \\ f(4) &= \frac{5-3}{4} f(3) && = 5/32 \\ f(5) &= 1/32. \end{aligned}$$

Note the symmetry (e.g., $f(1) = f(4)$).

No. of males	Observed frequency	Expected frequency (160 × probability)	$\frac{(O-E)^2}{E}$
0	10	5	5.0
1	15	25	4.0
2	40	50	2.0
3	55	50	0.5
4	30	25	1.00
5	10	5	5.0
	160		$X^2 = 17.5$

The 5 % significance point on the χ^2 distribution on $6-1=5$ d.f. is 11.07 and corresponding p value of 0.0036 which clearly indicates that we would have to reject the null hypothesis that $p = \frac{1}{2}$. So, we reject the hypothesis that the birth in a particular litter is equally likely i.e., accept that $p \neq \frac{1}{2}$.

Estimating p from the Sample Data

Further calculations to compare the observed frequencies with frequencies calculated on the assumption that the probability of a pig being male is p, where p is to be estimated from the data, are summarized below.

$$\begin{aligned}\hat{p} &= \frac{\text{Total number of males}}{\text{Total number of births}} \\ &= \frac{(15\times 1)+(40\times 2)+(55\times 3)+(30\times 4)+(10\times 5)}{160\times 5} \\ &= 0.5375.\end{aligned}$$

Both the binomial probabilities and expected frequencies are presented in the table below, together with the computed X^2 test statistic.

No. of males	Probability	Expected Frequency	$(O-E)^2/E$
0	$f(0) = (0.4625)^5 = 0.0212$	3.39	12.89
1	$f(1) = 5(1.1622)(0.0212) = 0.1230$	19.68	1.113
2	$f(2) = \frac{4}{2}(1.1622)0.1230 = 0.2859$	45.744	0.721
3	$f(3) = \frac{3}{3}(1.1622)0.2859 = 0.3323$	53.108	0.067
4	$f(4) = \frac{1}{2}(1.1622)0.3323 = 0.1931$	30.896	0.026
5	$f(5) = \frac{1}{5}(1.1622)0.1931 = 0.0449$	7.18	1.108

$X^2 = 15.925$ with $6-2=4$ d.f. which is highly significant. Thus, the data do not fit the binomial model. The calculations are implemented in MINITAB as follows:

```
MTB > set c1
DATA> 0 1 2 3 4 5
DATA> end
MTB > PDF 'x' c2;
SUBC>   Binomial 5 0.5375.
MTB > let c3=c2*160
MTB > let c5=((c4-c3)**2)/c3

MTB > print c1-c5

Data Display

Row  x         p         E   O        X2
  1  0  0.021162   3.3859  10  12.9199
  2  1  0.122969  19.6750  15   1.1108
  3  2  0.285820  45.7312  40   0.7182
  4  3  0.332169  53.1470  55   0.0646
  5  4  0.193017  30.8827  30   0.0252
  6  5  0.044863   7.1781  10   1.1093

MTB > Sum 'X2'.

Sum of X2

Sum of X2 = 15.9481
```

Again, the computed $X^2 = 15.9481$ and is based on 4 d.f., with corresponding p value of 0.0031which clearly indicates that we would strongly reject H_o.

8.8 Exercises

1. The data below are obtained by Catchside during his analysis of the secondary association of chromosomes in *Brassila oleracea*. The pollen mother cells were classified according to whether they had 3, 2, 1, or 0 pairs of bivalents showing secondary association at metaphase. Three preparations were studied to test the hypothesis that the classification of the pollen mother cells could be considered as constant from slide to slide.

Number of pairs	Slide		
	1	2	3
0	14	7	11
1	32	36	35
2	51	39	32
3	41	23	16

2. A psychologist tests coordination of hand and eye in 475 subjects. He finds that 30 can perform a certain task with the right hand but not

the left, 13 with the left hand but not the right. What light does this throw on whether success rates for the two hands differ in the population sampled?.

3. From a large sowing of seed, 480 plants are raised, and are classified for flower color and leaf type:

Leaf	Dark blue	Light blue	Yellow	Pink	White
Rough	41	105	36	39	151
Smooth	3	15	18	28	44

Test the hypothesis that color and leaf type are independent.

4. A therapeutic drug was tested against a placebo in terms of three subjectively evaluated patient categories: (1) much improved, (2) slightly improved, and (3) not improved. A total of 120 patients were assigned to the drug group and 90 other patients were given the placebo. All were judged to be in approximately the same initial condition. Physician evaluation was then made without knowing which treatment the patient received. The resulting data were organized in the following 2×3 contingency table.

	Patient categories			
Treatment	Much improved	Slightly improved	Not improved	Total
Drug	60	32	28	
Placebo	28	17	45	
Total				

What are the factors of interest in this study? Analyze the data and on the basis of your result, is the drug effective? Use $\alpha = 0.05$.

5. The data below relate to the study of the risk of cancer and cigarette smoking. Smoking is measured in pack-years (one pack-year is the equivalent of smoking a pack of cigarettes per day for 1 year). A light smoker is defined as a smoker with less than 31 pack-years of smoking, medium smoker, between 31 and 45 pack-years (inclusively); heavy smoker, more than 45 pack-years. Use MINITAB to answer the following questions:

	Smoking history				
Cancer Site	Never	Light	Medium	Heavy	Total
Lung	12	33	44	92	
Oral-bladder	37	37	42	36	
Other cancer	270	138	182	136	
No cancer	2025	996	880	738	
Total					

(a) What are the factors of interest in this study?
(b) What is the observed frequency of oral-bladder cancers among medium smokers?
(c) What is the contribution toward the overall X^2 value from individuals that are jointly classified with lung cancer and medium smoking history? How was it computed?
(d) What is the overall X^2 value?
(e) State the null and alternative hypotheses in this study and conduct the test (use $\alpha = 0.05$).
(f) How was the 9 d.f. obtained?
(g) On the basis of your result, is smoking bad for you?

6. Each of 126 individuals of a certain mammal species was placed in an enclosure containing equal amounts of each of six different foods. The frequency with which the animals chose each of the foods was:

Food item (i)	1	2	3	4	5	6
Frequency	13	26	31	14	28	14

(a) Test the hypothesis that there is no preference among the food items. Use $\alpha = 0.05$
(b) If the null hypothesis is rejected, ascertain which of the foods were preferred by the animals. Formulate the null and alternative hypotheses.

7. Over a specified period, observers sighted 300 birds at a particular location. The birds are classified into four species categories:

Species	1	2	3	4	Total
# of birds	60	120	97	23	300

Test the hypothesis that the composition of the species in the location has changed from the expected proportion of 3:3:3:1?

8. The following blood-type frequencies were obtained from a sample of 1000 subjects screened at a shopping mall over a period of one month.

Blood type	O	A	B	AB	Total
Frequency	465	394	96	45	1000

(a) Do the above data support the claim that less than 5 % of the population screened has blood type AB?
(b) Perform a test to test the hypothesis that the distribution of these blood types in the population should be in the ratio 9:8:2:1.

9. The Mendelian theory states that probability distribution of the color and shape of a variety of pea be in the ratio 9:3:3:1. A random sample of 200 peas has the following observed distribution:

Color and shape	Round and yellow	Round and green	Angular and yellow	Angular and green	Total
Frequency	110	40	42	8	200

Perform a hypothesis test to determine whether the observed data contradict the Mendelian theory.

10. The following are the numbers of a particular organism found in 100 samples of water from a pond.

Number of organisms per sample	Frequency
0	15
1	30
2	25
3	20
4	5
5	4
6	1
7	0
Total	100

Test the hypothesis that the above data follow a Poisson distribution. What is the p value for the test?

11. The number of tomato plants attacked by spotted wilt disease was counted in each of 160 areas of nine plants. The results are displayed below (Snedecor and Cochran 1973).

No of diseased plants	0	1	2	3	4	5	6	7	Total
Frequency	36	48	38	23	10	3	1	1	160

Fit a binomial distribution to these data and perform the relevant goodness-of-fit test.

12. The table below contains results of a study by Mendenhall et al. (1984) to compare *radiation* therapy with *surgery* in treating cancer of the larynx.

Treatment	Cancer controlled	Cancer not controlled
Surgery	21	2
Radiation	15	3

The distribution of the pivot cell n_{11} is also given by:

n_{11}	Probability
18	0.0449
19	0.2127
20	0.3616
21	0.2755
22	0.0939
23	0.0144

(a) Use Fisher's exact test to test the following hypotheses:

$$H_0 : \theta = 1 \quad \text{against}$$
$$H_a : \theta \neq 1$$

where θ is the odds ratio. Explain how you formed the p value and draw you conclusions based on your results.

13. If it is believed that treatment A is better than treatment B, list all possible outcomes that are extreme or more extreme than the observed table in the following fictitious table of data.

	Treatment		
Outcome	A	B	Total
Die	5	3	8
Live	9	15	24
Total	14	18	32

Conduct Fisher's exact test on these data.

14. The following are data from two studies that investigated the risk factors for epithelial ovarian cancer (Pagano and Gauvreau 1993).

Study I

Disease status	Term pregnancies		Total
	None	One or more	
Cancer	31	80	111
No cancer	93	379	472
Total	124	459	583

Study II

Disease status	Term pregnancies		Total
	None	One or more	
Cancer	39	149	188
No cancer	74	465	539
Total	113	614	727

(a) Estimate the odds ratio of developing ovarian cancer for women who have never had a term pregnancy versus women who have had one or more in the first study.
(b) If possible, you would like to combine the evidence in these two strata to make an overall statement about the relationship between ovarian cancer and term pregnancies. What would happen if you were to simply sum the entries in the tables?
(c) Conduct a test of homogeneity. Is it appropriate to use the Cochran–Mantel–Haenszel method to combine the information in these two tables?
(d) Obtain the Cochran–Mantel–Haenszel estimate of the common odds ratio.
(e) Test the null hypothesis that there is no significant association between ovarian cancer and term pregnancies at the 0.01 level of significance.

15. Geissler (1889) in a genetic study examined hospital records in Saxony and compiled data on gender ratio. The following table gives the number of male children in 6115 families with 12 children. If we assume that the genders in successive children follow a binomial distribution with constant probability p, estimate p from the data and test whether the data agree with a binomial model.

Number	Frequency
0	7
1	45
2	181
3	478
4	829
5	1112
6	1343
7	1033
8	670
9	286
10	104
11	24
12	3

16. Student (1907) conducted a study on errors made in counting yeast cells or blood corpuscles with a hemocytometer. In his study, yeast cells were killed and mixed with water and gelatin; the mixture was then spread on a glass and allowed to cool. Four different concentrations were used and counts were made on 400 squares, and the data are presented below:

Number of Cells	Concentration 1	Concentration 2	Concentration 3	Concentration 4
0	213	103	75	10
1	128	143	103	20
2	37	98	121	43
3	18	42	54	53
4	3	8	30	86
5	1	4	13	70
6	0	2	2	54
7	0	0	1	37
8	0	0	0	18
9	0	0	1	10
10	0	0	0	5
11	0	0	0	2
12	0	0	0	2

(a) Fit Poisson distributions to each set of data.
(b) Test the goodness-of-fit to these data using Pearson's X^2.
(c) What are the parameter estimates for each data set?
(d) Which data fit the Poisson model best?

17. Hoaglin (1980) suggested a "Poissoness plot" which is a simple visual method for assessing goodness of fit. For a Poisson model, the expected frequencies are given by:

$$E_x = n \Pr(X = x) = ne^{-\lambda} \frac{\lambda^x}{x!}, \text{or} \log E_x = \log n - \lambda + x \log \lambda - \log x!$$

Hence a plot of $\log(O_x) + \log x!$ versus x should yield nearly a straight line with a slope of $\log \lambda$ and an intercept of $\log n - \lambda$. Construct such plots for the data in the last exercise and comment on your results. Reconcile your result with the results in the last exercise.

Chapter 9
Experimental Design

9.1 Introduction

Before we formally describe the principles of experimental design, two terms need to be defined. The two terms are *Observational Study* and *Experiment*.

Observational study investigates what is present in the population. Any condition not represented in the population will not be observed in an observational study (we shall discuss this at the end of this chapter). In many investigations, however, it is desired to investigate conditions which do not appear in a population. In an *experimental investigation* or *experiment*, the experimenter may, and often does introduce conditions which do not exist in any naturally occurring population, i.e., it is a planned interference in the naturally occurring order of events by the investigator. The investigator controls the conditions in the experiment, whereas the conditions in a survey are those that prevail in the population.

An experiment is defined as the planning and collection of measurements or observations according to a prearranged plan, for the purpose of obtaining factual evidence for or against a stated theory or hypothesis. An experiment should be self-contained, i.e., it should provide an independent piece of information about a stated theory or hypothesis, and the conclusions should be based on the experimental data alone. Extraneous knowledge may help the investigator to understand the observed results, but it should not be injected into the statistical inferences. This certainly does not mean that one cannot use accumulated or prior knowledge in planning further experiments. However, experimentation is only one step in the continuous search for knowledge.

Some additional terms are defined below:

Treatment A treatment is a single entity or phenomenon under study in an experiment.

Absolute Experiment An absolute or single phenomenon experiment is one which contains a single treatment.

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_9,

Comparative Experiment A comparative experiment is one designed specifically to compare two or more treatments, which might be different varieties of corn (maize), different diets for cows, different drugs, or different acid concentrations in an industrial process.

Many comparative experiments are factorial, permitting simultaneous study of different *factors* each at different *levels*. Thus in an experiment to study the control of stem borers in corn or maize, three sowing dates (Factor A with three levels) might be tried in conjunction with four sprays (Factor B, four levels). In this "3 × 4 factorial experiment," the "twelve combinations of sowing dates and sprays are the twelve 'treatment combinations'". The treatment combinations are displayed in the following table.

Levels of factor A	Levels of factor B			
	1	2	3	4
1	11	12	13	14
2	21	22	23	24
3	31	32	33	34

A factor may be qualitative (e.g., different concentrations of nitrogen in a fertilizer experiment) or quantitative (e.g., different doses of vitamin C).

Each experiment comprises a set of "experimental units" called plots with different treatments or treatment combinations applied to different experimental units. Plots may be for example, pens of animals, individual animals or plants, individual leaves on growing plants, or (in an inoculation experiment on plants) half-leaves. Commonly, treatments are replicated, i.e., there is more than one plot per treatment. Factorial experiments include *single replicate experiments* (one plot per treatment combination; see Chap. 14) and *fractional replicate* (not all treatment combinations present; see Chap. 14).

The comparison between treatments, between factor levels, or treatment combinations are based on data recorded after the treatments have had a chance to affect the experimental material (the response). These data are one or more variates (often called variables) which as we discussed in Chap. 1 may be continuous (e.g., weights, heights, temperatures, and pressures) or discrete (e.g., counts, scores).

Variates may also be "primary" (obtained by direct reading of a scale by counting or by scoring) or "derived" (obtained from one or more other variates by arithmetic calculations).

For factorial experiments with quantitative factors, the analysis of a variate may consist of investigating a "response surface," and for our purpose in this text, experiments designed to study response surfaces will be regarded as comparative.

Experimental Unit It is the smallest unit in which a treatment is applied. It could be an individual animal or a group. For example, an individual mouse is considered as the experimental unit when a drug therapy or surgical

procedure is being tested, but an entire litter of mice is the experimental unit when an environmental teratogen is being tested. For the purpose of estimating error of variance or standard error for statistical analysis, it is necessary to consider the experimental unit (Weber and Skillings 2000). Many excellent sources provide discussions on the type of experimental units and their appropriateness (Dean and Voss 1999; Festing and Altman 2002; Keppel 1991; Wu and Hamada 2000).

Treatment Design A treatment design represents the arrangement and selection of treatments for comparative purposes or for ascertaining responses to several treatment variables and levels of variables.

9.2 Experimental Design

An experimental design is the arrangement of treatments in an experiment. It is used in all types of empirical investigations. Designing of an experiment is more difficult to define than its analysis, because it means both designing (part of planning) and a "specific design." Finney (1955) defined the "design" of an experiment to mean:

(i) The set of treatments selected for comparisons
(ii) The specification of the experimental unit (animals, field plots) to which the treatments are to be applied
(iii) The rules for allocating the treatments to the plots or units
(iv) The specification of the measurements or other records to be made on each unit
(v) Specification of details of the management of the experiment

The need for a statistical subject of "design and analysis of experiments" arises from variability inherent in the experimental material, environment, and management. This variability leads to uncontrolled, indeed uncontrollable variability in the observed variate values, i.e., to "experimental error" where "error" does *not* mean "mistake."

The four major components of design and analysis according to Preece (1982) are as follows:

(a) Planning, design, and layout
(b) Management
(c) Data recording
(d) Scrutiny and editing of data

We describe below the implications of the four components listed above.

9.2.1 Planning, Design, and Layout

This component includes the following:

(i) A statement of the objectives of the experiment. A clear and unambiguous hypothesis must be formulated. The investigator must be very familiar with the subject matter, and a thorough literature search and consultations with experts in the area are of extreme importance in developing the knowledge needed to formulate a testable hypothesis. The question you are trying to answer should be very specific and clearly stated. Examples of hypotheses are listed below:

   (a) Alternative Hypothesis

      – Groups are expected to show different results, e.g., rats will gain more weight on diet A rather than diet B.

   (b) Null Hypothesis

      – Groups are expected to be the same, e.g., rats on diet A will gain the same amount of weight as rats on diet B.

   (c) Untestable

      – A result can not be easily defined or determined, e.g., rats on diet A will look better than rats on diet B. What does better mean? A definition of "better" must be clearly stated.

   Once the question and hypothesis have been stated, the methods and techniques to be used can be determined, and evaluated for the best possible method to perform the research.

(ii) Definition of the population about which the inferences are to be made.

(iii) Selection of experimental treatments, which must include a control treatment/s. Sometimes a strictly "untreated" control may seem appropriate, and sometimes in medical vocabulary, a "placebo." Controls generally take four different forms viz., negative, vehicle, positive, and comparative.

(iv) Choosing the plot shape and size.

Choices of (iii) and (iv) must be guided by the principles of replication, randomization, and blocking (local control). These concepts will be fully discussed in section 9.3.

9.2.2 Management

Administration, supervision, and management of an experiment need systematic care if reliable results are to be obtained. If plots are pieces of land, they must be measured up and planted properly. If the treatments are different rates of doses, these must be determined accurately.

Care must be taken to avoid extra variability after treatments have been applied. Extra variability can be introduced into the results by different management of the plots, e.g., by having some plots weeded more carefully or frequently than others in a field experiment where treatment differences are not intended to include weeding differences. Such circumstances may require different blocks to be weeded by different people, so that unwanted *between-laborer* differences are assimilated in *between-block* differences. Likewise, if an experiment has to be irrigated throughout, but there is insufficient equipment to irrigate it all at once, the irrigation should be "by blocks.". Also included in this category are, operations such as harvesting or scoring for diseases.

9.2.3 Data Recording

There is a need for suitable balance for obtaining accurate, precise data values. When weights, heights, and other continuous variables are to be recorded, a "degree of precision of recording" should be specified and adhered to, e.g., "to the nearest 5 g," "to the nearest 25 cm," etc. Data should be recorded on prepared data sheets having proper headings. Copying of data introduces errors and should be avoided if possible.

9.2.4 Scrutiny and Editing of Data

This involves searching for outliers. Quality assurance procedures to identify data entry errors should be developed and incorporated into the experimental design before data analysis.

9.3 Principles of Experimentation

The experimental designs that will be considered in this text are for comparative experiments involving two or more treatments where the object of investigation is to obtain information on the treatments relative to each other. In other words, the interest is on differences between treatment averages rather than on the averages per se. We shall consider the characteristics of designing the experimental arrangements or procedures as follows.

9.3.1 Randomization

Randomization is an objective or fair method of random allocation of the experimental material or treatments in an experiment to the experimental

units. Randomization is also used to ensure that the order in which the trials are performed are random. Thus, randomization subjects all treatments to as nearly equal conditions as possible, and by utilizing chance allotments thereafter. It is known in field experiments for instance, that the errors in adjacent plots are usually positively correlated. Randomization is used to circumvent much of this difficulty. By randomization, the treatments are allocated at random to the experimental unit subject to the design restrictions, so that there is an equal chance of any two treatments appearing in both adjacent and nonadjacent plots. The expected value of the total error for any treatment is, hence, independent of that for any other treatment. Randomization helps:

1. Protect us from "systematic" error that might be caused by subjectively assigning the treatments to the experimental units
2. In averaging out the effects of all uncontrollable conditions or extraneous factors that might exist
3. To obtain unbiased estimates of differences among treatment responses (means or effects)
4. To obtain an unbiased estimate of the error variation in the experiment
5. Validate the underlying statistical assumption that the errors are randomly independently distributed

9.3.2 Replication

Replication refers to the repetition of the basic experiment with the same assignment of treatments. Thus, an experiment that has eight rats allocated to treatment A, say, has eight replicates for treatment A. However, an increase in the number of replicates of a treatment tends to decrease the variation in the estimate of a difference between two treatment means from orthogonal designs. This is the manner in which replication leads to a reduction in the experimental error of differences of treatment effects. Replications allow us to:

1. Estimate the error variance σ^2 due to uncontrollable or assignable causes in the experiment
2. Make experimental result "powerful" enough to recognize true differences between groups (statistical significance) by increasing the accuracy of estimates of means, and other functions of the response variable

However, knowing the exact number of replications is extremely important. It might turn out to be worse to use too few replications than too many. Choosing the correct number of replications is a function of four determinants, viz.,

- d, the minimal difference that is necessary to detect differences in treatment means
- s^2, an estimate of the error variance in the experimental material obtained from a previous experiment or by observing a group of untreated animals

- α, the probability of making a type I error (when you believe that there is a true difference when there is none) which is typically 0.05
- β, the probability of making a type II error (when you believe there is no difference between the treatments, in reality there is) which is typically 0.1. To prevent this type of error, use sufficient replications to prove there is a difference. You must define how much difference is a true difference.

Formulas for determining the proper number of animals using the four variables above can be found in standard statistical texts, but it is wise to consult a statistician. A notable consequence of several replications of an experiment is often an increase in the heterogeneity of the variances of the treatments due to increase variability of the response variable. However, local control or blocking is a technique for dealing with this situation. We discuss the concept of local control or blocking in the next subsection.

9.3.3 Blocking

Grouping of the experimental material in such a manner that the units within a group are more alike than are units in different groups. This kind of grouping is called *blocking* or *stratification*. It allows considerable reduction in error of treatment associated with experimental material. An investigator might decide whether to group experimental units by gender, age, litters, breed, sbp(systolic blood pressure), or by any other factor that may be deemed to influence the response variable.

The above three characteristics will be exemplified with two to three examples, and we shall then later present various types of experimental designs which control various types of heterogeneity among the individual items or units in the investigation.

We have not said anything about the size and the shape of the smallest unit of observation, i.e., the *sampling unit*, nor about the smallest unit to which one treatment is applied, i.e., the *experimental unit*. In some cases, the sampling and experimental units are the same, and in others the experimental and/or the sampling unit size is fixed and cannot be varied. When the size and shape can be varied according to certain criteria, one can talk about the optimum size and shape, but this is a topic unto itself which will not be discussed in this text. We shall assume that the size and shape of the sampling and experimental unit are given. Examples of investigations wherein the unit is fixed are the animal in physiological and nutritional studies, the plant in physiological studies, the individual in learning experiences, a cake or pie in baking studies (a whole cake or whole pie must be baked, even if size and shape can be varied), the classroom for teaching methods (the number of students can be varied, but classroom is fixed), the automobile for road endurance tests, a piece of equipment used to produce or evaluate a product, fixed farms or pastures in certain management investigations, etc.

Example 9.2.1

The first experimental design we shall consider has to do with weighing very light objects. Suppose that one has seven objects (a, b, c, d, e, f, g) to be weighed. One experimental design for weighing the seven objects would be (any order of weighing could be utilized, but they are ordered here for easy reading; Table 9.1):

Table 9.1 First weighing experiment

Weighing	Object weighed
1	Determination for zero correction
2	a
3	b
4	c
5	d
6	e
7	f
8	g

The second design also utilizes eight weighings. In the first weighing all the objects are weighed. In the second weighing, only objects $a, b,$ and d are weighed, in the third only objects $a, c,$ and e, and so on. We see that in the eighth weighing, only objects d, e, f are weighed. The weights of each of the objects are obtained from the following weighings (Table 9.2).

Table 9.2 Second weighing experiment

	Sum or differences of weighing							
Weight of object	1	2	3	4	5	6	7	8
a	+	+	+	+	−	−	−	−
b	+	+	−	−	+	+	−	−
c	+	−	+	−	+	−	+	−
d	+	+	−	−	−	−	+	+
e	+	−	+	−	−	+	−	+
f	+	−	−	+	+	−	−	+
g	+	−	−	+	−	+	+	−

In the second design, therefore, we note that if an object is present in the weighing it receives a +, and a − if not present. The sum of the first four weighings minus the sum of the last four weighings gives the weight of object a for instance, while the sum of weighings 1, 2, 5, and 6 minus the sum of weighings 3, 4, 7, and 8 gives the weight of object b, etc.

In the second design, the *same number* of weighings are used here as for the previous weighing design, but each object has been weighed *four times* rather than only once as in the previous design. This means that the variation in weights from the above design is only one-fourth that of the first design. The weights of each of the objects are obtained as follows:

$$\hat{a} = \frac{1+2+3+4-5-6-7-8}{4}$$
$$\hat{b} = \frac{1+2+5+6-3-4-7-8}{4}$$
$$\hat{c} = \frac{1+3+5+7-2-4-6-8}{4}$$
$$\hat{d} = \frac{1+2+7+8-3-4-5-6}{4}$$
$$\hat{e} = \frac{1+3+6+8-2-4-5-7}{4}$$
$$\hat{f} = \frac{1+4+5+8-2-3-6-7}{4}$$
$$\hat{g} = \frac{1+4+6+7-2-3-5-8}{4}$$

Example 9.2.2

As a second illustrative example which is used to illustrate characteristics (i) and (ii) above, let us suppose that the investigator is comparing four nutritional treatments and is using the rat as the experimental animal. Suppose that he/she randomly selects five rats for each treatment without any regard to the rat's parentage for design I. This allows all four treatments a fair or equal chance to be allotted any 5 of the 20 rats. Design I could look like this:

01 (A)	10 (D)	06 (B)	08 (D)	17 (C)
18 (B)	03 (C)	16 (A)	12 (B)	14 (A)
05 (C)	15 (B)	19 (D)	20 (A)	11 (C)
09 (A)	13 (D)	04 (B)	02 (D)	07 (C)

Design I

Alternatively, suppose that another investigator takes account of the rat's parentage, and uses five litters of four male rats each. (The word litter is used to designate the members born at the same time from the mating of two parents.) Thus, twins in humans would be a litter of size two, triplets would be a litter of size three, etc. In certain types of animals like rabbits, dogs, cats, swine, etc., the members of a litter are brothers and sisters, and usually not identical in genetic composition. The four treatments are then allotted by chance to the four male rats of each of the five litters to form design II. This is "fair" to all four treatments as each has an equal chance at any rat in the litter. In this design, the comparison among treatments is within a litter (i.e., on members of the same litter) and on rats of the same sex. The variation among members of the same litter or family for many characteristics including nutritional response is less than the members of different litters. Hence, design II would be expected to yield treatment means which are less variable than the corresponding means from design I.

Litters				
1	2	3	4	5
A	B	D	B	C
C	A	B	D	B
B	D	C	C	A
D	C	A	A	D

Design II

In fact, from nutritional experiments on swine, it was found that the variation of treatment means compared on individuals of the same litter was about one-half of that obtained when the animals were not grouped or stratified into litters. Practically, this means that the investigators using design II would require only one-half as many animals to obtain the same degree of variation among treatment means as for design I. A simple change of design from I to II would cut the cost of experimentation one-half, or alternatively for a fixed amount of experimentation, it would decrease the variation among treatment means by one-half.

The use of blocking or stratifying experimental material into relatively homogenous groups can greatly increase the efficiency of experimentation. Since total variation is equal to that due to assignable causes, plus bias plus random error, by blocking, a portion of the random error is put into the assignable or controllable category, thereby, reducing the amount of variation in the chance or random category.

9.4 Methods of Increasing the Accuracy of an Experiment

Accuracy refers to the success of estimating the true value of a quantity. It is often confused with precision which refers to the clustering of sample values about their average. Precision can be thought of as the inverse of variance, while accuracy involves both biasedness and precision.

Often the experimenter has to choose between an unbiased estimate with rather low precision (high variance), and a slightly biased one with high precision. The choice of the proper estimate is often dictated by circumstances beyond his/her control, but certain methods of increasing the accuracy of the experiment should be kept in mind.

(i) Accuracy can be generally increased by increasing the size of the experiment. One must, however, be careful not to introduce heterogeneity into the experiment by poor supervision with a possible biased result.

(ii) Experimental techniques should be refined as much as possible by making sure that:

   (a) Uniform method of applying treatment to the experimental units is adopted.
   (b) Sufficient control over external influence so that every treatment operates under as nearly the same conditions as possible.

(c) Checks are set up to avoid gross errors in recording and analyzing the data.

(iii) Experimental material should be selected to suit the experiment, i.e.,

(a) Size and shape of the experimental unit (plot) are prepared to achieve maximum accuracy, and unbiasedness.
(b) Often, additional measurements can be taken to help explain the final results e.g., covariance analysis techniques.
(c) Treatments should be grouped together in the best manner. In other words, the proper selection of the experimental design is of the utmost importance, e.g., if too many treatments or the experimental units are quite heterogeneous, then an incomplete block design will be suitable or if interactions are assumed to be important, then factorial system will be suitable. If high-order interactions are not important, then some system of confounding might be used, etc.

9.5 Random and Fixed Effect Models

The model for most designs is of the form

$$Y_{ij} = \mu + t_i + e_{ij} \tag{9.1}$$

with $i = 1, 2, \ldots, t$ (the number of treatments) and $j = 1, 2, \ldots, r_i$, the number of replications for treatment i. Equation (9.1) is a regression model with $\sum t_i = 0$. The e_{ij} (the errors) are independent of the t_i. One method of distinguishing between the t's and the e's has been to call the t's the fixed effects, and e's the random effects.

By fixed effects, we imply that all the treatments about which inferences are to be made are included in the experiment. A random effect is assumed to be one of a large number of possible effects: in general we shall refer to the number of possible effects to be infinite, i.e., a random sample from a larger population of treatments. In this situation, we should be able to extend the conclusions (which are based on the sample of treatments) to all treatments in the population, whether they were explicitly considered in the analysis or not. Here, the t_iS are random variables and knowledge about the particular ones investigated is relatively useless. Instead, we would test hypotheses about the variability of the t_i, and try to estimate this variability. This is called the *random effects* or *components of variance* models.

9.6 Control of Error

Error control can be accomplished by:

1. Blocking (e.g., split plot, incomplete block designs, etc.)
2. Plot technique, that is size and shape of the experimental units
3. Data analysis (use of concomitant observations as in analysis of covariance).

9.7 Summary of Principles of Experimental Designs

At the beginning of this chapter, we listed three desirable characteristics for designing experiments or what we are sometimes described as the three principles of design, viz., *replication*, *randomization*, and *local control*.

Randomization and replication are necessary to obtain a valid estimate or measure of the experimental variation. Replication and "local control" (= blocking or grouping) are necessary to achieve a reduction in the random variation among treatment effects in the experiment. The use of "local control" has been made throughout this book in blocking or grouping to eliminate or to control the various sources of variation.

In addition to the three basic principles of experimental design mentioned above, we can also consider perhaps a fourth principle, viz., *orthogonality*. Orthogonality is important in order to estimate the random variation between treatment means is the same for all pairs of treatments having equal replication, and having the same degree or magnitude of random error variation.

If orthogonal designs are not possible, then we strive for balanced designs which will ensure that differences between pairs of treatment effects for all have the same variance. In balanced designs, all treatment pairs occur equally, frequently with each other in the b blocks of size k for the v treatments. Since $bk =$ total number of experimental units and since there are v treatments each repeated r times then $bk = vr$ in balanced designs. Orthogonal designs are balanced designs, but the reverse is not true. The randomized complete block design is a balanced design, but the balanced incomplete block design is not an orthogonal design. We shall discuss the concept of orthogonality further when we become familiar with the randomized complete block design.

9.8 Observational Studies

In observational study, unlike in experimental study, the researcher does not have control over the assignments of treatments or exposure to a certain disease (e.g., cancer). Thus, neither the subjects under study or any of the factors of interest are determined by the investigator. There are two types of observational data, viz., *prospective studies* and *retrospective studies*.

9.8.1 Prospective Studies

A prospective study or *cohort study* or *follow-up study* is a study in which two random samples of subjects were selected, one having the presence (the exposed) of the suspected antecedent or risk factor (e.g., smoking), and the

other sample consists of subjects which do not have the suspected antecedent or risk factor. The subjects are then followed up to some time into the future (i.e., prospectively), and the proportions of subjects developing the disease (outcome variable or response, e.g. lung cancer) at some point in time are then estimated for both samples. The 2×2 Table 9.3 gives the distribution subjects into each of the four categories over a specified period of time.

Table 9.3 Classification of a sample of n subjects

	Disease status		
Risk factor	Present	Absent	Total at risk
Present (exposed)	a	b	$a+b$
Absent (unexposed)	c	d	$c+d$
Total	$a+c$	$b+d$	n

Here we note that for this scheme, the sum of probabilities across each row adds to 1, i.e.,

$$\left(\frac{a}{a+b}\right) + \left(\frac{b}{a+b}\right) = 1$$
$$\left(\frac{c}{c+d}\right) + \left(\frac{d}{c+d}\right) = 1$$

We may also note that this scheme is often very expensive and time-consuming to undertake.

9.8.2 Relative Risk

For data arising from prospective studies, the risk of development of the disease is computed as $\left(\frac{a}{a+b}\right)$. Similarly, the risk among the unexposed subjects is similarly computed as $\left(\frac{c}{c+d}\right)$. Hence, the *relative risk* is the ratio of the risk of developing the diseases among exposed subjects to the risk of developing the diseases among unexposed subjects i.e.,

$$\widehat{RR} = \frac{\Pr(\text{disease}|\text{exposed})}{\Pr(\text{disease}|\text{unexposed})} = \frac{a/(a+b)}{c/(c+d)} = \frac{a(c+d)}{c(a+b)} \tag{9.2}$$

and a $100(1-\alpha)\,\%$ confidence interval (CI) for RR or ψ as is sometimes succinctly represented is computed as:

$$100(1-\alpha)\%\ CI = \hat{\psi}^{1\pm(z_\alpha/\sqrt{X^2})} \tag{9.3}$$

where, z_α is the two-sided z value and X^2 corresponds to Pearson's test statistic.

Consider the following example that examines the risk factors for breast cancer among women participation in the US National Health and Nutrition Examination Survey (Carter et al. 1989). In this study, a woman is considered exposed if she first gave birth at age 25 or more. In a sample of 4550 women who gave birth to their first child before the the age of 25, 65 developed breast cancer. Of the 1628 women who gave birth at age 25 or more, 31 were diagnosed with breast cancer. The data from this study are displayed in Table 9.4.

Table 9.4 Study data on breast cancer in women

	Disease status		
Risk factor	Present	Absent	Total
Present (exposed)	31	1597	1628
Absent (unexposed)	65	4485	4550
Total	96	6082	6178

$$\begin{aligned}\hat{\psi} &= \frac{\Pr(\text{disease}|\text{exposed})}{\Pr(\text{disease}|\text{unexposed})} \\ &= \frac{31/1628}{65/4550} \\ &= 1.3329.\end{aligned}$$

The proportion having breast cancer, therefore, was 1.33 times higher for those women who had their first birth at a later age than those who gave birth at an earlier age.

$$X^2 = \frac{6176[(31)(4485) - (65)(1597)]^2}{(96)(6082)(1628)(4550)} = 1.7729.$$

Hence, the 95 % CI is calculated as:

$$\begin{aligned}\hat{\psi}^{1\pm(z_\alpha/\sqrt{X^2})} &= 1.3329^{1\pm(1.96/\sqrt{1.7726})} \\ &= 1.3329^{1\pm1.4721} \\ &= (1.3329^{-0.4721}, 1.3329^{2.4721}) \\ &= (0.8731, 2.0348).\end{aligned}$$

Thus the 95 % CIs for ψ are computed to be (0.8731,2.0348).

9.8.3 Retrospective Studies

A retrospective study or *case-control study* is the opposite of prospective study. Here, the samples are selected from those having the outcome variable

(disease of interest), and the researcher then looks back (i.e., takes a retrospective look) at the subjects and classify them according to those that have (or had) or have not been exposed to the risk factor. Under this scheme, we again have the following as given in Table 9.5.

Table 9.5 Classification of a sample of n subjects

	Sample		
Risk factor	Present (cases)	Absent (controls)	Total
Present	a	b	$a+b$
Absent	c	d	$c+d$
Total	$a+c$	$b+d$	n

Here we note that for this scheme, the sum of probabilities across each columns adds to one, i.e.,

$$\left(\frac{a}{a+c}\right)+\left(\frac{c}{a+c}\right)=1$$
$$\left(\frac{b}{b+d}\right)+\left(\frac{d}{b+d}\right)=1$$

9.8.4 Odds Ratio

For data arising from retrospective studies, relative risk measure would not be appropriate. Instead, we would compute what is known as the odds ratio. First, we recall that the odds of an event is defined as the ratio of P(A)/[1-P(A)]. Thus,

(a) The odds of being a case (disease present) in Table 9.5 to being a control among subjects with the risk factor is computed as:

$$\frac{a/(a+b)}{b/(a+b)}=\frac{a}{b}$$

(b) The odds of being a case (disease present) to being a control (disease absent) among subjects without the risk factor is computed as:

$$\frac{c/(c+d)}{d/(c+d)}=\frac{c}{d}$$

(c) The estimated odds ratio OR or simply θ is therefore computed as:

$$\hat{\theta}=\frac{a/(b)}{c/(d)}=\frac{ad}{cd} \tag{9.4}$$

It is sometimes advocated that for those situations where there are zero cell frequencies, the estimate can be computed as:

$$\hat{\theta}^* = \frac{(a+0.5)(d+0.5)}{(c+0.5)(d+0.5)}$$

(d) A $100(1-\alpha)\,\%$ CI for θ is again computed as:

$$100(1-\alpha)\%\ CI = \hat{\theta}^{1\pm(z_\alpha/\sqrt{X^2})} \tag{9.5}$$

Alternatively, we can use the fact that the asymptotic variance of $\hat{\theta}$ is given by:

$$\hat{V}(\hat{\ln\theta}) = \left[\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right] \tag{9.6}$$

and hence, the CI is computed as:

$$e^{\ell \pm z_{\alpha/2}\sqrt{\hat{V}}}$$

where ℓ is the log of the estimated odds ratio. We give an example below.

Example

The following example is a case-control study (Hennekens et al. 1984), where two samples of women were identified with and without breast cancer, and their records were retrospectively examined to determine whether they have been exposed to the use of oral contraceptive. Among the 989 women in the study who had breast cancer, 273 had in the past used oral contraceptives and 716 had not. Of the 9901 women who did not have breast cancer, 2641 had used oral contraceptives and 7260 had not.

Table 9.6 Study data on breast cancer in women

	Sample		
Risk factor	Cases	Controls	Total
Present (exposed)	273	2641	2914
Absent (unexposed)	716	7260	7976
Total	989	9901	10,890

Hence,

$$\hat{\theta} = \frac{273 \times 7260}{716 \times 2641} = 1.0481$$

Thus, women who had used oral contraceptives in the past have odds of developing breast cancer, i.e., only 1.048 times the odds of nonusers. This odds may not be significant. The Pearson's X^2 statistic is computed as:

$$X^2 = \frac{10890[(273)(7260)-(716)(2641)]^2}{(2914)(7976)(989)(9901)} = 0.3965$$

Hence, the 95 % CI is calculated as:

$$\begin{aligned}\hat{\theta}^{1\pm(z_\alpha/\sqrt{X^2})} &= 1.0481^{1\pm(1.96/\sqrt{0.3965})} \\ &= 1.0481^{1\pm 3.1127} \\ &= (1.0481^{-2.1127}, 1.0481^{4.1127}) \\ &= (0.9055, 1.2131)\end{aligned}$$

The 95 % CIs for θ are computed to be (0.9055, 1.2131). Since this CI does not include zero, we can conclude that the use of oral contraceptives in the past by women does not seem to cause significant difference in the rate of breast cancer in the female population. MINITAB can compute the odds ratio as well as X^2 for us but unfortunately, could not compute the CIs, etc. We present a typical MINITAB out for this problem.

```
Tabulated statistics: risk, case

Using frequencies in f
Rows: risk   Columns: case

          n    y    All

a      7260  716   7976
p      2641  273   2914
All    9901  989  10890

Cell Contents:      Count

Pearson Chi-Square = 0.396, DF = 1, P-Value = 0.529
Likelihood Ratio Chi-Square = 0.394, DF = 1, P-Value = 0.530

Results for all 2x2 tables

Common odds ratio  1.04814

MHCstatistic  DF   P-Value
     0.350407   1  0.553883
```

Employing the alternative approach, we have

$$\hat{V} = \frac{1}{273} + \frac{1}{2641} + \frac{1}{716} + \frac{1}{7260} = 0.00557334$$

and hence, $\sqrt{\hat{V}} = 0.0747$. A 95 % CI is computed as:

$$e^{\ln(1.0481)\pm 1.96(0.0747)} = e^{0.0470\pm 0.1464} = e^{-0.0994, 0.1934} = (0.9053, 1.2134)$$

9.9 Exercises

1. Some terms that occur rather frequently in the literature are:

   (a) Accuracy
   (b) Precision
   (c) Validity
   (d) Reliability
   (e) Bias

   Restricting your remarks to the theory of statistics or to applications of statistical methods, define and discuss each of these terms.
2. What is meant by a "control treatment"?
3. Define:

   (a) Absolute experiment
   (b) Comparative experiment

   Give examples of each.
4. Describe the role of replication in the design of experiments.
5. Describe how we interpret the estimated values of the following measures:

   - Relative risk
   - Odds ratio

6. Define the following terms: relative risk, prospective study, and odds ratio.
7. Two treatments, heparin and enoxaparin, were compared in a double-blind, randomized clinical trial of patients with coronary artery disease (Samuels and Witmer 1999). The subjects are classified as having a positive or negative response to treatment. The data is presented below.

	Treatments		
Outcome	Heparin	Enoxaparin	Total
Negative	309	266	
Positive	1255	1341	
Total	1564	1607	

   (a) Estimate the odds ratio θ
   (b) Construct a 95 % CI for the population value of θ.
   (c) Based on your results in (a) and (b), is there any evidence of significant association between the outcome and treatment at the 0.05 level of significance?

Chapter 10
The Completely Randomized Design

10.1 Introduction

The completely randomized design (CRD) is the simplest of all experimental designs, both in terms of analysis and experimental layout. Here, treatments are randomly allocated to the experimental units entirely at random. Thus if a treatment is to be applied to five experimental units, then each unit is deemed to have the same chance of receiving the treatment as any other unit. The CRD is often used if we believe that the experimental material is homogeneous or uniform. In this case, the experimental units are regarded as a group and the investigator believes that the experimental material available contains only nonassignable variation, and that it would be impossible to try to group the material into blocks or some other subgroups such that the variation among subgroups is larger than among units within subgroups as far as the response variable under investigation is concerned. Usually, though not necessarily, the random assignment is restricted in such a manner as to have an equal number of experimental units assigned to each treatment. The CRD should therefore be used where extraneous factors can easily be controlled, such as in laboratories or green houses. The CRD is usually the choice design in pilot studies where experimental units and conditions are homogeneous.

To illustrate the above, suppose that an animal nutritionist has four diets, A, B, C, and D and he wants to allocate five rats to each diet. The response is the weight gained after 6 weeks. The diets will then be randomly allocated to the 20 rats. The 20 rats are treated alike in all other respects except for type of diet, that is, they all are in the same pen and have the same food and water sources. The intermingling of 20 rats results in the rats all being subjected to the elements of the environment in the enclosure. The diets (treatments) are compared in as nearly equitable manner as possible. One possible random assignment is to do the following.

Step 1 Label the animals from 1 to 20 and put these in Column C1.

Step 2 Randomly select five animals using MINITAB from C1 and place them in C2. These are {11, 4, 1, 7, 12}.

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_10,

Step 3 Remove the five digits in step 2 from C1 and put the remaining digits in C3 where again you perform step 2, yielding the digits {2, 10, 19, 18, 14}.

Step 4 Repeat step 3 by removing the digits generated in steps 2 and 3 from C1 and again randomly select five digits from the remaining 10 digits yielding once again {6, 17, 5, 9, 3}. The remaining five digits are therefore {8, 13, 15, 16, 20}.

Step 5 If we denote diet A as the first group, is then assigned to the animals selected from step 2. The last group, which is diet D received rats numbered {8, 13, 15, 16, 20}.

Table 10.1 gives the final random allocation for this experiment.

The above layout for an experiment in a completely randomized design might be appropriate for 20 pots on a greenhouse bench or a series of soil analyses involving four treatments.

The design is the simplest of all experimental designs because it involves zero-way or no elimination of heterogeneity in the experimental material. The total variation in the experiment may be written as:

Table 10.1 A CRD layout with four treatments and five replications

Rats	Diet	Rats	Diet	Rats	Diet	Rats	Diet
1	A	6	C	11	A	16	D
2	B	7	A	12	A	17	C
3	C	8	D	13	D	18	B
4	A	9	C	14	B	19	B
5	C	10	B	15	D	20	D

Total variation = variation among treatment means + error variation

Observed yield = treatment + random (unit) variation.

The yield of any experimental unit may be written as treatment mean + an error term. It is permissible to use the above form when the different components of variation are additive in their effects.

The above design is sometimes called a one-way classification or single-factor design because the data can be classified in only one way, namely, according to the treatment, and treatment is the only factor involved.

The model for the design is given below as:

$$Y_{ij} = \mu + t_i + \varepsilon_{ij} \quad \begin{matrix} i = 1, 2, \cdots, t \\ j = 1, 2, \cdots, r \end{matrix} \tag{10.1}$$

where:

Y_{ij} is the yield for treatment i in the jth replicate

μ is the overall mean

t_i is the effect of the ith treatment, and,

ε_{ij} is a random error component.

The data from such an experiment will appear as in Table 10.2.

Table 10.2 Table of observations for a completely randomized design

Treatments	Observations				Total
1	y_{11}	y_{12}	$\cdots$	y_{1r}	Y_{1+}
2	y_{21}	y_{22}	$\cdots$	y_{2r}	Y_{2+}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
t	y_{t1}	y_{t2}	$\cdots$	y_{tr}	Y_{t+}

10.1.1 Analysis of Variance of Table 10.2

There are $r \times t = rt$ observations in the experiment. Hence,

$$\text{The correction factor (CF)} = \frac{Y_{++}^2}{rt} = \frac{G^2}{rt}$$

$$\text{Total SS} = y_{11}^2 + y_{12}^2 + \cdots + y_{tr}^2 - CF = \sum_i \sum_j y_{ij}^2 - \frac{Y_{++}^2}{rt} = TSS$$

$$GSS = \frac{Y_{1+}^2}{r} + \frac{Y_{2+}^2}{r} + \cdots + \frac{Y_{t+}^2}{r} - CF = \sum_i \frac{Y_{i+}^2}{r} - \frac{Y_{++}^2}{rt}$$

The error sum of squares (SS) is obtained by subtraction as Total SS−Treatment SS, or as:

$$SSE = \sum_i \sum_j y_{ij}^2 - \frac{Y_{i+}^2}{r}$$

Hence the analysis of variance table is given in Table 10.3.

Table 10.3 Analysis of variance table for a CRD

Source of variation	d.f.	SS	MS	F
Treatments	$t-1$	GSS	$\frac{\text{GSS}}{t-1} = A$	$\frac{A}{S^2}$
Error	$t(r-1)$	SSE	$\frac{\text{SSE}}{t(r-1)} = S^2$	
Total	$rt-1$	TSS		

Since we are interested in testing the equality of the t treatment effects, the appropriate hypotheses are

$$\begin{aligned} H_0: &\quad t_1 = t_2 = \cdots = t_t = 0 \\ H_a: &\quad t_i \neq 0 \quad \text{for at least one } i \end{aligned} \tag{10.2}$$

The hypotheses in (10.2) are equivalent to the following in terms of the population means of the treatments.

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = \cdots = \mu_t \\ H_a: &\quad \text{at least two of the means are not equal} \end{aligned} \tag{10.3}$$

The appropriate test procedure is derived from the analysis of variance table in Table 10.3. The value under the F column $\frac{A}{S^2}$ is, when H_0 is true, distributed as an F distribution with $(t-1)$ and $t(r-1)$ degrees of freedom. This value can be compared with the tabulated F value with the corresponding pairs of degrees of freedom at a specified α level (Table 4 in the Appendix).

10.2 Example 10.1

In an experiment to compare melon variates, six plots of each of four varieties were grown, the plots being allotted to varieties in a completely random manner, and the results are given below in Table 10.4.

$$\text{Total for all observations} = 643.69 = \text{G. Here } r = 6, t = 4 \text{ and } rt = 24.$$

$$\text{CF} = (643.93)^2/24 = 17264.034$$

$$\begin{aligned} \text{Total variation (SS)} &= 2629.23 + 8472.09 + 2434.12 + 5387.73 - \text{CF} \\ &= 1659.1303 \end{aligned}$$

Alternatively, this could be computed as,

Table 10.4 Results of the experiment

	Varieties			
	A	B	C	D
	25.12	40.25	18.30	28.05
	17.25	35.25	22.60	28.55
	26.42	31.98	25.90	33.20
	16.08	36.52	15.05	31.68
	22.15	43.32	11.42	30.32
	15.92	37.10	23.68	27.58
$\sum y$	122.94	224.42	116.95	179.38
$\sum y^2$	2629.23	8472.09	2434.12	5387.73
Mean $\bar{y}$	20.49	37.40	19.49	29.90
Variance (S_i^2)	22.04	15.61	30.91	4.97

$$25.12^2 + 17.25^2 + \cdots + 27.58^2 - CF = 18,923.1643 - CF$$
$$= 1659.1303$$

$$\text{Treatment SS} = \frac{122.94^2}{6} + \frac{224.42^2}{6} + \frac{116.95^2}{6} + \frac{179.38^2}{6} - \text{CF}$$
$$= 1291.4771$$

Note that the divisor 6 comes about because each sum in the $\sum y$ unit comes from six observations. The error SS is obtained by subtraction as:

$$\text{Error SS (SSE)} = \text{Total SS} - \text{Treatment SS} = 367.66$$

The complete analysis of variance table for the data is presented in Table 10.5.

Table 10.5 Analysis of variance table

Source	d.f.	SS	MS	F
Between varieties	3	1291.477	430.492	23.4179
Error	20	367.655	$18.383 = (S^2)$	
Total	23	1659.130		

The estimated standard error of a variety mean is

$$\sqrt{\frac{S^2}{r}} = \sqrt{\frac{18.383}{6}} = 1.7504.$$

The estimated standard error of a difference between two variety means based on r_1 and r_2 observations is

$$\sqrt{\frac{S^2}{r_1} + \frac{S^2}{r_2}} \quad \text{or} \quad \sqrt{\frac{2S^2}{r}} \quad \text{if } r_1 = r_2 = r.$$

Hence in our example where all treatments are equally replicated, this becomes,

$$\sqrt{\frac{2S^2}{r}} = \sqrt{\frac{2 \times 18.383}{6}} = 2.4754 \tag{10.4}$$

The standard error in (10.4) is based on 20 d.f. (the error d.f.).

We can therefore present the results as follows:

Variety	Mean yield
A	20.5
B	37.4
C	19.5
D	29.9

For an overall test of whether the varieties give different yields, we cross-check the computed F value of 23.42 with an F distribution with 3 and 20 degrees of freedom. At $\alpha = 0.01$, the tabulated F value is 3.10, and there is therefore strong evidence that there are real differences in yielding abilities between the varieties.

Without further knowledge about the four varieties, the making of particular comparisons between pairs of varieties is rather dangerous. However it is clear that B is the best variety and that D is probably better than A and C.

10.2.1 Students' t Test

In order to compare any two treatment means say μ_1 and μ_2 respectively, we need to compute the following:

$$t^* = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{2S^2}{r}}} \tag{10.5}$$

and compare with the tabulated t value with the error degrees of freedom. Thus in our example above, the value of t for the comparison between A and D is,

$$t^* = \frac{29.9 - 20.5}{2.48} = \frac{9.4}{2.48} = 3.79$$

which is significant since the 1 % significance level for t on 20 d.f. is 2.852 from Appendix Table 2.

Alternatively, we could have obtained

$$t_{.01}(20\ d.f.) \times \sqrt{\frac{2S^2}{r}} = 2.85 \times 2.48 = 7.068$$

In the foregoing, $t_{\alpha/2}\ (20\ d.f.) \times \sqrt{\frac{2S^2}{r}}$ is called the *least significance difference* (LSD). And our hypothesis is rejected whenever $|(\bar{y}_i - \bar{y}_j)|$ is greater than the LSD for all $i \neq j$.

An immediate consequence of the above t test is that, in order to determine the best variety, we need to have made comparisons on all possible pairs of treatment means (AB, AC, AD, BC, BD, and CD) that is six possible pairs of means, where for equally replicated experiments, the difference for each pair of means is compared with the LSD. However, it has been recognized that if several t tests were performed, the probability that at least one of these is apparently significant is greater than 0.05. If the t tests are independent, this probability is 0.23 for five tests, 0.40 for ten tests and 0.64 for 20 tests. The implication here is that with five comparisons for instance, the level

of significance is no longer 5 % but 23 % (inadvertently high!). Conclusions therefore arrived at will certainly be erroneous and these observations lead to the problem that is commonly referred to as *Multiple Comparison*.

10.3 Multiple Comparisons of Means

Among several methods suggested for combating the problem of multiple comparison, three methods are mostly favored by biologists and investigators in agriculture. These are described in this section; the first of these is *Duncan's Multiple Range* test.

10.3.1 Duncan's Multiple Range Test

Instead of making all comparisons in relation to a single significance difference (LSD) as in the t test, the size of the LSD is adjusted depending upon whether the two means being compared are adjacent or whether one or more other means fall between those being compared.

To apply Duncan's multiple range test for equal sample sizes, we first compute:

$$\text{Least Significance Ranges} = \text{LSR} = K_r \sqrt{\frac{S^2}{r}}$$

where the K values are obtained from Duncan's table of significant ranges (Appendix Tables 7 and 8).

For the data in our example, the standard error of a mean is $\sqrt{\frac{S^2}{6}} = 1.75$. From Duncan's table with 20 d.f. and $\alpha = 0.05$, we obtain readings (K) for the different ranges of mean (r) to be compared. The values are

$$\begin{aligned} r &= 2 \quad K = 2.95 \\ r &= 3 \quad K = 3.10 \\ r &= 4 \quad K = 3.18 \end{aligned}$$

The least significant ranges are thus for

$$\begin{aligned} r &= 2 \quad R_2 = 2.95 \times 1.75 = 5.160 \\ r &= 3 \quad R_3 = 3.10 \times 1.75 = 5.425 \\ r &= 4 \quad R_4 = 3.18 \times 1.75 = 5.750. \end{aligned}$$

To test the differences between the various means, we next rank the means from the smallest to the largest resulting in,

$$\bar{Y}_C = 19.5, \quad \bar{Y}_A = 20.5, \quad \bar{Y}_D = 29.9, \quad \bar{Y}_B = 37.4.$$

The comparisons would yield

$$\text{B vs. C} = 37.4 - 19.5 = 17.9 > 5.75 \quad (R_4)$$
$$\text{B vs. A} = 37.4 - 20.5 = 16.9 > 5.43 \quad (R_3)$$
$$\text{B vs. D} = 37.4 - 29.9 = 7.5 > 5.16 \quad (R_2)$$
$$\text{D vs. C} = 29.9 - 19.5 = 10.4 > 5.43 \quad (R_3)$$
$$\text{D vs. A} = 29.9 - 20.5 = 9.4 > 5.16 \quad (R_2)$$
$$\text{A vs. C} = 20.5 - 19.5 = 1.0 < 5.16 \quad (R_2)$$

From the above analysis, it can be concluded that there are significant differences between all pairs of means except A and C

Results of Duncan's Multiple Range Test

$\bar{Y}_A$	$\bar{Y}_C$	$\bar{Y}_D$	$\bar{Y}_B$
20.5	19.5	29.9	37.4

It is obvious that variety B is the best of all the other treatments. To prevent contradictions, no differences between a pair of means are considered significant if the means involved fall between two other means that do not differ significantly.

10.3.2 Tukey's Test

Here we compute:

$$\text{Significance Difference (SD)} = q_r \times \sqrt{\frac{S^2}{r}} \tag{10.6}$$

where the needed q_r value is obtained from tables of significant studentized ranges (two tailed, Appendix Table 6). Thus for our example, $r = 4, \alpha = 0.05$, and $\sqrt{\frac{S^2}{r}} = 1.75$; hence $q_4(20\text{d.f.}) = 3.96$ where $r = 4$ is the number of means to be compared. Now $\text{SD} = 1.75 \times 3.96 = 6.93$. In testing differences between the various means, in all instances, if the difference between any two means is larger than the SD = 6.93, then the means are assumed to be significantly different. The results from this test are presented in Table 10.6. Note that this procedure was earlier introduced in Chap. 6.

10.3.3 Scheffé's Test

Scheffé's test is similar to Tukey in that the same significant difference is also computed. The only difference is that the Scheffé's test makes use of F

tables. It is also more stringent than the Tukey test. Thus the probability of type I error is lower.

To use this test, we compute,

$$\text{SD} = \sqrt{(t-1)\, F_{(t-1,E\ d.f.)}} \times \sqrt{\frac{2S^2}{r}}$$

Where t is the number of treatments and E is the error degrees of freedom. For our data in Table 10.4, $\sqrt{\frac{2S^2}{r}} = 2.48$ and $F_{(3,20)}$ at $\alpha = 0.05 = 3.10$, hence,

$$\text{SD} = \sqrt{(4-1)F_{(3,20)}} \times \sqrt{\frac{2S^2}{r}} = \sqrt{3 \times 3.10} \times 2.48 = 7.56$$

Table 10.6 Summary of results for the three methods when applied to our example

Comparison	Difference	Tukey	Scheffé	Duncan
B vs. C	17.9	SG	SG	SG
B vs. A	16.9	SG	SG	SG
B vs. D	7.5	SG	NS	SG*
D vs. C	10.4	SG	SG	SG
D vs. A	9.4	SG	SG	SG
A vs. C	1.0	NS	NS	NS
		6.93	7.56	

where

SG – Significant and

NS – Not significant

While the Tukey and Duncan tests produced significant difference between varieties B and D, the Scheffé test indicates that these are not significantly different.

Other tests that have been employed in multiple comparison problems are the Newman–Keuls' and Dunnett's procedures which are not the subject of discussion in this text. It must be noted however that the comparisons to be studied should be selected in advance of any analysis of the data—indeed before conducting the experiment. The analysis of variance for the data in Table 10.4 is carried out in MINITAB in two ways. The first consists of reading the data into four columns C1–C4. This produces the ANOVA table, table of means, and individual $100(1-\alpha)\%$ confidence intervals but does not do pairwise comparisons. The results are presented below.

```
MTB > PRINT C1-C4

Data Display
 Row       A       B       C       D

   1   25.12   40.25   18.30   28.05
   2   17.25   35.25   22.60   28.55
   3   26.42   31.98   25.90   33.20
   4   16.08   36.52   15.05   31.68
   5   22.15   43.32   11.42   30.32
   6   15.92   37.10   23.68   27.58

MTB > AOVOneway 'A' 'B' 'C' 'D'.

One-way ANOVA: A, B, C, D

Analysis of Variance
Source     DF        SS        MS        F        P
Factor      3    1291.5     430.5    23.42    0.000
Error      20     367.7      18.4
Total      23    1659.1
                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
Level      N      Mean     StDev  --------+---------+---------+---------
A          6    20.490     4.694    (----*----)
B          6    37.403     3.950                              (----*-----)
C          6    19.492     5.560   (----*----)
D          6    29.897     2.230                 (-----*----)
                                  --------+---------+---------+---------
Pooled StDev =    4.288               21.0      28.0      35.0
```

The ANOVA table results agree with those presented in Table 10.5. We present an alternative way for the analysis in MINITAB. Here, we have read in the varieties in one column (C1) and the yield in another column (C2). The analysis of variance is then implemented with the request for Fisher's LSD and Tukey's Test. Again the results are presented in the following:

```
MTB > print c1-c2

Data Display


 Row   VART   YIELD

   1      A   25.12
   2      A   17.25
   3      A   26.42
   4      A   16.08
   5      A   22.15
   6      A   15.92
   7      B   40.25
   8      B   35.25
   9      B   31.98
  10      B   36.52
  11      B   43.32
  12      B   37.10
  13      C   18.30
  14      C   22.60
```

```
  15      C   25.90
  16      C   15.05
  17      C   11.42
  18      C   23.68
  19      D   28.05
  20      D   28.55
  21      D   33.20
  22      D   31.68
  23      D   30.32
  24      D   27.58

MTB > Oneway 'YIELD' 'VART';
SUBC>   Tukey 5;
SUBC>   Fisher 5;
SUBC>   MCB 5 +1.

One-way ANOVA: YIELD versus VART

Source  DF      SS     MS      F      P
VART     3  1291.5  430.5  23.42  0.000
Error   20   367.7   18.4

Total   23  1659.1

S = 4.288   R-Sq = 77.84%   R-Sq(adj) = 74.52%


                        Individual 95% CIs For Mean Based on
                        Pooled StDev
Level  N    Mean  StDev  -------+---------+---------+---------+--
A      6  20.490  4.694   (----*----)
B      6  37.403  3.950                            (----*-----)
C      6  19.492  5.560  (----*----)
D      6  29.897  2.230               (-----*----)
                        -------+---------+---------+---------+--
                            21.0      28.0      35.0      42.0

Pooled StDev = 4.288


Grouping Information Using Tukey Method

VART  N    Mean  Grouping
B     6  37.403  A
D     6  29.897     B
A     6  20.490      C
C     6  19.492      C

Means that do not share a letter are significantly different.


Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of VART

Individual confidence level = 98.89%
```

```
VART = A subtracted from:

VART   Lower  Center   Upper  -------+---------+---------+---------+--
B      9.982  16.913  23.845                           (---*----)
C     -7.930  -0.998   5.933              (---*----)
D      2.475   9.407  16.338                     (---*----)
                             -------+---------+---------+---------+--
                                  -15         0         15        30


VART = B subtracted from:

VART    Lower   Center    Upper  -------+---------+---------+---------+--
C     -24.843  -17.912  -10.980  (----*----)
D     -14.438   -7.507   -0.575         (----*----)
                             -------+---------+---------+---------+--
                                  -15         0         15        30


VART = C subtracted from:

VART  Lower  Center   Upper  -------+---------+---------+---------+--
D     3.474  10.405  17.336                    (----*----)
                           -------+---------+---------+---------+--
                                -15         0         15        30
```

For the four varieties A, B, C, and D, there are $\binom{4}{2} = 6$ pairwise comparisons. MINITAB does these by computing the confidence interval for two means, $(\mu_1 - \mu_2)$ and check whether the interval includes zero. If zero is included, then we would fail to reject H_0. Otherwise, the two means are significantly different. In the above results from the Tukey's test, only A and C are not significantly different from one another, which agrees with our result displayed in Table 10.6.

Similarly, the results from Fisher's LSD test are also presented below.

```
Grouping Information Using Fisher Method

VART  N    Mean  Grouping
B     6  37.403  A
D     6  29.897     B
A     6  20.490        C
C     6  19.492        C

Means that do not share a letter are significantly different.

Fisher 95% Individual Confidence Intervals
All Pairwise Comparisons among Levels of VART

Simultaneous confidence level = 80.83%


VART = A subtracted from:

VART   Lower  Center   Upper  ---------+---------+---------+---------+
B     11.750  16.913  22.077                                (---*---)
C     -6.162  -0.998   4.165                    (---*---)
D      4.243   9.407  14.570                             (---*---)
                           ---------+---------+---------+---------+
                                  -12         0         12        24
```

```
VART = B subtracted from:

VART    Lower   Center    Upper  ---------+---------+---------+---------+
C     -23.075  -17.912  -12.748  (---*---)
D     -12.670   -7.507   -2.343           (----*---)
                                ---------+---------+---------+---------+
                                       -12         0         12        24


VART = C subtracted from:

VART  Lower  Center   Upper  ---------+---------+---------+---------+
D     5.241  10.405  15.569                      (----*---)
                             ---------+---------+---------+---------+
                                    -12         0         12        24
```

Here again, only the pair A and C are found not to be significant as the computed interval $(-4.165, 6.162)$ again includes zero. This result is again consistent with our earlier results presented in Table 10.6.

10.4 The Unbalanced Case

In some single-factor experiments, the number of observations (replications) taken on each treatment may be different. We then say that the design is unbalanced. The analysis of variance described above may still be used, but slight modifications must be made in the SS formula. We illustrate these modifications with an example below.

10.4.1 Example 10.2

Three fertilizers A, B, and C were applied to 15 plots chosen at random in a field of strawberries such that fertilizer A is applied to 4 plots, B to 6 plots and C to 5 plots. The total crop yields (lbs) (coded) by these plots over the entire season were recorded and are presented in Table 10.7.

Table 10.7 Yields from application of three fertilizers on a field of strawberries

Fertilizer	Yields	Total
A	4, 7, 6, 6	23
B	5, 1, 3, 5, 3, 4	21
C	8, 6, 8, 9, 5	36
		80

10.4.2 Analysis

The CF is computed as: CF $= \frac{80^2}{15} = 426.67$ and, (That is after Total SS $=$ 65.33).

The Analysis of variance Table for the data in Table 10.7 is presented in Table 10.8.

$$\text{Total SS} = 4^2 + 7^2 + \cdots + 5^2 - \frac{80^2}{15} = 65.33$$

$$\text{Treatment SS} = \frac{23^2}{4} + \frac{21^2}{6} + \frac{36^2}{5} - \frac{80^2}{15} = 38.283$$

Table 10.8 Analysis of variance table for the data in Table 10.7

Source	d.f.	SS	MS	F
Fertilizers (treatments)	2	38.28	19.14	8.49
Error	12	27.05	$2.254 = S^2$	
Total	14	65.33		

Standard error for comparing treatment A with treatment B is given by

$$\sqrt{\frac{S^2}{4} + \frac{S^2}{6}} = 0.9691$$

and B with C is given by

$$\sqrt{\frac{S^2}{6} + \frac{S^2}{5}} = 0.9091.$$

Note that these are no longer $\sqrt{\frac{2S^2}{r}}$ since $r_1 \neq r_2 \neq r_3$. The analysis is implemented in MINITAB with the following results. We have only employed Tukey's method in this example.

```
MTB > PRINT C1-C2

Data Display
 Row   FERT  YIELD

   1      A      4
   2      A      7
   3      A      6
   4      A      6
   5      B      5
   6      B      1
   7      B      3
   8      B      5
   9      B      3
  10      B      4
  11      C      8
  12      C      6
  13      C      8
  14      C      9
  15      C      5

MTB > Oneway 'YIELD' 'FERT';
SUBC>   Tukey 5.

One-way ANOVA: YIELD versus FERT
```

```
Analysis of Variance for YIELD
Source     DF       SS       MS        F        P
FERT        2    38.28    19.14     8.49    0.005
Error      12    27.05     2.25
Total      14    65.33

                                Individual 95% CIs For Mean
                                Based on Pooled StDev
Level       N     Mean    StDev ----------+---------+---------+------
A           4    5.750    1.258           (-------*-------)
B           6    3.500    1.517  (------*-----)
C           5    7.200    1.643                   (------*------)
                                ----------+---------+---------+------
Pooled StDev =    1.501                   4.0       6.0       8.0

Tukey's pairwise comparisons

    Family error rate = 0.0500
Individual error rate = 0.0206

Critical value = 3.77

Intervals for (column level mean) - (row level mean)
                A          B

       B     -0.334
              4.834

       C     -4.135     -6.124
              1.235     -1.276
```

Results from Tukey's test indicate that both A & B and A & C are not significant. These are represented in terms of the population means in the following table.

μ_B	μ_A	μ_C
3.50	5.75	7.20

However, Tukey's test also shows that B and C are significant. This represents a contradiction. As mentioned earlier, in this example, we would be at great pains to say that there are significant differences in the three fertilizers means even though the overall F value of 8.49 with a p value of 0.005 indicates significance. In this case at least, the results are inconclusive and we either go and take more replications or we could conduct some other partitioning of the treatments SS as in the next section.

10.5 Tests on Individual Treatment Means

The hypotheses of interest in the analysis of variance table are of the form for t treatments

$$H_0: \quad \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_t$$

$$H_a: \quad \text{at least two of these are unequal}$$

Suppose in conducting such an hypotheses, the null hypothesis is rejected; there are then significant differences between the treatment means, but exactly which treatments differ is not specified. In this situation, further comparisons between groups of treatment means may be useful. However, with t treatments, we can make at most $t-1$ comparisons, that is, the treatments degree of freedom in the analysis of variance. Any attempt to make more than this number of comparisons will result in the earlier problem of multiple comparisons with its attendant problems.

For example, if $t = 5$, this means that we can make at most $(5-1) = 4$ comparisons in order to arrive at our conclusion. In practice, one can make a total of $\binom{5}{2} = 10$ paired comparisons with this set of treatments. Our problem therefore is on how to choose these four comparisons in such a way that at the end of the analysis and comparisons, we would have been able to solve the problem posed above.

In Example 10.1, for instance, suppose our hypothesis is of the form

$$\begin{aligned} H_0: &\quad \mu_B = \mu_D \qquad \text{or} \qquad \mu_2 = \mu_4 \\ H_a: &\quad \mu_B \neq \mu_D \qquad \text{or} \qquad \mu_2 \neq \mu_4 \end{aligned}$$

Where A, B, C, D are represented respectively by 1–4. This hypothesis can be tested by investigating an appropriate linear combination of population means or treatment totals. Here we would prefer to use the population means, namely:

$$L_1: \quad \mu_2 - \mu_4 = 0$$

If the hypothesis had been for instance of the form:

$$\begin{aligned} H_0: &\ \mu_1 + \mu_3 = \mu_2 + \mu_4 \\ H_1: &\ \mu_1 + \mu_3 \neq \mu_2 + \mu_4 \end{aligned}$$

this implies that the linear combination is now

$$L_2: \quad \mu_1 + \mu_3 - \mu_2 - \mu_4 = 0.$$

In general, and as previously discussed in Chap. 6, the comparison of treatment means of interest will imply a linear combination of treatment means such as

$$C = \sum_{i=1}^{t} c_i \mu_i \tag{10.7}$$

with the restriction that $\sum_{i=1}^{t} c_i = 0$. Such linear combinations are called *contrasts*. In L_1 for instance, $c_1 = 1$, $c_2 = -1$, and hence $\sum c_i = 0$. Similarly, in L_2, we have, $c_1 = 1, c_2 = -1, c_3 = 1$, and $c_4 = -1$. Hence again, $\sum c_i = 0$. Thus, both L_1 and L_2 are therefore contrasts by definition.

The sum of squares (SS) for any contrast is computed as,

$$SS_C = \frac{(\sum c_i Y_{i+})^2}{r \sum c_i^2} \quad i = 1, 2, \cdots, t \tag{10.8}$$

and has a single degree of freedom. Further, if we have two contrasts with coefficients $\{c_i\}$ and $\{d_i\}$, then the contrasts are said to be *orthogonal* if

$$\sum_{i=1}^{t} c_i d_i = 0 \tag{10.9}$$

For t treatments, a set of $t-1$ orthogonal contrasts will partition the SS due to treatments into $t-1$ independent single degree of freedom components. Tests performed on orthogonal contrasts are independent.

Example 10.3

A botanist observes that the stalks of primroses are of different lengths in different habitats. He has three habitats, h_0, h_1, and h_2 and random samples of plants from each habitat is drawn, and grown in good potting loam in posts in a glass house under uniform moisture and other environmental conditions. The experiment was conducted using a one-way classification design with six replications per habitat. His results are shown in Table 10.9.

10.5.1 Analysis

Our analysis starts by computing the necessary SSs for the ANOVA table. Here, $r = 6$ and $t = 3$, therefore $rt = 18$ and the CF is calculated as CF $= \frac{1589^2}{18} = 140,273.39$. Hence,

Table 10.9 Mean length of primrose stalks from three habitats. (Source: Ridgman, Experimentation in Biology, p. 55)

Habitats		Length (mm)						Total
h_0	(dry)	83	82	98	76	66	64	469
h_1	(wet)	106	96	107	94	87	92	582
h_2	(wet)	81	93	79	98	111	76	538
Total								1598

$$\begin{aligned}\text{Total SS} &= 83^2 + 82^2 + \cdots + 76^2 - CF \\ &= 143,367 - 140,273.39 = 3093.61\end{aligned}$$

$$\begin{aligned}\text{Treatment SS} &= \frac{469^2}{6} + \frac{582^2}{6} + \frac{538^2}{6} - CF \\ &= 141,354.83 - 140,273.39 = 1081.44\end{aligned}$$

Note the division by 6 because 469, 582, and 538 each is the sum of six observations. Therefore,

$$\begin{aligned}\text{Error SS} &= \text{Total SS} - \text{Treatment SS}\\ &= 3094 - 1081.44\\ &= 2012.17\end{aligned}$$

Table 10.10 Analysis of variance table for the data in Table 10.9

Source	d.f.	SS	MS	F
Habitats	2	1081.44	540.72	4.03
Error	15	2012.17	134.14	
Total	17	3093.61		

$F_{(2,15)}$ at $\alpha = 0.05$ equals $= 3.68$. Since the computed F value of 4.03 is greater than 3.68, hence we conclude that significant differences exist between the means of the habitats.

In order to ascertain which of the habitats is best, we need to make further comparisons. Since we have only 2 d.f. for this, this means that we could make at most only two comparisons. The comparisons chosen by Ridgman are:

(i) Comparison between the two wet habitats, viz., h_1 versus h_2, that is,

$$\begin{aligned}H_0 &: \mu_1 = \mu_2\\ H_a &: \mu_1 \neq \mu_2\end{aligned}$$

and,

(ii) The mean of the dry habitat compared with the average means of the wet habits. That is, h_0 versus $h_1 + h_2$, which translates to the following hypotheses:

$$\begin{aligned}H_0 : \quad \mu_0 &= \frac{\mu_1 + \mu_2}{2}, \quad \text{that is}\\ 2\mu_0 &= \mu_1 + \mu_2 \quad \text{versus}\\ H_a : \quad \mu_0 &\neq \frac{\mu_1 + \mu_2}{2}, \quad \text{that is}\\ 2\mu_0 &\neq \mu_1 + \mu_2\end{aligned}$$

H_0 under (i) can be rewritten as $L_1 = \mu_1 - \mu_2 = 0$. Similarly that of (ii) can be written as $L_2 = 2\mu_0 - \mu_1 - \mu_2 = 0$.

These can be conveniently written in terms of the population means in the form:

	μ_0	μ_1	μ_2	$\sum c_i^2$
(i)	0	1	−1	2
(ii)	2	−1	−1	6

We note that for (i) $\sum c_i = 0 + 1 - 1 = 0$, that is, L_1 is a contrast by (10.7). Similarly for (ii) $\sum d_i = 2 - 1 - 1 = 0$, that is, L_2 is also a contrast.

Further for contrasts L_1 and L_2, we have

$$\sum c_i d_i = 0 \times 2 + 1(-1) + (-1)(-1) = 0 - 1 + 1 = 0.$$

That is, the two contrasts are orthogonal.

10.5.2 Computing Contrasts SS

To calculate the two contrasts SS, we first tabulate the totals for each treatment as in the Table below:

Treatments totals	469	582	538	$\sum c_i^2$
(i)	0	1	−1	2
(ii)	2	−1	−1	6

The partitioning of the habitat SS can be obtained using the expression in (10.8). For hypothesis (i), we have from the above table of totals,

$$\frac{\{(+1) \times 582 + (-1) \times 538\}^2}{2 \times 6} = \frac{(-44)^2}{12} = 161.33.$$

For hypothesis (ii), we also have

$$\frac{\{(+2) \times 469 + (-1) \times 582 + (-1) \times 538\}^2}{6 \times 6} = \frac{(-182)^2}{36} = 920.1$$

Since the two contrasts are orthogonal, we see that $161.33 + 920.1 = 1081.4$. This equals the original habitat SS given earlier in Table 10.10.

Alternative Calculations

The SS calculated above can alternatively be calculated as follows:

For Hypothesis (i)

$$\text{The SS} = \frac{582^2}{6} + \frac{538^2}{6} - \frac{(582+538)^2}{12} = \frac{628,168}{6} - \frac{1120^2}{12}$$
$$= 161.33$$

For Hypothesis (ii) Hypothesis (ii) has $2\mu_0 = \mu_1 + \mu_2$. We would therefore, first add the totals for habitats h_1 and h_2, to give $582 + 538 = 1120$ and we note that this total comes from 12 observations. Hence, the SS is computed as:

$$\text{The SS} = \frac{469^2}{6} + \frac{1120^2}{12} - \frac{(469+1120)^2}{18}$$
$$= \frac{219,961}{6} + \frac{1,254,400}{12} - \frac{2,524,921}{18}$$
$$= 920.11$$

Table 10.11 Revised analysis of variance table

Source	d.f.	SS	MS	F
h_0 Vs h_2 (within wet habitats)	1	161.33	161.33	1.20
h_0 Vs $h_1 + h_2$ (wet Vs. dry habitats)	1	920.11	920.11	6.87
Error	15	2012.17	134.14	
Total	17	3093.61		

which agrees with the earlier results. A new analysis of variance is given below in Table 10.11. $F_{(1,15)}$ at $\alpha = 0.05$ equals $= 4.54$. With our calculated value of 6.87 for the second contrast we can say that if the null hypothesis that there is no difference in stalk length between primroses obtained from wet and dry habitats is true, we have witnessed a very unlikely event and would prefer to believe that there is a difference. With an F value of 1.20 for the other contrast, however, we would be quite content to go on believing that within the wet area the primroses form a homogeneous population.

Yet another method for conducting the above hypothesis is to go via students' t test. For hypothesis (i), we first obtain the standard error (S.E.) for comparing the means of h_1 and h_2, viz.

$$\text{Required S.E.} = \sqrt{\frac{2S^2}{r}} = \sqrt{\frac{2 \times 134.14}{6}} = 6.683$$

Hence,

$$t = \frac{\bar{h}_1 - \bar{h}_2}{6.683} = \frac{\frac{(582)}{6} - \frac{(538)}{6}}{6.683} = \frac{97 - 89.67}{6.683} = 1.097$$

Compare with Student's t distribution t_{15} at $\alpha = 0.05 = 2.131$ (two-tailed value). Since $1.097 < 2.132$, therefore, we would fail to reject H_0. That is, there are no significant differences between the means of the two wet habitats.

For hypothesis (ii), S.E. is computed as

$$\text{S.E.} = \sqrt{\frac{S^2}{6} + \frac{S^2}{12}} = \sqrt{\frac{134.14}{6} + \frac{134.14}{12}} = 5.788$$

Since the mean of h_0 comes from 6 observations and the mean of $h_1 + h_2$ comes from 12 observations. That is,

$$\bar{h}_0 = \frac{469}{6} = 78.17$$

$$\overline{(h_0 + h_2)} = \frac{582 + 538}{12} = 93.33.$$

Hence,

$$t = \frac{93.33 - 78.17}{5.788} = 2.619$$

Since $2.619 > 2.132$ at $\alpha = 0.05$, we conclude that H_0 is not tenable and conclude that there is sufficient evidence to conclude that there are significant differences between the combined mean of the wet habitats and the mean of the dry habitat.

It is worth mentioning here that the computed values of t obtained under (i) and (ii) will, when squared equal the original F values obtained in Table 10.11. That is,

$$1.097^2 = 1.20 \quad \text{and}$$

$$2.619^2 = 6.86$$

which shows that in general $t_d = \sqrt{\mathrm{F}_{(1,d)}}$.

Of course we could have saved ourselves a lot of calculations by utilizing the capability of MINITAB to compute the two contrasts SS. We can accomplish this as follows:

(a) First code the levels of habitat based on the contrasts in (i) and (ii) and designate them as hypothesis $H1$ and $H2$, respectively.
(b) Now run an ANOVA analysis indicating $H1$ and $H2$ to be covariates.
(c) The resulting analysis, presented below would give us the similar results obtained earlier from our manual calculations—except that this is more reliably accurate.
(d) The results again indicate, based on the calculated p values that we would strongly reject $H2$.

```
MTB > Code (0) 0 (1) 1 (2) -1 'Hab' c3
MTB > Code (0) 2 (1) -1 (2) -1 'Hab' c4
MTB > GLM 'length' = H1 H2;
SUBC>   Covariates 'H1' 'H2';
SUBC>   Brief 2 .

General Linear Model: length versus

Factor  Type  Levels  Values


Analysis of Variance for length, using Adjusted SS for Tests

Source  DF  Seq SS  Adj SS  Adj MS     F      P
H1       1   161.3   161.3   161.3  1.20  0.290
H2       1   920.1   920.1   920.1  6.86  0.019
Error   15  2012.2  2012.2   134.1
Total   17  3093.6


S = 11.5821   R-Sq = 34.96%   R-Sq(adj) = 26.28%


Term        Coef  SE Coef       T      P
Constant  88.278    2.730   32.34  0.000
H1         3.667    3.343    1.10  0.290
H2        -5.056    1.930   -2.62  0.019
```

10.6 Design with a Quantitative Treatment

The single treatment factor investigated in the one-way classification analysis of variance can be either quantitative or qualitative. A quantitative factor is one whose level can be associated with points in a numerical scale, such as temperature, pressure, time, and dosage levels. Qualitative factors, on the other hand, are factors in which the levels cannot be arranged in order of magnitude.

In so far as the initial design and analysis of the experiment are concerned, both types of factors are treated identically. However, in experiments relating to quantitative factors, we are interested not only in the differences in the treatment means, but also in determining if the treatment means are functionally related to the ordered values of the factor. In general, we are interested in finding a mathematical relationship between the factor and the response. The general procedure for this problem, we will recall, is called regression analysis. However, if the levels of the factor are equally spaced, a simple procedure using orthogonal polynomial coefficients may be readily employed.

The procedure consists of computing a linear, quadratic, cubic, quartic, quintic, etc. effect and SS for the factor. Each effect has a single degree of freedom contrast and they are computed from the treatment totals at the t factor levels as in the earlier section, and the corresponding sum of squares is found from Eq. (10.8). It is possible to extract polynomial effects up through order $t-1$ if there are t factor levels used in the experiment.

Example 10.6.1

The data in Table 10.12 relate to the outcome of an experiment with four equally spaced dosages of a drug.

Table 10.12 Effects of four equally spaced dosage levels

	T_1	T_2	T_3	T_4
	0	5	10	15
	10	9	14	17
	8	13	13	15
	12	12	11	14
	11	10	12	18
	9	11	15	16
$\sum Y_i$	50	55	65	80
$\bar{Y}_i$	10	11	13	16

The dependent variable Y is a physiological measure that presumably is influenced by the amount of the drug administered. Table 10.12 gives the outcome of the experiment in which $r = 5$ subjects were assigned at random to each of the four dosages of the drugs. The initial analysis of the data is presented in Table 10.13

10.6.1 Analysis of Variance for the Experiment

$$\text{Total SS} = 10^2 + 8^2 + \cdots + 16^2 - \frac{250^2}{20} = 145.00$$

$$\text{Treatments SS} = \frac{50^2}{5} + \frac{55^2}{5} + \frac{65^2}{5} + \frac{80^2}{5} - \frac{250^2}{20} = 105.00$$

$$\text{Error SS} = \text{Total SS} - \text{Treatments SS} = 40.00$$

Table 10.13 ANOVA table for the data in Table 10.12

Source	d.f.	SS	MS	F
Treatments	3	105	35	14.0
Error	16	40	2.5	
Total	19	145		

```
MTB > set c1
DATA> 5(0:15/5)
DATA> end
MTB > set c2
DATA> 10 9 14 17 8 13 13 15
DATA> 12 12 11 14 11 10 12 18
DATA> 9 11 15 16
DATA> end
MTB > Oneway 'y' 'dosage';
SUBC>   Tukey 5.

One-way ANOVA: y versus dosage

Analysis of Variance for y
Source     DF        SS        MS        F        P
dosage      3    105.00     35.00    14.00    0.000
Error      16     40.00      2.50
Total      19    145.00

                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
Level      N      Mean     StDev  ------+---------+---------+---------+
 0         5    10.000     1.581  (-----*-----)
 5         5    11.000     1.581      (-----*-----)
10         5    13.000     1.581               (-----*-----)
15         5    16.000     1.581                          (-----*-----)
                                  ------+---------+---------+---------+
Pooled StDev =   1.581               10.0      12.5      15.0      17.5

Tukey's pairwise comparisons

    Family error rate = 0.0500
Individual error rate = 0.0113

Critical value = 4.05

Intervals for (column level mean) - (row level mean)

                0           5          10
```

```
 5      -3.864
         1.864

10      -5.864      -4.864
        -0.136       0.864

15      -8.864      -7.864      -5.864
        -3.136      -2.136      -0.136
```

The graph of Y means against the values of X-dosage levels is given in Fig. 10.1.

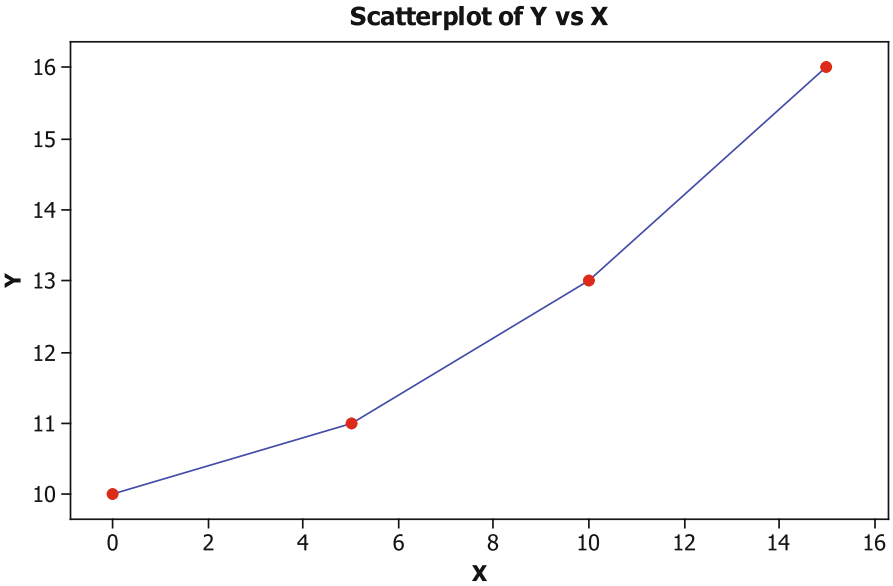


Fig. 10.1 Plot of Y means against the dosage values

With t treatments, we can fit a $(t-1)$th degree polynomial. In our example $t = 4$, hence the appropriate model will be of the form,

$$\bar{Y}_i = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \epsilon \tag{10.10}$$

10.6.2 Use of Orthogonal Polynomials

Table 6 in the Appendix gives the coefficients for orthogonal polynomials for up to $t = 10$. For our case, $t = 4$, and since we have 3 degrees of freedom for treatments, the three components are appropriately called linear, quadratic, and cubic. From this table the required coefficients are

Linear	−3	−1	1	3
Quadratic	1	−1	−1	1
Cubic	−1	3	−3	1

Note that each component forms a contrast as $\sum c_i = 0$. Further, each pairs of contrasts are orthogonal. Thus the three contrasts are said to be mutually or pairwise orthogonal. This means that the treatment SS can be partitioned into these three components. That is,

$$\text{Treatment SS} = \text{Linear SS} + \text{Quadratic SS} + \text{Cubic SS}.$$

For the above example, the treatment totals are

$$50 \qquad 55 \qquad 65 \qquad 80$$

Hence, the sum of squares is calculated as follows:

$$\text{The Linear SS} = \frac{\{(-3)\times 50 + (-1)\times 55 + (+1)\times 65 + (+3)\times 80\}^2}{5\times 20} = \frac{100^2}{100}$$

$$= 100.0$$

$$\text{Quadratic SS} = \frac{\{(+1)\times 50 + (-1)\times 55 + (-1)\times 65 + (+1)\times 80\}^2}{5\times 4} = \frac{10^2}{20}$$

$$= 5.0$$

$$\text{Cubic SS} = \frac{\{(-1)\times 50 + (+3)\times 55 + (-3)\times 65 + (+1)\times 80\}^2}{5\times 20} = \frac{0^2}{100}$$

$$= 0.$$

Of course we could have obtained the Cubic SS from the fact that

$$\text{Cubic SS} = \text{Treatment SS} - \text{Linear SS} + \text{Quadratic SS} = 105 - 105 = 0.$$

Our revised analysis of variance table is given in Table 10.13 and $F_{(1,16)}$ at $\alpha = 0.05$ is 4.49. Hence, only the linear component is significant, that is the response can be fitted by an equation of the form

$$Y_i = \beta_0 + \beta_1 X_1 \tag{10.11}$$

Table 10.14 Revised analysis of variance table

Source	d.f.	SS	MS	F
Linear	1	100	100	40
Quadratic	1	5	5	2.0
Cubic	1	0	0	0
Error	16	40	25	
Total	19	145		

rather than Eq. (10.10). The method of construction of these equations had earlier been discussed in Chap. 6.

A third degree polynomial applied to the data in Table 10.12 using MINITAB gives the following results. Again, we see that the SS due to the quadratic and cubic are not significant, indicating that a simple linear regression model will be adequate for these data. A simple regression model applied to the data gives the following estimated equation

$$\hat{y}_i = 9.5 + 0.4X_i$$

With $R^2 = 0.69$ and $X_i = (0, 15)$ in the above-estimated equation.

```
MTB > %Fitline 'y' 'dosage';
SUBC>   Poly 3;
SUBC>   Confidence 95.0.

Analysis of Variance

Source            DF         SS         MS         F      P
Regression         3        105       35.0        14  0.000
Error             16         40        2.5
Total             19        145

Source      DF     Seq SS         F       P
Linear       1        100    40.000   0.000
Quadratic    1          5     2.125   0.163
Cubic        1          0     0.000   1.000
```

Again, the MINITAB implementation of the above is presented below. Again we recode the levels of dosage based on the orthogonal coefficients and declaring the components designated L, Q, and C as covariates as in the previous example.

```
MTB > Code (0) -3 (5) -1 (10) 1 (15) 3 'dosage' c3
MTB > Code (0) 1 (5) -1 (10) -1 (15) 1 'dosage' c4
MTB > Code (0) -1 (5) 3 (10) -3 (15) 1 'dosage' c5


MTB > print c1-c5

Data Display

Row  dosage   y   L   Q   C
  1       0  10  -3   1  -1
  2       5   9  -1  -1   3
  3      10  14   1  -1  -3
  4      15  17   3   1   1
  5       0   8  -3   1  -1
  6       5  13  -1  -1   3
  7      10  13   1  -1  -3
  8      15  15   3   1   1
  9       0  12  -3   1  -1
 10       5  12  -1  -1   3
 11      10  11   1  -1  -3
 12      15  14   3   1   1
 13       0  11  -3   1  -1
 14       5  10  -1  -1   3
 15      10  12   1  -1  -3
```

```
 16      15  18   3   1   1
 17       0   9  -3   1  -1
 18       5  11  -1  -1   3
 19      10  15   1  -1  -3
 20      15  16   3   1   1

MTB > GLM 'y' = L Q 'C';
SUBC>   Covariates 'L' 'Q' 'C';
SUBC>   Brief 2 .

General Linear Model: y versus

Factor  Type  Levels  Values


Analysis of Variance for y, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS   Adj MS      F      P
L        1  100.000  100.000  100.000  40.00  0.000
Q        1    5.000    5.000    5.000   2.00  0.176
C        1    0.000    0.000    0.000   0.00  1.000
Error   16   40.000   40.000    2.500
Total   19  145.000


S = 1.58114   R-Sq = 72.41%   R-Sq(adj) = 67.24%


Ter m           Coef    SE Coef        T        P
Constant     12.5000     0.3536    35.36    0.000
L             1.0000     0.1581     6.32    0.000
Q             0.5000     0.3536     1.41    0.176
C             0.0000     0.1581     0.00    1.000



MTB > Fitline 'y' 'dosage';
SUBC>   Confidence 95.0.

Regression Analysis: y versus dosage

The regression equation is
y = 9.500 + 0.4000 dosage


S = 1.58114   R-Sq = 69.0%   R-Sq(adj) = 67.2%


Analysis of Variance

Source      DF   SS     MS      F      P
Regression   1  100  100.0  40.00  0.000
Error       18   45    2.5
Total       19  145
```

The fitted regression line is also implemented in MINITAB in the bottom part of the above MINITAB partial output.

Example 10.6.2

This example is taken from the book by D.C. Montgomery (1982),"Design and Analysis of Experiments."

The tensile strength of synthetic fiber used to make cloth for men's shirts is of interest to a manufacturer. It is suspected that strength is affected by the percentage of cotton in the fiber. Five levels of cotton percentage are of interest: 15, 20, 25, 30, and 35 %. Five observations are taken at each level of cotton percentage. Table 10.15 gives the data for this experiment.

Table 10.15 Tensile strength of synthetic fiber at five levels

Percentage of carbon	Observations 1	2	3	4	5	Total Y_i
15	7	7	15	11	9	49
20	12	17	12	18	18	77
25	14	18	18	19	19	88
30	19	25	22	19	23	108
35	7	10	11	15	11	54
Total						376

Analysis of Variance for the Experiment

$$\begin{aligned} \text{CF} &= 376^2/25 \\ \text{Total SS} &= 7^2 + 12^2 + 14^2 + \cdots + 11^2 - \text{CF} \\ &= 636.96 \\ \text{Treatment SS} &= \frac{49^2}{5} + \frac{77^2}{5} + \frac{88^2}{5} + \frac{108^2}{5} + \frac{54^2}{5} - \text{CF} \\ &= 475.76 \end{aligned}$$

The results of our analysis are presented in Table 10.16.

Table 10.16 ANOVA table for the data in Table 10.15

Source	d.f.	SS	MS	F
Treatments	4	475.75	118.94	14.76
Error	20	161.20	8.06	
Total	24	636.96		

$F_{(4,20)}$ at $\alpha = 0.05 = 2.87$. Thus we would reject H_0 and conclude that the percentage of cotton in the fiber significantly affects the strength.

In order to partition the treatment SS into four components (corresponding to the treatments degree of freedom), we may use the fact that the percentage of cotton are equally spaced. This enables us to make use of table of coefficients of orthogonal polynomials. From Table 6 in the Appendix we have for d.f. $= (5 - 1) = 4$. Those coefficients are,

Linear	−2	−1	0	1	2
Quadratic	2	−1	−2	−1	2
Cubic	−1	2	0	−2	1
Quartic	1	−4	6	−4	1

Like in the previous section, we note here too that each component forms a contrast and the four contrasts are mutually orthogonal. Since they are orthogonal, the addition of the sum of squares for the four components will equal the original treatment SS. The coefficients and treatment totals are displayed in Table 10.17

Table 10.17 Calculation of components SS

Treatment Totals	49	77	88	108	54
Linear	−2	−1	0	1	2
Quadratic	2	−1	−2	−1	2
Cubic	−1	2	0	−2	1
Quartic	1	−4	6	−4	1

$$\text{Linear SS} = \frac{\{(-2) \times 49 + (-1) \times 77 + (0) \times 88 + (+1) \times 108 + (+2) \times 54\}^2}{5 \times 10}$$

$$= 33.62$$

Similarly, the Quadratic SS $= 343.21$ and the Cubic SS $= 64.98$, while the Quartic SS $= 33.95$. The revised analysis of variance therefore is as shown in Table 10.18.

Table 10.18 Revised analysis of variance table

Source	d.f.	SS	MS	F
Percentage cotton	4	475.76	118.94	14.76
Linear	1	33.62	33.62	4.17
Quadratic	1	343.21	343.21	42.58
Cubic	1	64.98	64.98	8.06
Quartic	1	33.95	33.95	4.21
Error	20	161.20	8.06	
Total	24	636.96		

From Table 10.18, we note that $F_{(1,20)}$ at $\alpha = 0.05 = 4.35$ and therefore both the quadratic and cubic effects of cotton percentage are statistically

significant. We will therefore fit a cubic polynomial of the form

$$Y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon \tag{10.12}$$

to the data. The MINITAB implementation for the model above is presented below with the estimated regression function plotted in Fig. 10.2. The regression equation is,

$$\hat{y}_i = 62.6114 - 9.0114x + 0.4814x^2 - 0.0076x^3$$

```
MTB > set c1
DATA> 5(15:35/5)
DATA> end
MTB > set c2
DATA> 7 12 14 19 7 7 17 18 25 10
DATA> 15 12 18 22 11 11 18 19 19 15
DATA> 9 18 19 23 11
DATA> end
MTB > Name c3 = 'COEF1'

MTB > %Fitline 'y' 'x';
SUBC>   Poly 3;
SUBC>   Confidence 95.0;
SUBC>   Coef  'COEF1'.

Polynomial Regression Analysis: y versus x


The regression equation is
y = 62.6114 - 9.01143 x
 + 0.481429 x**2 - 0.0076 x**3

S = 3.04839      R-Sq = 69.4 %      R-Sq(adj) = 65.0 %

Analysis of Variance

Source            DF         SS         MS         F      P
Regression         3    441.814    147.271   15.8482  0.000
Error             21    195.146      9.293
Total             24    636.960


Source      DF     Seq SS          F      P
Linear       1     33.620     1.2816  0.269
Quadratic    1    343.214    29.0272  0.000
Cubic        1     64.980     6.9926  0.015
```

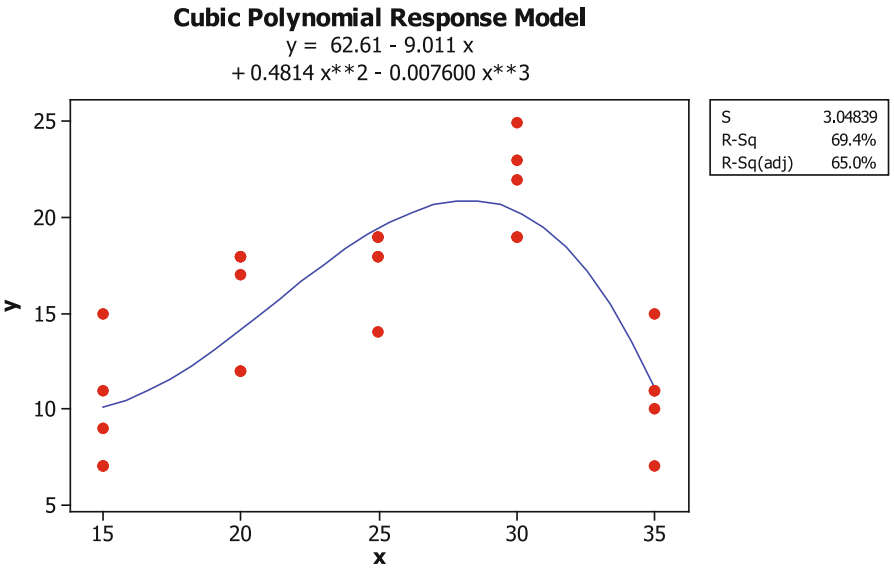


Fig. 10.2 The plotted cubic polynomial to the data in Table 10.15

A MINITAB computation of the components SS is again accomplished by first coding the levels of 'pct' the percentage of fiber and then run the appropriate ANOVA model, declaring the components as covariates. The following partial results are obtained.

```
Results for: fiber.MTW

MTB > Code (15) -2 (20) -1 (25) 0 (30) 1 (35) 2 'pct' c3
MTB > Code (15) 2 (20) -1 (25) -2 (30) -1 (35) 2 'pct' c4
MTB > Code (15) -1 (20) 2 (25) 0 (30) -2 (35) 1 'pct' c5
MTB > Code (15) 1 (20) -4 (25) 6 (30) -4 (35) 1 'pct' c6

MTB > print c1-c6

Data Display

Row  pct   y   L   Q   C  Qt
  1   15   7  -2   2  -1   1
  2   20  12  -1  -1   2  -4
  3   25  14   0  -2   0   6
  4   30  19   1  -1  -2  -4
  5   35   7   2   2   1   1
  6   15   7  -2   2  -1   1
  7   20  17  -1  -1   2  -4
  8   25  18   0  -2   0   6
  9   30  25   1  -1  -2  -4
 10   35  10   2   2   1   1
 11   15  15  -2   2  -1   1
 12   20  12  -1  -1   2  -4
 13   25  18   0  -2   0   6
 14   30  22   1  -1  -2  -4
 15   35  11   2   2   1   1
 16   15  11  -2   2  -1   1
 17   20  18  -1  -1   2  -4
```

```
 18   25  19   0  -2   0   6
 19   30  19   1  -1  -2  -4
 20   35  15   2   2   1   1
 21   15   9  -2   2  -1   1
 22   20  18  -1  -1   2  -4
 23   25  19   0  -2   0   6
 24   30  23   1  -1  -2  -4
 25   35  11   2   2   1   1

MTB > GLM 'y' = L Q 'C' Qt;
SUBC>   Covariates 'L' 'Q' 'C' 'Qt';
SUBC>   Brief 2 .

General Linear Model: y versus

Factor  Type  Levels  Values


Analysis of Variance for y, using Adjusted SS for Tests

Source  DF  Seq SS  Adj SS  Adj MS      F      P
L        1   33.62   33.62   33.62   4.17  0.055
Q        1  343.21  343.21  343.21  42.58  0.000
C        1   64.98   64.98   64.98   8.06  0.010
Qt       1   33.95   33.95   33.95   4.21  0.053
Error   20  161.20  161.20    8.06
Total   24  636.96


MTB > Fitline 'y' 'pct';
SUBC>   Poly 3;
SUBC>   Confidence 95.0.

Polynomial Regression Analysis: y versus pct

The regression equation is
y = 62.61 - 9.011 pct + 0.4814 pct**2 - 0.007600 pct**3
```

10.7 Model Adequacy Checking

The following are some of the assumptions underlying the analysis of variance for the CRD.

(i) The observations in the experiment are adequately represented by the model in (10.1), that is,

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

(ii) The errors, ϵ_{ij}, are normally and independently distributed.
(iii) These errors all have a constant variance σ^2.

Generally, we can never be sure if all these assumptions are satisfied when we handle data daily and slight departures from these assumptions are of little concern. The assumption of homogeneity of variance ceases to be valid if some treatments are erratic in their effects or if data follow a non-normal, skewed distribution, as for instance, in the skewed distributions, the variance tends to be a function of the mean. We would expect that inferences not be

made until these assumptions have been validated. Violations of the above assumptions can be investigated by investigating the residuals, which we recall from Chap. 7 are defined as:

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

where $\hat{y}_{ij}$ is the fitted values under the model and y_{ij} is the observed data for treatment i in replicate j.

For the analysis of the data in Table 10.4, the normality test 10.3 using the Anderson–Darling test gives a p value of 0.696 and is based on the following hypotheses,

$$H_0 : \text{The } e_{ij} \text{ are normally distributed}$$
$$H_a : \text{The } e_{ij} \text{ are not normally distributed}$$

The p value of 0.696 indicates that we would fail to reject H_0, hence the normality assumption is satisfied (Fig. 10.3).

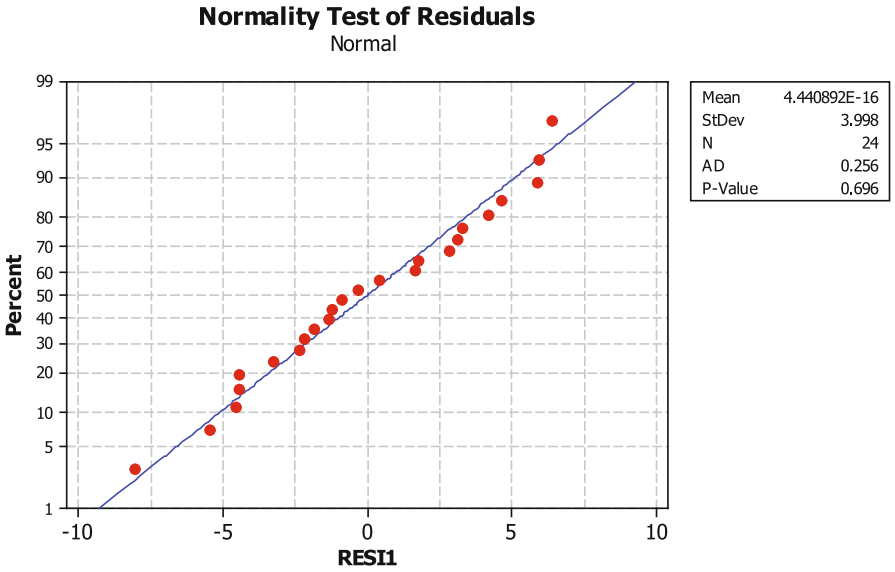


Fig. 10.3 Normality test and plot for the residuals for the data in Table 10.4

The test of homogeneity of variances of the residuals for each treatments (varieties) is conducted with Bartlett's test of homogeneity of variances discussed earlier in Chap. 6. This test is sometimes referred to as *Bartlett's Test of Homogeneity of Variances*. The implication of the constant variance assumption under the assumptions above is that the treatments all came from the same population. In essence what this means is that the treatments are assumed to have equal variances. If this were not so, then all the inferences,

t tests etc., are invalid. It is therefore important to check whether this assumption of equality of variances or homogeneity is invalidated, and in order to do this, we use Bartlett's test of homogeneity. We are interested in testing the hypotheses:

$$\begin{aligned} \mathrm{H}_0 &: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 \\ \mathrm{H}_a &: \sigma_1^2 \neq \sigma_2^2 \neq \cdots \neq \sigma_t^2 \end{aligned}$$

where t is the number of treatments. To implement Bartlett's test, we perform the following calculations for the data in Table 10.4.

$$\chi_0^2 = \log 10 \, \frac{q}{c} = 2.3026 \, \frac{q}{c}$$

where

$$\begin{aligned} q &= (N-t)\log S_p^2 - \sum_{i=1}^{t}(n_i - 1)\log s_i^2 \\ c &= 1 + \frac{1}{3(t-1)}\left[\sum (n_i - 1)^{-1} - (N-t)^{-1}\right] \\ S_p^2 &= \frac{1}{N-t}\sum_{i=1}^{t}(n_i - 1)s_i^2. \end{aligned}$$

$N = \sum_{i=1}^{t} n_i$ and s_i^2 is the sample variance of the ith treatment. We would therefore reject H_0, whenever $\chi_0^2 > \chi_{\alpha,t-1}^2$.

10.7.1 Example 10.62

In example 10.1, we have four treatments: A, B, C, and D. From Table 10.4, we have $s_A^2 = 22.04$, $s_B^2 = 15.61$, $s_C^2 = 30.91$, and $s_D^2 = 4.97$.

Here $t = 4$, $n_1 = n_2 = n_3 = n_4 = 6$ (number of replications). Hence, $N = \sum n_i = 4 \times 6 = 24$ and therefore,

$$\begin{aligned} S_p^2 &= \frac{(6-1)s_1^2 + (6-1)s_2^2 + (6-1)s_3^2 + (6-1)s_4^2}{24-4} \\ &= \frac{5(s_1^2 + s_2^2 + s_3^2 + s_4^2)}{20} \\ &= \frac{5 \times 73.53}{20} \\ &= 18.38. \end{aligned}$$

Note that the value of $S_P^2 = 18.33$ is the same as the error mean square value in Table 10.5. In general, $S_P^2 =$ error mean square. Hence,

$$\begin{aligned} q &= (24-4)\log 18.38 - (5\log 22.04 + 5\log 15.61 + 5\log 30.91 + 5\log 4.97) \\ &= 58.225 - 5(10.87527) \\ &= 3.849 \end{aligned}$$

That is, $q = 3.849$.

$$\begin{aligned} c &= 1 + \frac{1}{3\times(4-1)}\left[\left(\frac{1}{5}+\frac{1}{5}+\frac{1}{5}+\frac{1}{5}\right) - \frac{1}{20}\right] \\ &= 1 + \frac{1}{9}\left(\frac{4}{5} - \frac{1}{20}\right) \\ &= 1.083. \end{aligned}$$

Hence,

$$X_0^2 = 2.3026\left(\frac{3.849}{1.083}\right) = 3.5540.$$

But, $\chi^2_{0.01,3} = 11.34$, and since $3.5540 \ll 11.34$, we would therefore fail to reject H_0 and conclude that indeed, the treatment populations all have the same variance at the 1 % significance point. Bartlett's test is implemented in MINITAB with the following statements and modified output.

```
MTB > Name c3 = 'STDV1' c4 = 'VARS1'
MTB > %Vartest 'YIELD' 'VARIETY';
SUBC>   Confidence 95.0;
SUBC>   Stdevs 'STDV1';
SUBC>   Variances 'VARS1'.
Executing from file: C:\PROGRAM FILES\MTBWIN\MACROS\Vartest.MAC

Test for Equal Variances

Response    YIELD
Factors     VARIETY
ConfLvl     95.0000

Bonferroni confidence intervals for standard deviations

  Lower     Sigma     Upper     N  Factor Levels

 2.60661   4.69442   15.5997    6  A
 2.19354   3.95050   13.1276    6  B
 3.08726   5.56005   18.4762    6  C
 1.23816   2.22989    7.4100    6  D

Bartlett's Test (normal distribution)

Test Statistic: 3.554
P-Value       : 0.314
```

```
Levene's Test (any continuous distribution)

Test Statistic: 2.090
P-Value       : 0.134

Test for Equal Variances: YIELD vs VARIETY
```

The calculated statistic under Bartlett's test agrees exactly with our result of 3.554. The p value indicates that we would fail to reject H_0. It should be pointed out here that Bartlett's test is often used when it is assumed that the errors follow a normal distribution. When this is not the case, Levene's test is most appropriate.

10.8 Exercises

1. The results of trials of three varieties of gooseberry bush in eight regions are as follows:

Region	1	2	3	4	5	6	7	8
Control variety	21	24	20	14	15	18	17	28
New variety A	24	28	17	18	25	19	25	28
New variety B	22	28	21	16	17	17	19	30

   The two purposes of the trial were:

   (i) To compare the two new varieties.
   (ii) To examine the average differences between the control variety and the new varieties.

   Complete the analysis of variance below, explaining how the calculations already made are obtained and write a report of your conclusion, giving S.E.s and confidence intervals when appropriate.

Source	SS	d.f.	MS	F
Regions	400	7		3.98
Varieties	46	2	23	
Error	81	14	5.79	
Total	527	23		
CF	1088			

2. The amount of carbon used in the manufacture of steel is assumed to have an effect on the tensile strength of the steel. Given the following data, perform the appropriate analysis and interpret your results. The tensile strengths of six specimens of steel for each of three different percentages of carbon are shown (The data have been coded for easy calculation).

Percentage of carbon		
0.10	0.20	0.30
23	42	47
36	26	43
31	47	43
33	34	39
31	37	42
31	31	35

3. It is suspected that five filling machines in a certain plant are fillings cans to different levels. Random samples of the production from each machine were taken with the following results. Analyse the data and state your conclusions.

Machine				
A	B	C	D	E
11.95	12.18	12.16	12.25	12.10
12.00	12.11	12.15	12.30	12.04
12.25		12.08	12.10	12.02
12.10				12.01

4. The data below relate to the activated lives of 20 batteries. There are four treatments investigated in the experiment. Carry out an analysis of variance and use either Duncan's, Scheffé, or Tukey tests to draw your conclusions.

Activated lives of twenty thermal batteries resulting from experiment

	Treatment				
	1	2	3	4	Total
Observations (sec)	73	74	68	71	
	73	74	69	71	
	73	74	69	72	
	75	74	69	72	
	75	75	70	73	
Total	369	371	345	359	1444
Number of observations	5	5	5	5	20
Mean	73.5	74.2	69.0	71.8	72.2

5. Consider the data below. Carry out a full analysis of variance and partition your treatments SS into linear, quadratic, and cubic models. What would be the appropriate response model?

Obs.	Treatment levels			
	10 lb/ plot	20 lb/ plot	30 lb/ plot	40 lb/ plot
1	25	36	35	43
2	29	37	39	40
3	31	29	31	36
4	30	40	42	48
5	27	33	44	47

6. A biologist wished to study the effects of ethanol on sleep time. A sample of 20 rats matched for age and other characteristics, was selected, and each rat was given an oral injection having a particular concentration of ethanol per body weight. The rapid eye movement (REM) sleep time for each rat was then recorded for a 24-h period. The data are presented below (Source: Devore and Peck 2001).

Treatment levels	Observations on rats				
	1	2	3	4	5
0 (control)	88.6	73.2	91.4	68.0	75.2
1 g/kg	63.0	53.9	69.2	50.1	71.5
2 g/kg	44.9	59.5	40.2	56.3	38.7
4 g/kg	31.0	39.6	45.3	25.2	22.7

Carry out a full analysis of variance and partition your treatments SS into the following components (i) control vs. others (ii) between others. Also conduct a multiple comparison procedure on the means of the treatment levels.

7. Nineteen pigs were divided into four groups, and each group was given a different feed. The data are weights, in kilograms. The data are presented below (Source: Zar 2000).

Feed1	Feed2	Feed3	Feed4
60.8	68.7	102.6	87.9
57.0	67.7	102.1	84.2
65.0	74.0	100.2	83.1
58.6	66.3	96.5	85.7
61.7	69.8		90.3

Carry out an analysis of variance on the above data and test the necessary assumptions. Which feed would you recommend?

8. The MINITAB printout for an experiment utilizing a completely randomized design is shown below:

Analysis of variance table				
Source	d.f.	SS	MS	F
Factor	3	57,258	19,086	14.80
Error	34	43,836	1289	
Total	37	101,094		

a. How many treatments are involved in the experiment? What is the total sample size?
b. State the null and alternative hypotheses for this experiment and conduct a test of the null hypothesis that the treatment means are equal. Use $\alpha = .01$.
c. What assumptions must be satisfied before the analysis above can be valid? (State these only).
d. Are the treatments equally replicated?

9. The partially completed ANOVA table given here is for a one-way experiment:

Source	d.f.	SS	MS	F
Treatments	–	2193.442	548.3605	–
Error	–	–	–	
Total	29	2437.572		

a. Give the number of levels for the treatments.
b. How many observations were collected for each treatment level?
c. Complete the ANOVA table.
d. Test to determine whether the treatment means differ. Use $\alpha = .10$.

10. The partially completed ANOVA table for an experiment is shown below:

Source	d.f.	SS	MS	F
Treatments	–	2236.44	–	–
Error	11	–	546.96	
Total	14	–		

(a) Complete the above table. What design was employed?
(b) How many treatments are involved in the experiment? What is the total sample size?
(c) Conduct a test of the null hypothesis that the treatment means are equal. Use $\alpha = .05$.
(d) Are the treatments equally replicated in this experiment (with reasons)?

11. The following data from Steel and Torrie (1960) give the results of effects of inbreeding on plant weight in red clover. The results are the average weight in grams of non-inbred (F_1) lines and three groups of inbred families arranged in increasing order of inbreeding.

```
F1:                254 263 266 249 337 277 289 244 265
Slightly inbred:   236 191 209 252 212 224
F2:                253 192 141 160 229 221 150 215 232 234 193 188
F3:                173 164 183 138 146 125 178 199 170 177 172 198
-------------------------------------------------------------------
```

Carry out the analysis of variance for the data and draw your conclusions.

12. Four strains of rats were selectively bred for differences in blood pressure in order to determine the possible effect of heredity on blood pressure A, B, C, and D are given in the table below:

Strains			
A	B	C	D
84	87	89	89
82	84	94	86
86	84	92	88
89	92	91	93
85	88	92	85
85	89	91	85
92	92	95	89
80	89	89	90
79	87	87	90
83	88	91	93

(i) Analyze the above data for statistical differences among the means.
(ii) If a statistically significant F value is found, then test for statistically significant differences among all possible pairs of means.
(iii) Test for the underlying assumptions for your analysis to be valid.
(iv) From a biological point of view, which strain(s) would you recommend?

Chapter 11
The Randomized Complete Block Design

11.1 Introduction

If it is possible to group the experimental material or conditions in a manner such that the variation among experimental units within a group is less than the variation would have without grouping, this should be done in order to compare treatments on the less variable material or under less variable conditions.

Example 11.1.1

The first illustrative example in the introduction of Chap. 10 illustrates this point. Suppose, the twenty rats in that example were selected differently. Suppose, 20 rats are selected, four from each of five litters ensuring that we have four rats each from the same mother. Again, suppose that the four nutritional treatments were A, B, C, and D. The four rats in each litter would be randomly allocated to a treatment. One possible arrangement could be:

	Rats and treatment numbers							
Litter no	Rat no	Trt no	Rat no	Trt no	Rat no	Trt no	Rat no	Trt no
1	1	*B*	2	*A*	3	*D*	4	*C*
2	5	*B*	6	*C*	7	*A*	8	*D*
3	9	*C*	10	*A*	11	*B*	12	*D*
4	13	*A*	14	*B*	15	*D*	16	*C*
5	17	*D*	18	*C*	19	*A*	20	*B*

In the above arrangement, the rats have been stratified into five groups (litters). The rats could be housed in different cages which will be kept under nearly identical conditions or rats in each litter could be exposed to the same environment, resulting in possibly five different environments. Then, the observed variation among the five groups is composed of variation among litters + variation among environments. However, as far as the treatments

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_11,

are concerned they are compared within a group, and the variation among treatment means is less than it would have been without grouping by litter + environment.

Example 11.1.2

Suppose, for the second example, the interest is in the effectiveness of nine different herbicides in eliminating dandelions from home lawns, and that eight different lawns have been selected for the investigation; each of the eight lawns are relatively uniform in topography, grass cover, and dandelion infestation. Each lawn forms a relatively uniform *block* of land. (It was for situations like this that the randomized complete block design was first described by Sir Ronald A Fisher, and hence, the name). Each of the eight blocks, or lawns, are divided into nine experimental units each of which are as alike as possible. Then, the nine treatments are randomly allocated to the nine experimental units in each of the eight blocks or lawns. With the numbers $1, 2, \cdots, 9$ representing the treatments, one possible arrangement could be:

Block 1									Block 2								
1	9	2	7	6	8	3	4	5	6	2	8	1	3	4	9	5	7

Block 3									Block 4								
3	5	9	1	6	8	2	4	7	7	5	2	4	6	8	1	9	3

Block 5									Block 6								
8	1	2	6	3	7	5	9	4	9	3	6	4	7	1	5	8	2

Block 7									Block 8								
3	7	9	6	1	4	5	2	8	4	1	9	2	7	5	3	8	6

The characteristic to be measured is number of dandelions in the *plot* or experimental unit at monthly intervals up to one year after spraying with the herbicide. In the above experiment, there could be considerable variation in dandelion count among the eight lawns, but this would not affect the differences between treatments, since all treatments are compared with each other on each of the 8 lawns.

If the treatments had been randomly allocated to the $8 \times 9 = 72$ experimental units, or plots, and if the lawns differed in dandelion count, all eight plots of some treatments could by chance have been allocated to lawns with low dandelion count, and other treatments to lawns with a high dandelion count. The comparison between treatments would then be mixed up with differences between lawns. In the randomized block design, each of the nine treatments appears on each of the eight lawns. Given that the lawns differ in dandelion count this arrangements makes the differences between mean less variable than if there had been no stratification.

The count at a given time t ($t = 0, 1, 2, \cdots, 12$ months) may be expressed as the sum of the block and treatments, and an error term minus the overall mean, thus:

$$\text{Count} = \text{Block mean} + \text{Treatment mean} - \text{Overall mean} + \text{Error}.$$

The sum of all counts for two given treatments, say 1 and 2, in the above described experiment is:

$$\begin{aligned}&\text{trt. one sum} + \text{sum of 8 block means} - 8\ (\text{overall mean}) + 8 \text{ error terms}\\ &\text{trt. two sum} + \text{sum of 8 block means} - 8\ (\text{overall mean}) + 8 \text{ other error terms.}\end{aligned}$$

The difference of two treatment means is therefore:

$$\frac{1}{2}(\text{treatment one sum} - \text{treatment two sum} + 8 \text{ error terms} - 8 \text{ other error terms})$$

This equals:

$$\bar{Y}_{1+} - \bar{Y}_{2+} + \frac{1}{8}\sum_{j=1}^{8}(e_{1j} - e_{2j}).$$

Here, we may note that the effects of the overall mean, and of the block means do not appear in the difference between two treatment means. Since we are comparing herbicide treatments for effectiveness of dandelion control the differences between means are the statistics of interest. All designs having this property for all treatments in the experiment are known as *orthogonal designs*.

In another form, we may write the observation or count as:

$$\bar{Y} + (\bar{Y}_{i+} - \bar{Y}) + (\bar{Y}_{+j} - \bar{Y}) + \text{error}$$

where, the treatment mean minus overall mean is defined to be the treatment effect, and the block mean minus the overall mean is defined to be the block effect. Note that the observation is assumed to be the sum of four terms. This need not be the case, as some observations may be the product of these four terms rather than the sum. The appropriateness of the assumption of additive effects must be questioned for every type of experiment. If the observation is the product of terms instead of the sum, one could use another function of the observations to obtain additive effects. In this case, one could transform the observation to log of observation. Thus, if

$$Y = \text{abcd}, \quad \text{then}$$
$$\log Y = \log a + \log b + \log c + \log d.$$

One might wonder why additivity of effects is desirable. The answer is simply that life is easier, i.e., computations are simpler, on the additive scale.

In field and laboratory experimentation on biological material the randomized complete block design is probably the most frequently used experimental design. Ease of construction, lay-out, and analysis of result contributes heavily to its frequent use. Also, for the above type of experiment it has been found to be considerably more efficient than the completely randomized design. Summarization of several hundred experiments over a period of years for field experiments indicates that nine blocks or replicates of a randomized complete block design are approximately equivalent to ten replicates of a completely randomized design in attaining the same degree of variability among treatment means. For this kind of experimentation the blocking, or stratification into blocks almost halves the variability among treatment means.The value of blocking material is dependent upon the type of experimental material under consideration. Each type of experimentation requires evaluation. One can always block as a form of insurance against heterogeneity, but overstratification results in some disadvantages which will be discussed later. As a rule, one should use the minimum blocking to control the heterogeneity, or the suspected heterogeneity present in the experimental material.

11.1.1 Why RCBD?

Blocking is an experimental technique to control the variability of the experimental material. On the fields however, variability takes different form: either soil heterogeneity or sloping. Thus before blocking is considered, an appropriate and effective blocking technique must be designed. Gomez and Gomez (1984) have described most appropriately, what must be done to arrive at this decision. These are:

- The recognition of the sources of variability to be used for blocking
- The selection of the block shape and orientation

Gomez and Gomez further stated that: 'an ideal source of variation to use as the basis for blocking is one that is large, and highly predicate.' They give the following examples:

(i) Soil heterogeneity, in a fertilizer or variety trial where yield data is the primary character of interest
(ii) Direction of insect migration in an insecticide trial where insect infestation is the primary character of interest
(iii) Slope of the field in a study of plant reaction to water stress.

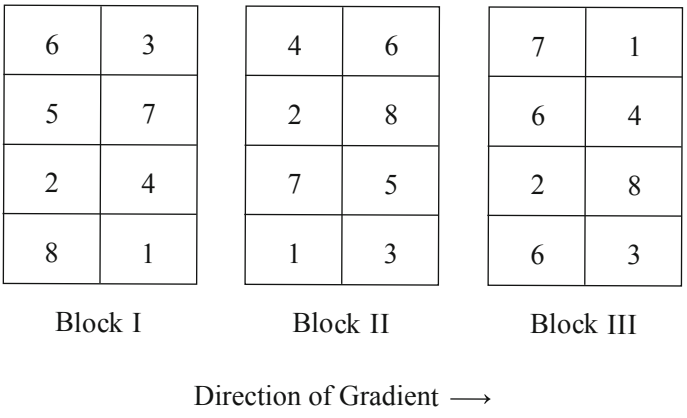
Once the sources of blocking are identified, the next thing will be to consider the appropriate size, and shape of the block to maximize the variability among the blocks. Gomez and Gomez gave the following guidelines which are reproduced here with the permission of the authors.

1. When the gradient is uni-directional (i.e., there is only one gradient), use long and narrow blocks, and orient these blocks so that their length is perpendicular to the direction of the gradient.
2. When fertility gradients occur in two directions with one gradient much stronger than the other one, ignore the weaker gradient and follow the preceding guideline for the case of unidirectional gradient.
3. When the fertility gradient occurs in both directions with both gradients equally strong and perpendicular to each other, choose one of the following alternatives:

   - Use blocks that are as square as possible.
   - Use long and narrow blocks with their length perpendicular to the direction of one gradient (as in (1) above) and use the covariance analysis technique in Chap. 13 to take care of the other gradient.
   - Use the Latin square design discussed in the next chapter to take care of variability in both directions.

4. When the pattern of variability is not predictable, blocks should be as square as possible.

As an example, suppose we wish to conduct a variety trial, but the gradient of the field is from left to right as indicated in the following figure. If we have eight treatments and we wish to employ three replicates or blocks, then blocks are laid out so that they are perpendicular to the direction of the gradient. If we number the treatments 1 to 8, then we can readily randomize these treatments within blocks as shown below. Note that we generated a random digit from an integer distribution [1, 8]. We generated 30 random numbers for each block. We then go down each replicate until we have come across all the treatments 1–8. We will skip a random number if it is already used. In this way, we can randomly allocate all the treatments to each block.



```
MTB > Base 10000.
MTB > Random 30 c1 c2 c3;
SUBC>   Integer 1 8.
MTB > print c1-c3

Data Display

           REP  REP
Row  REP I   II  III
  1      6    4    7
  2      3    6    1
  3      6    4    6
  4      5    2    4
  5      6    8    1
  6      6    7    1
  7      7    5    7
  8      2    8    2
  9      4    1    4
 10      8    8    8
 11      1    3    7
 12      7    3    6
 13      3    1    2
 14      6    2    1
 15      1    4    8
 16      7    8    3
 17      3    7    5
 18      2    2    6
 19      6    1    5
 20      7    8    3
 21      7    5    3
 22      3    3    7
 23      5    3    4
 24      7    4    6
 25      3    1    8
 26      7    3    5
 27      1    6    5
 28      5    1    5
 29      1    2    4
 30      7    2    8
```

On the other hand, if variability is from North to South, then the blocks will be laid out as in the following—being perpendicular to the direction of the gradient.

6	5	2	8
3	7	4	1

Block I

6	5	2	8
3	7	4	1

Block II

6	5	2	8
3	7	4	1

Block III

Example 11.1.3

As an illustration of the above consideration, suppose that one was interested in only three herbicides instead of nine, and that the size of the experimental unit was fixed, i.e., the lawns were divided into nine plots or experimental units instead of three. One could use blocks of size three and have three blocks per lawn.

Block 1			Block 2			Block 3			
1	2	3	4	5	6	7	8	9	Lawn I

Block 4			Block 5			Block 6			
1	4	7	2	5	8	3	6	9	Lawn II

Block 7			Block 8			Block 9			
1	5	9	7	2	6	4	8	3	Lawn III

Block 10			Block 11			Block 12			
1	8	6	4	2	9	7	5	3	Lawn IV

In the above layout for instance, the lawns have been stratified into three blocks of size three each. The nine treatments are then allocated according to the above layout. We notice that each treatment is replicated four times in the whole experiment, there are 12 blocks each of $k = 3$ plots. Each pair of treatments occur together only once in the entire layout. Further, the lawns each constitute a single replicate of the experiment. The above design is called a *balanced incomplete block design* and we shall discuss this design further in Chap. 16.

However, if the nine experimental units in each lawn were relatively homogeneous, one could use a completely randomized design of three treatments, and three replicates on each treatment for each lawn. This would result in minimum blocking which would control the lawn to lawn variability. It should be pointed out, however, that we will probably divide the lawn into thirds and have larger experimental units.

Example 11.1.4

To illustrate another variation of the randomized complete block design, suppose that only five herbicides were of interest, four of these (1, 2, 3, and 4) were of more interest than the fifth one (no. 5), and suppose that nine experimental units were available on each lawn. Treatments 1, 2, 3, and 4 could be included twice on each lawn and treatment 5 could be put in once. If we let number 1 and 6 be the plots for treatment 1, numbers 2 and 7 be the plots for treatment 2, numbers 3 and 8 be the plots for treatment 3, numbers 4 and 9 be the plots for treatment 4, and number 5 be the plot for treatment 5 in the original design, then the arrangements in the first three blocks or lawns would be:

Block 1								
1	4	2	2	1	3	3	4	5

Block 2								
1	2	3	1	3	4	4	5	2

Block 3								
3	5	4	1	1	3	2	4	2

Both of the above two variations on the randomized complete block design are orthogonal designs i.e., differences between treatment means do not involve the block effects. As long as the orthogonality of block and treatment effects is a property of the design, the analysis remains simple.

11.1.2 Model and Analysis for the RCBD

From the discussion in Sect. 11.1, it is clear that a randomized complete block design (RCBD) is one in which there are t treatments per block, and the treatments are randomized within each of the blocks. Each block contains the same number of experimental units which are assumed to be homogeneous. Usually, homogeneity within each block is achieved through matching (e.g., by age group among human subjects, litters among animals, neighboring plots of land, or sets of similar trees etc).

The statistical model for this design is:

$$Y_{ij} = \mu + t_i + b_j + e_{ij} \quad i = 1, 2, \cdots, t, \ j = 1, 2, \cdots, b \tag{11.1}$$

where, i is the general mean, t_i is the effect of the i-th treatment, b_j is the effect of the j-th block and e_{ij} is the random error term. Both treatments and blocks are viewed as being fixed factors.

A typical table of observations for a randomized block design experiment is given in Table 11.11.

Table 11.1 Typical table of observations

	Blocks					
Treatments	1	2	3	$\cdots$	b	Total
1	Y_{11}	Y_{12}	Y_{13}	$\cdots$	Y_{1b}	Y_{1+}
2	Y_{21}	Y_{22}	Y_{23}	$\cdots$	Y_{2b}	Y_{2+}
3	Y_{31}	Y_{32}	Y_{33}	$\cdots$	Y_{3b}	Y_{3+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
t	Y_{t1}	Y_{t2}	Y_{t3}	$\cdots$	Y_{tb}	Y_{t+}
Total	Y_{+1}	Y_{+2}	Y_{+3}	$\cdots$	Y_{+b}	$Y_{++} = G$

11.1.3 Analysis

There are a total of bt experimental units (plots) in this experiment. Each treatment is said to be replicated b times.

$$\text{Correction factor C.F} = \frac{Y_{++}^2}{bt} = \frac{G^2}{bt}$$

$$\text{Total SS} = Y_{11}^2 + Y_{12}^2 + \cdots + Y_{bt}^2 - \text{CF} \qquad \text{based on (bt-1) d.f.}$$

$$\text{Blocks SS} = \frac{Y_{+1}^2}{t} + \frac{Y_{+2}^2}{t} + \cdots + \frac{Y_{+b}^2}{t} - \text{CF} \qquad \text{based on (b-1) d.f.}$$

$$\text{Treatments SS} = \frac{Y_{1+}^2}{b} + \frac{Y_{2+}^2}{b} + \cdots + \frac{Y_{t+}^2}{b} - \text{CF} \qquad \text{based on (t-1) d.f.}$$

The Error SS is obtained by subtraction as:

$$\text{Error SS (SSE)} = \text{Total SS} - \text{Treatments SS}$$

The Error SS is based on $(bt-1)-(b-1)-(t-1) = (b-1)(t-1)$ degrees of freedom. The structure of the analysis of variance table is presented in Table 11.2.

Table 11.2 Analysis of variance for a randomized block design

Source	d.f.	SS	MS	F
Blocks	$b-1$	$\frac{\sum Y_{+j}^2}{t} - \frac{Y_{++}^2}{bt}$	$\frac{\text{SS Blocks}}{b-1}$	
Treatments	$t-1$	$\frac{\sum Y_{i+}^2}{b} - \frac{Y_{++}^2}{bt}$	$\frac{\text{SS Trt.}}{t-1} = A$	A/S^2
Error	$(b-1)(t-1)$	SSE	$\frac{\text{SSE}}{(b-1)(t-1)} = S^2$	
Total	$bt-1$	$\sum_i \sum_j Y_{ij}^2 - \frac{Y_{++}^2}{bt}$		

Example 11.2.1

The data in Table 11.3 relates to a trial of four strains of Galliapolli Wheat (Snedecor Page 16) laid out in five randomized blocks of four treatments each.

Table 11.3 Yields of wheat in (lb/plot)

Strains	Blocks 1	2	3	4	5	Strain total
A	32.3	34.0	34.3	35.0	36.5	172.1
B	33.3	33.0	36.3	36.8	34.5	173.9
C	30.8	34.3	35.3	32.3	35.8	168.5
D	29.3	26.0	29.8	28.0	28.8	141.9
Total	125.7	127.3	135.7	132.1	135.6	656.4

$$\text{C.F.} = \frac{656.4^2}{20} = 21543.048$$

$$\text{Total SS} = 32.3^2 + 33.3^2 + \cdots + 28.8^2 - \text{CF} = 182.17$$

$$\text{Blocks SS} = \frac{125.7^2}{4} + \frac{127.3^2}{4} + \cdots + \frac{135.6^2}{4} - \text{CF} = 21.46$$

$$\text{Treatments SS} = \frac{172.1^2}{5} + \frac{173.9^2}{5} + \cdots + \frac{141.9^2}{5} - \text{CF} = 134.45$$

Hence the Error SS = Total SS−Blocks SS−Treatments SS $= 182.17 - 21.46 - 134.45 = 26.26$ on $(5 - 1)(4 - 1) = 12$ degrees of freedom. The analysis of variance table for the data in Table 11.3 is presented in Table 11.4.

$F_{(3,12)}$ at $\alpha = 0.05 = 3.49$. Since $20.48 > 3.49$, we thus conclude that there is strong evidence that there are significant differences between the yielding abilities of the four strains. Standard error for comparing any two treatment means

$$\text{S.E} = \sqrt{\frac{2S^2}{b}} = \sqrt{\frac{2 \times 2.188}{5}} = 0.936$$

Notice that the number of blocks in the experiment correspond to the number of replicates of the treatments, viz., five in this study. The above analysis is implemented in MINITAB with the following together with a corresponding output.

Table 11.4 Analysis of variance table for the experiment

Source	d.f.	SS	MS	F
Blocks	4	21.46		
Treatments	3	134.45	44.817	20.48
Error	12	26.26	2.188	
Total	19	182.17		

```
MTB > SET C2
DATA> 4(1:5)
DATA> END
MTB > SET C3
DATA> 32.3 34.0 34.3 35.0 36.5
DATA> 33.3 33.0 36.3 36.8 34.5
DATA> 30.8 34.3 35.3 32.3 35.8
DATA> 29.3 26.0 29.8 28.0 28.8
DATA> END

Data Display
 Row  STRAINS  BLOCKS  YIELD

   1        A       1   32.3
   2        A       2   34.0
   3        A       3   34.3
   4        A       4   35.0
   5        A       5   36.5
   6        B       1   33.3
   7        B       2   33.0
   8        B       3   36.3
   9        B       4   36.8
```

```
10        B       5   34.5
11        C       1   30.8
12        C       2   34.3
13        C       3   35.3
14        C       4   32.3
15        C       5   35.8
16        D       1   29.3
17        D       2   26.0
18        D       3   29.8
19        D       4   28.0
20        D       5   28.8


MTB > GLM 'YIELD' = BLOCKS STRAINS;
SUBC>   Brief 1 ;
SUBC>   Means STRAINS;
SUBC>   Pairwise STRAINS;
SUBC>     Tukey;
SUBC>     NoCI.

General Linear Model: YIELD versus BLOCKS, STRAINS

Factor     Type Levels Values
BLOCKS    fixed      5 1 2 3 4 5
STRAINS   fixed      4 A B C D

Analysis of Variance for YIELD, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
BLOCKS      4     21.462     21.462      5.365    2.45  0.103
STRAINS     3    134.448    134.448     44.816   20.48  0.000
Error      12     26.262     26.262      2.188
Total      19    182.172

Least Squares Means for YIELD

STRAINS      Mean   SE Mean
A           34.42    0.6616
B           34.78    0.6616
C           33.70    0.6616
D           28.38    0.6616

Tukey Simultaneous Tests
Response Variable YIELD
All Pairwise Comparisons among Levels of STRAINS

STRAINS = A subtracted from:

Level      Difference       SE of             Adjusted
STRAINS      of Means  Difference   T-Value    P-Value
B               0.360      0.9356     0.385     0.9797
C              -0.720      0.9356    -0.770     0.8666
D              -6.040      0.9356    -6.456     0.0002

STRAINS = B subtracted from:

Level      Difference       SE of             Adjusted
STRAINS      of Means  Difference   T-Value    P-Value
C              -1.080      0.9356    -1.154     0.6649
D              -6.400      0.9356    -6.840     0.0001

STRAINS = C subtracted from:

Level      Difference       SE of             Adjusted
STRAINS      of Means  Difference   T-Value    P-Value
D              -5.320      0.9356    -5.686     0.0005
```

The results from Tukey's pairwise comparison tests indicate that A is significantly different from D, so are B and C from D. The other pairs are not significant. The results of the test are summarized in Table 11.5.

Table 11.5 Results of the pairwise comparisons

Comparisons	A & B	A & C	A & D	B & C	B & D	C & D
Result			*		*	*

* Significant at $\alpha = 0.05$ level of significance

The results in Table 11.5 are succinctly displayed in the following table.

μ_D	μ_C	μ_A	μ_B
28.38	33.70	34.42	34.78

11.1.4 Calculation of the Error SS from Residuals

The following results in Table 11.6 could have been obtained from a randomized block experiment with four treatments in each of the three blocks. The figures were actually chosen to simplify the calculations to be made.

Table 11.6 Results of an hypothetical experiment

	Treatments			
Blocks	*A*	*B*	*C*	*D*
1	16	14	20	10
2	14	23	25	18
3	15	17	18	14

The analysis of the above data gives the following ANOVA Table.

The error mean square is the amount of random variation among the results that cannot be explained by differences between blocks and differences between treatments. To see this, we calculate it in another way.

First, calculate the overall mean, the four treatment and the three block means. Each of the 12 results are influenced by the block in which they occur, and the treatment they receive. To remove these effects subtract from each result the overall mean, the deviation of the appropriate block mean from the overall mean, and the deviation of the appropriate treatment mean from the overall mean i.e.,

$$\begin{aligned}\text{Result} - \text{Overall mean} &- (\text{Block mean} - \text{Overall mean}) \\ &- (\text{Treatment mean} - \text{Overall mean})\end{aligned}$$

Table 11.7 ANOVA table for in Table 11.6

Source	d.f.	SS	MS
Blocks	2	56	28
Treatments	3	90	30
Error	6	46	7.7
Total	11	192	
Correction factor	1	3468	

or more simply

$$\text{Result} - \text{Block mean} - \text{Treatment mean} + \text{Overall mean}.$$

The results of this is given below in Table 11.8

Blocks	Total	Means	Treatments	Total	Means
1	60	15	A	45	15
2	80	20	B	54	18
3	64	16	C	63	21
	204	17	D	42	14
				204	17

Table 11.8 Table of residuals

Blocks	A	B	C	D	Total
1	$+3$	-2	$+1$	-2	0
2	-4	$+2$	$+1$	$+1$	0
3	$+1$	0	-2	$+1$	0
Total	0	0	0	0	0

where for example $+3 = 16 - 15 - 15 + 17$ and the error sum of squares are calculated as;

$$\begin{aligned}\text{Error SS} &= (+3)^2 + (-2)^2 + (+1)^2 + (-2)^2 + (-4)^2 + (+2)^2 + (+1)^2 \\ &\quad + (+1)^2 + (0)^2 + (-2)^2 + (+1)^2 \\ &= 46\end{aligned}$$

Notice that each block and treatment residuals adds to zero. Those twelve values in Table 11.8 are called the residuals.It should be noted that this is not the best way to calculate the error sum of squares but simply to illustrate where it comes from. The correct way to calculate it is from an analysis of variance. The analysis of variance is simpler and quicker, and avoids difficulties with rounding-off errors that would occur with real data.

11.2 Missing Values in a RCB Design

Sometimes an observation in one of the blocks of a randomized block design may be missing. The reason for these could be:

1. Due to carelessness or error (e.g., in accurate recording of yields or wrongful application of a treatment to a wrong plot). Many times, data are recorded, whose values are not consistent with common sense or expectation or with the rest of the data so collected in the experiment. In some

cases, such an outlier may be a true reflection of true differences in yields of the treatments but in most cases, these data are often wrongly recorded, data in this category are of measurement type (such as plant height, seed weight, or protein content). If we can not correct this anomaly, such a data value/s should be considered missing.
2. Due to reasons beyond our control such as unavoidable damage such as pest destruction or damage of a crop in a plot (e.g., stem borers destroying a plot of planted maize).
3. Due to no or poor germination, or an animal death in an experiment.
4. Other common causes could be excessive water logging in a plot which ultimately destroyed the plants or destruction by stray grazing cattle.
5. Due to loss of harvested samples because sometimes certain traits of a crop can only be determined outside the experimental field (where the original samples were taken). For example, the grain yields of maize/plot can only be measured after dihusking and threshing. Other examples such as measuring plant heights, leaf areas, and protein content of a cowpea can only be measured in a laboratory. In such situations, it is not uncommon, in spite of best efforts of the researcher to avoid missing values, that some samples may be missing between the field and the laboratory or in some cases, completely misplaced. When no measurement of traits from such missing samples are possible, then such sample/s should be considered missing.

A missing observation introduces a new problem into the analysis, since treatments are no longer orthogonal to blocks, i.e., every treatment does not occur in every block. The general approach to missing value problem is an approximate analysis in which the missing observation is estimated, and then the usual analysis of variance is performed proceeding just as if the missing observation were real data, with both the error and total degrees of freedom reduced by one each.

If the plot corresponding to the k-th treatment and the l-th block (Y_{kl}) is missing, then it can be estimated approximately from the remaining data from the expression

$$x_0 = \frac{tT_k^{'} + bB_l^{'} - G^{'}}{(b-1)(t-1)} \tag{11.2}$$

where, $T_k^{'}$ is the total for all known yields for plots receiving treatment k, $B_l^{'}$is the sum of all known yields from plots in Block l and $G^{'}$ is the sum of all known yields.

The variance of the mean of a treatment estimate $\hat{t}_k$ is given by

$$\text{Var}\,(\hat{t}_k) = \frac{\sigma^2}{b}\left[1 + \frac{t}{(b-1)(t-1)}\right] \tag{11.3}$$

The variance of any other treatment mean is still given by $\dfrac{\sigma^2}{b}$.

Situations when more than one plot is missing is complicated and the standard procedure is to use Yates' algorithm to estimate the missing values. This will not be described in this text.

11.2.1 Example 11.2.2 Let us consider the data in example 11.1. Suppose the value for strain D in block 5 is 'missing'. Then without this value, we have

$$\text{Strain D total} = T_D^{'} = 113.1$$

$$\text{Block 5 total} = B_5^{'} = 106.8$$

$$\text{Grand Total} = G^{'} = 627.6$$

Missing value estimate from (11.2) is therefore estimated as,

$$x_0 = \frac{4 \times 113.1 + 5 \times 106.8 - 627.6}{4 \times 3} = 29.9$$

This value of 29.9 compares favorably with the real value of 28.8. This estimated value is entered in the table with the observed values and the analysis of variance is performed as usual with reduced degrees of freedom in both the total and error lines.

Including this missing value, the new value will be,

$$\text{Treatment D Total} = 143.0$$

$$\text{Block 5 Total} = 136.7$$

$$\text{Grand Total} = 657.5$$

In Table 11.9 are presented the analysis of variance with the estimated missing value included.

Table 11.9 ANOVA table for a missing value analysis

Source	d.f.	SS	MS	F
Blocks	4	24.08	6.02	
Treatments	3	124.86	41.62	17.94
Error	11	25.04	2.32	
Total	18	174.48		
CF	1	21615.31		

Note the new degrees of freedom for Error $=(12-1) = 11$ and for Total $=(19-1) = 18$ since one parameter x_0 was already estimated from the data.

Strain means	A	34.4
	B	34.8
	C	33.7
	D	28.6

S.E. for comparing any two means of strains A,B, and/or C is computed as:

$$\sqrt{\frac{2S^2}{b}} = \sqrt{\frac{2S^2}{5}} = \sqrt{\frac{2 \times 2.32}{5}} = 0.96$$

S.E. for comparing strain D with any other strains from the expression in (11.3) is calculated as:

$$\sqrt{S^2\left[\frac{2}{b} + \frac{t}{b(b-1)(t-1)}\right]} = \sqrt{\left\{\frac{S^2}{5}(2 + \frac{4}{4 \times 3}\right\}} = \sqrt{\left(\frac{7 \times S^2}{15}\right)} = 1.04$$

Both standard errors are based on 11 d.f. A MINITAB implementation of the missing value problem is implemented in what follows, but first the data are read in with the particular observation read in with an asterisks indicating to MINITAB to treat that observation as missing. The analysis is then carried out as usual.

```
Data Display


 Row  STRAINS  BLOCKS  YIELD

   1        A       1   32.3
   2        A       2   34.0
   3        A       3   34.3
   ..........................
  17        D       2   26.0
  18        D       3   29.8
  19        D       4   28.0
  20        D       5      *


MTB > GLM 'YIELD' = BLOCKS STRAINS;
SUBC>   Brief 1 ;
SUBC>   Means STRAINS;
SUBC>   Pairwise STRAINS;
SUBC>     Tukey;
SUBC>     NoCI.

General Linear Model: YIELD versus BLOCKS, STRAINS


Factor     Type Levels Values
BLOCKS    fixed      5 1 2 3 4 5
STRAINS   fixed      4 A B C D

Analysis of Variance for YIELD, using Adjusted SS for Tests

Source      DF     Seq SS     Adj SS     Adj MS        F      P
BLOCKS       4     39.131     21.967      5.492     2.37  0.117
STRAINS      3    100.494    100.494     33.498    14.43  0.000
Error       11     25.536     25.536      2.321
Total       18    165.161

Least Squares Means for YIELD
```

```
STRAINS      Mean   SE Mean
 A          34.42    0.6814
 B          34.78    0.6814
 C          33.70    0.6814
 D          28.60    0.7868

Tukey Simultaneous Tests
Response Variable YIELD
All Pairwise Comparisons among Levels of STRAINS

STRAINS = A subtracted from:

Level       Difference          SE of              Adjusted
STRAINS       of Means     Difference    T-Value     P-Value
B                0.360         0.9636      0.374      0.9813
C               -0.720         0.9636     -0.747      0.8760
D               -5.820         1.0408     -5.592      0.0008

STRAINS = B subtracted from:

Level       Difference          SE of              Adjusted
STRAINS       of Means     Difference    T-Value     P-Value
C               -1.080         0.9636     -1.121      0.6850
D               -6.180         1.0408     -5.938      0.0005

STRAINS = C subtracted from:

Level       Difference          SE of              Adjusted
STRAINS       of Means     Difference    T-Value     P-Value
D               -5.100          1.041     -4.900      0.0023
```

The analysis of variance table produced from MINITAB agrees with the one presented in Table 11.9, and the standard errors computed also agree with those computed earlier by hand.

11.2.1 Summary of Results of Analysis

The summary of results again indicate that treatment D is significantly different from each of A, B, and C, while there are no significant differences on the means of A, B and C at the 0.1 % significance levels. It could therefore be concluded based on the analysis above that the overall differences among the strain means were significant at 0.1 %. The differences among the strains A, B, and C were not significant. Strain D yielded 5.7 lb/plot less than the average yield of strains A, B, and C i.e., the result can be succinctly summarized as follows:

34.78	34.42	33.70	28.38
B	A	C	D

11.3 Partitioning Treatment SS in a RCBD

Example 11.3.1

An experiment to investigate the use of soil fumigants for the control of eelworm used four blocks of six plots each. Two plots in each blocks were untreated (control), the other four treatments consisted of two fumigants, F_1, F_2 each at two levels, F_{11} and F_{12}; and F_{21} and F_{22}. The numbers of cysts per 100 g of soil were counted and these were given in Table 11.10.

Table 11.10 Number of cysts in 100 g of soil

Treatments	Blocks I	II	III	IV	Total
Control C_1	86	91	108	76	361
Control C_2	141	71	113	47	372
F_1 level 1 F_{11}	48	55	108	20	231
F_1 level 2 F_{12}	93	42	78	7	220
F_2 level 1 F_{21}	64	59	67	33	223
F_2 level 2 F_{22}	70	36	102	73	281
Total	502	354	576	256	1688

The initial analysis of variance is calculated in the usual way, and the results are given in Table 11.11.

Table 11.11 Analysis of variance table for the data in table 11.10

Source	d.f.	SS	MS	F
Blocks	3	1,0382	3461	
Treatments	5	6066	1213	2.60
Error	15	6871	450	
Total	23	2,3319		
CF	1	11,8723		

The MINITAB implementation of the analysis is presented, together with a partial output in the following,

```
MTB > SET C1
DATA> (1:6)4
DATA> END
MTB > SET C2
DATA> 6(1:4)
DATA> END
MTB > SET C3
DATA> 86 91 108 76 141 71 113 47
DATA> 48 55 108 20 93 42 78 7
DATA> 64 59 67 33 70 36 102 73
DATA> END
MTB > PRINT C1-C3

Data Display
 Row   TRT  BLOCKS  COUNT

   1     1       1     86
   2     1       2     91
   3     1       3    108
   4     1       4     76
   5     2       1    141
   6     2       2     71
   7     2       3    113
   8     2       4     47
   9     3       1     48
  10     3       2     55
  11     3       3    108
  12     3       4     20
  13     4       1     93
  14     4       2     42
  15     4       3     78
  16     4       4      7
  17     5       1     64
  18     5       2     59
  19     5       3     67
  20     5       4     33
  21     6       1     70
  22     6       2     36
  23     6       3    102
  24     6       4     73

MTB > GLM 'COUNT' = BLOCKS TRT;
SUBC>   Brief 1 .

General Linear Model: COUNT versus BLOCKS, TRT


Factor     Type Levels Values
BLOCKS    fixed      4 1 2 3 4
TRT       fixed      6 1 2 3 4 5 6

Analysis of Variance for COUNT, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
BLOCKS      3   10382.7    10382.7     3460.9    7.56  0.003
TRT         5    6066.3     6066.3     1213.3    2.65  0.066
Error      15    6868.3     6868.3      457.9
Total      23   23317.3
```

The analysis indicates that there are significant differences in the means of the six treatments at $\alpha = 0.10$ level of significance. The treatment sum of squares can be sub-divided into components corresponding to comparisons of interest.

(i) Comparisons between the average means of the controls and average means of the the fumigants, which has the null hypothesis and equivalent accompanying contrast C_1 as:

$$H_0 : \frac{(\mu_1 + \mu_2)}{2} = \frac{(\mu_3 + \mu_4 + \mu_5 + \mu_6)}{4}$$

$$C_1 : = 4\mu_1 + 4\mu_2 - 2\mu_3 - 2\mu_4 - 2\mu_5 - 2\mu_6$$

$$\text{Total for the control plots} = 733$$

$$\text{Total for the fumigant plots} = 955$$

Hence the SS corresponding to this contrast is calculated after as,

$$\text{SS} = \frac{733^2}{8} + \frac{955^2}{16} - \text{CF} = 5440 \quad \text{on 1 d.f.}$$

(ii) Comparison between the average means of the controls. This is equivalent to the following hypothesis and contrast

$$H_0 : \mu_1 = \mu_2$$

$$C_2 : = \mu_1 - \mu_2$$

$$\text{SS} = \frac{361^2}{4} + \frac{372^2}{4} - \frac{733^2}{8} = 15 \quad \text{on 1 d.f.}$$

(iii) Comparison between the four fumigants.

$$\text{SS} = \frac{231^2}{4} + \frac{220^2}{4} + \frac{223^2}{4} + \frac{281^2}{4} - \frac{955^2}{16} = 612 \quad \text{on 3 d.f.}$$

It is obvious that this last sum of squares is so small compared with the error mean square that any component of the SS must be non-significant.

It would however be possible to subdivide this further as follows:

(a) Comparison between fumigants F_1 and F_2 which is equivalent to contrast:

$$H_0 : \frac{(\mu_3 + \mu_4)}{2} = \frac{(\mu_5 + \mu_6)}{2}$$

$$C_3 : = \mu_3 + \mu_4 - \mu_5 - \mu_6$$

$$\text{SS} = \frac{451^2}{8} + \frac{504^2}{8} - \frac{955^2}{16} = 176 \quad \text{on 1 d.f.}$$

(b) Comparison between levels of fumigants F_1. Equivalent to the contrast

$$C_4 : \mu_3 - \mu_4$$

$$\text{SS} = \frac{231^2}{4} + \frac{220^2}{4} - \frac{451^2}{8} = 16 \quad \text{on 1 d.f.}$$

(c) Comparison between levels of fumigants F_2. Equivalent to the contrast

$$C_5 : \mu_5 - \mu_6$$

$$\text{SS} = \frac{223^2}{4} + \frac{281^2}{4} - \frac{504^2}{8} = 420 \quad \text{on 1 d.f.}$$

Note that the sum of SS for contrasts C_3, C_4, and C_5 add up to 612. Similarly, the sum SS for the five contrasts add up to 6066, that is the total treatment SS.

The subdivision of the treatment sum of squares can be summed up in a second analysis of variance in Table 11.12.

Table 11.12 Second analysis of variance table

Comparisons	d.f.	SS	MS	F
Control Vs fumigants	1	5440	5440	11.9
Between controls	1	15	15	0.03
F_1 Vs F_2	1	176	176	0.38
F_{11} Vs F_{12}	1	16	16	0.03
F_{21} Vs F_{22}	1	420	420	0.9
Error	15	6871	458	

Clearly the only effect shown is that the fumigants reduce the numbers of cysts. The different fumigants and different levels do not appear to act differentially. The mean number of cysts per 100 g of soil are 91.6 for untreated soil and 59.7 for treated soil; the difference between the means is 31.9 and the standard error for this difference is $\sqrt{458(\frac{1}{8} + \frac{1}{16})} = 9.2$.

We can write the five contracts in the form:

Contrasts	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
L_1	2	2	−1	−1	−1	−1
L_2	1	−1	0	0	0	0
L_3	0	0	1	1	−1	−1
L_4	0	0	1	−1	0	0
L_5	0	0	0	0	1	−1

We see that each pair of contrasts is orthogonal and hence we should not be surprised that the five contrasts sum of squares add up to the original treatments sum of squares based on 5 d.f. Note that the original F values of 2.60 is the average of the five F values, i.e., $2.6 = \dfrac{11.9 + 0.03 + 0.38 + 0.03 + 0.9}{5}$.

To implement the above contrasts in MINITAB, first create the five contrasts into C1–C5 as covariates L1–L5 respectively. The covariates are created as in the program below. Where for L1 for instance, we have 2's for levels 1 and 2 for treatment and −1 for the remaining levels. The model is then fitted by declaring that L1–L5 are covariates in the GLM model statement.

```
MTB > SET C5
DATA> 4(1) 4(-1) 16(1)
DATA> END
MTB > SET C6
DATA> 8(0) 8(1) 8(-1)
DATA> END
MTB > SET C7
DATA> 8(0) 4(1) 4(-1) 8(0)
DATA> END
MTB > SET C8
DATA> 16(0) 4(1) 4(-1)
DATA> END
MTB > PRINT C1-C8

Data Display


 Row   TRT  BLOCKS  COUNT    L1    L2    L3    L4    L5

   1     1       1     86     2     1     0     0     0
   2     1       2     91     2     1     0     0     0
   3     1       3    108     2     1     0     0     0
   4     1       4     76     2     1     0     0     0
   5     2       1    141     2    -1     0     0     0
   6     2       2     71     2    -1     0     0     0
   7     2       3    113     2    -1     0     0     0
   8     2       4     47     2    -1     0     0     0
   9     3       1     48    -1     1     1     1     0
  10     3       2     55    -1     1     1     1     0
  11     3       3    108    -1     1     1     1     0
  12     3       4     20    -1     1     1     1     0
  13     4       1     93    -1     1     1    -1     0
  14     4       2     42    -1     1     1    -1     0
  15     4       3     78    -1     1     1    -1     0
  16     4       4      7    -1     1     1    -1     0
  17     5       1     64    -1     1    -1     0     1
  18     5       2     59    -1     1    -1     0     1
  19     5       3     67    -1     1    -1     0     1
  20     5       4     33    -1     1    -1     0     1
  21     6       1     70    -1     1    -1     0    -1
  22     6       2     36    -1     1    -1     0    -1
  23     6       3    102    -1     1    -1     0    -1
  24     6       4     73    -1     1    -1     0    -1

MTB > GLM 'COUNT' = BLOCKS ;
SUBC>   Covariates 'L1' 'L2' 'L3' 'L4' 'L8';
SUBC>   SSquares 1;
SUBC>   Brief 1 .

General Linear Model: COUNT versus BLOCKS


Factor     Type Levels Values
BLOCKS    fixed      4 1 2 3 4

Analysis of Variance for COUNT, using Sequential SS for Tests
```

```
Source     DF     Seq SS     Adj SS     Seq MS       F      P
 L1         1     5440.0     2989.0     5440.0   11.88  0.004
 L2         1       15.1       15.1       15.1    0.03  0.858
 L3         1      175.6      175.6      175.6    0.38  0.545
 L4         1       15.1       15.1       15.1    0.03  0.858
 L8         1      420.5      420.5      420.5    0.92  0.353
BLOCKS      3    10382.7    10382.7     3460.9    7.56  0.003
Error      15     6868.3     6868.3      457.9
Total      23    23317.3
```

Note that we have requested for the sequential SS for this analysis in MINITAB. The results agree with those obtained by hand calculations earlier.

Alternatively, we could use the following coding system and subsequent analysis to achieve the same goal. Here, the coefficients for L1–L5 are stored in columns C5–C9. The print out of these columns agree with those presented earlier. We have printed the first and last five observations here. The ANOVA analysis agrees with what we had earlier.

```
MTB > code (1) 2 (2) 2 (3) -1 (4) -1 (5) -1 (6) -1 c1 c5
MTB >  code (1) 1 (2) -1 (3) 0 (4) 0 (5) 0 (6) 0 c1 c6
MTB >  code (1) 0 (2) 0 (3) 1 (4) 1 (5) -1 (6) -1 c1 c7
MTB >  code (1) 0 (2) 0 (3) 1 (4) -1 (5) 0 (6) 0 c1 c8
MTB >  code (1) 0 (2) 0 (3) 0 (4) 0 (5) 1 (6) -1 c1 c9
MTB > GLM 'COUNT' = BLOCKS L1 L2 L3 L4 L5;
SUBC>   Covariates 'L1' 'L2' 'L3' 'L4' 'L5';
SUBC>   Brief 2 .

MTB > print c1-c3 c5-c9

Data Display

Row  TRT  BLOCKS  COUNT  L1  L2  L3  L4  L5
  1    1       1     86   2   1   0   0   0
  2    1       2     91   2   1   0   0   0
  3    1       3    108   2   1   0   0   0
  4    1       4     76   2   1   0   0   0
  5    2       1    141   2  -1   0   0   0
  .....................................
 20    5       4     33  -1   0  -1   0   1
 21    6       1     70  -1   0  -1   0  -1
 22    6       2     36  -1   0  -1   0  -1
 23    6       3    102  -1   0  -1   0  -1
 24    6       4     73  -1   0  -1   0  -1

General Linear Model: COUNT versus BLOCKS

Factor  Type   Levels  Values
BLOCKS  fixed       4  1, 2, 3, 4
```

```
Analysis of Variance for COUNT, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS  Adj MS     F      P
BLOCKS   3  10382.7  10382.7  3460.9  7.56  0.003
L1       1   5440.0   5440.0  5440.0 11.88  0.004
L2       1     15.1     15.1    15.1  0.03  0.858
L3       1    175.6    175.6   175.6  0.38  0.545
L4       1     15.1     15.1    15.1  0.03  0.858
L5       1    420.5    420.5   420.5  0.92  0.353
Error   15   6868.3   6868.3   457.9
Total   23  23317.3


S = 21.3983   R-Sq = 70.54%   R-Sq(adj) = 54.83%


Term        Coef  SE Coef      T      P
Constant  70.333    4.368  16.10  0.000
L1        10.646    3.089   3.45  0.004
L2        -1.375    7.565  -0.18  0.858
L3        -3.312    5.350  -0.62  0.545
L4         1.375    7.565   0.18  0.858
L5        -7.250    7.565  -0.96  0.353
```

This author will recommend the latter form of coding as it less cumbersome than the earlier method.

Example 11.3.2

In an experiment to compare the performance of six newly introduced varieties of maize, it was thought fit to introduce four control varieties a, b, c and d. The experiment was laid out in a randomized block design with ten plots per block and five blocks. Table 11.13 gives the data for this experiment.

Table 11.13 The yields in (kg/plot) for the experiment in this example

	Blocks					
Treatment	I	II	III	IV	V	Total
1a	1.63	1.48	1.43	1.76	1.17	7.47
1b	1.73	1.42	1.50	1.06	0.76	6.47
lc	1.49	1.70	1.52	1.48	0.85	7.04
1d	1.25	1.36	0.93	1.38	0.68	5.60
2	1.07	1.28	1.28	1.83	1.16	6.62
3	0.73	1.42	1.30	1.38	0.70	5.53
4	0.69	1.69	1.61	1.61	1.17	6.67
5	0.52	1.50	1.17	1.05	1.04	5.28
6	1.63	1.34	1.07	1.01	0.87	5.92
7	1.08	1.33	0.63	1.21	0.51	4.76
Block total	11.82	14.42	12.44	13.77	8.91	61.36

The structure of the analysis of variance (degrees of freedom only) for this example is given in the following table.

Source	d.f
Blocks	4
Treatments	9
Error	36
Total	49

The 9 d.f for treatments can be partitioned into the following components

(i) Control Vs rest with 1 d.f.
(ii) Between controls with 3 d.f.
(iii) Between other varieties with 5 d.f.

Thus, the total degrees of freedom is $1+3+5=9$ d.f.

11.3.1 Analysis

First we obtain the total for variety 1 and varieties 2 to 7 viz.,

$$\text{Total for variety 1 (Controls)} = 26.58$$
$$\text{Total for varieties 2–7} = 34.78$$

Next we compute the various SS as shown below,

$$\text{Correction Factor (CF)} = \frac{61.46^2}{50} = 75.5466$$
$$\text{Control Vs Rest SS} = \frac{26.58^2}{20} + \frac{34.88^2}{30} - \text{CF} = 0.3320$$
$$\text{Between Control SS} = \frac{7.47^2}{5} + \frac{6.47^2}{5} + \frac{7.04^2}{5} + \frac{5.60^2}{5} - \frac{26.58^2}{20} = 0.3919$$
$$\text{Between other Varieties} = \frac{6.62^2}{5} + \cdots + \frac{4.76^2}{5} - \frac{34.88^2}{30} = 0.6103$$
$$\text{Total SS} = 1.63^2 + 1.48^2 + \cdots + 0.51^2 - \frac{61.46^2}{50} = 5.8892$$
$$\text{Blocks SS} = \frac{11.82^2}{10} + \frac{14.42^2}{10} + \cdots + \frac{8.91^2}{10} - \text{CF} = 1.8831$$

The analysis of variance of the data in this example is displayed in Table 11.14.

Table 11.14 Analysis of variance table

Source	d.f.	SS	MS	F
Blocks	4	1.8831		
Treatments	9	1.3342	0.1482	2.00
(a) Controls vs rest	1	0.3320	0.3320	4.47*
(b) Between controls	3	0.3919	0.1306	1.76
(c) Between others	5	0.6103	0.1221	1.65
Error	36	2.6719	0.0742	
Total	49	5.8892		

*Significant at 5 % point

Variety	Means (kg/plot)
1	1.33
2	1.33
3	1.11
4	1.33
5	1.06
6	1.18
7	0.95

From the ANOVA table $S^2 = 0.0738$. Hence, the standard error (s.e) for comparing control (variety 1) mean with any other treatment mean equals

$$\sqrt{\left(\frac{S^2}{20} + \frac{S^2}{5}\right)} = \sqrt{\left(\frac{0.0738}{20} + \frac{0.0738}{5}\right)} = 0.136 \quad (36 \text{ d.f})$$

S.E for comparing any two means, not including control equals

$$\sqrt{\left(\frac{S^2}{5} + \frac{S^2}{5}\right)} = \sqrt{\left(\frac{0.0738}{5} + \frac{0.0738}{5}\right)} = 0.172 \quad (36 \text{ d.f})$$

The controls mean $= 1.33$ and the other varieties mean $= 1.16$, hence the standard error of difference equals,

$$\sqrt{\left(\frac{S^2}{20} + \frac{S^2}{30}\right)} = \sqrt{\left(\frac{0.0738}{20} + \frac{0.0738}{30}\right)} = 0.078 \quad (36 \text{ d.f})$$

The results in Table 11.14 show that none of the six new varieties is better than the control. Variation between the mean yields for the six varieties is not significant and the control gives a significantly better yield than the average variety mean. The above analysis is implemented in MINITAB as follows:

```
MTB > SET C1
DATA> (1:10)5
DATA> END
MTB > SET C2
DATA> 10(1:5)
DATA> END
MTB > SET C3
DATA> 1.63 1.48 1.43 1.76 1.17
DATA> 1.73 1.42 1.50 1.06 0.76
DATA> 1.49 1.70 1.52 1.48 0.85
DATA> 1.25 1.36 0.93 1.38 0.68
DATA> 1.07 1.28 1.28 1.83 1.16
DATA> 0.73 1.42 1.30 1.38 0.70
DATA> 0.69 1.69 1.61 1.61 1.17
DATA> 0.52 1.50 1.17 1.05 1.04
DATA> 1.63 1.34 1.07 1.01 0.87
DATA> 1.08 1.33 0.63 1.21 0.51
DATA> END
MTB > SET C4
DATA> 20(6) 30(-4)
DATA> END
MTB > PRINT C1-C4

Data Display

 Row    TRT  BLOCKS  YIELDS   CT1

   1      1       1    1.63     6
   2      1       2    1.48     6
   3      1       3    1.43     6
   4      1       4    1.76     6
   5      1       5    1.17     6
   6      2       1    1.73     6
   7      2       2    1.42     6
   8      2       3    1.50     6
   9      2       4    1.06     6
  10      2       5    0.76     6
  11      3       1    1.49     6
  12      3       2    1.70     6
  13      3       3    1.52     6
  14      3       4    1.48     6
  15      3       5    0.85     6
  16      4       1    1.25     6
  17      4       2    1.36     6
  18      4       3    0.93     6
  19      4       4    1.38     6
  20      4       5    0.68     6
  21      5       1    1.07    -4
  22      5       2    1.28    -4
  23      5       3    1.28    -4
  24      5       4    1.83    -4
  25      5       5    1.16    -4
  26      6       1    0.73    -4
  27      6       2    1.42    -4
  28      6       3    1.30    -4
  29      6       4    1.38    -4
```

```
  30      6       5    0.70    -4
  31      7       1    0.69    -4
  32      7       2    1.69    -4
  33      7       3    1.61    -4
  34      7       4    1.61    -4
  35      7       5    1.17    -4
  36      8       1    0.52    -4
  37      8       2    1.50    -4
  38      8       3    1.17    -4
  39      8       4    1.05    -4
  40      8       5    1.04    -4
  41      9       1    1.63    -4
  42      9       2    1.34    -4
  43      9       3    1.07    -4
  44      9       4    1.01    -4
  45      9       5    0.87    -4
  46     10       1    1.08    -4
  47     10       2    1.33    -4
  48     10       3    0.63    -4
  49     10       4    1.21    -4
  50     10       5    0.51    -4

MTB > GLM 'YIELDS' = BLOCKS TRT;
SUBC>   SSquares 1;
SUBC>   Brief 1 .

General Linear Model: YIELDS versus BLOCKS, TRT


Factor     Type Levels Values
BLOCKS    fixed      5 1 2 3 4 5
TRT       fixed     10  1  2  3  4  5  6  7  8  9 10

Analysis of Variance for YIELDS, using Sequential SS for Tests

Source     DF      Seq SS      Adj SS     Seq MS       F       P
BLOCKS      4     1.88311     1.88311    0.47078    6.34   0.001
TRT         9     1.33417     1.33417    0.14824    2.00   0.069
Error      36     2.67189     2.67189    0.07422
Total      49     5.88917
```

11.4 Paired Comparisons

Data frequently occur in pairs. For instance, two treatments A and B may be applied to a pair of identical twins, treatment A to one twin, and treatment B to the other twin. We can then consider a situation in which we have n pairs of twins in this particular example. Here pairing was done before the commencement of the experiment on the basis of expected similar responses when there were no treatment effects and the individuals are matched. Thus members of each pair are similar to each other with respect to any extraneous factors.

Another example for instance is an experiment involving two feed rations applied to two animals from each of four litters of pigs by assigning the animals of each litter at random, one to each ration. Yet another example is the response of two sugar cane varieties grown on paired plots at each of six locations in Kwara State, to the infestation of a certain root nematode (Hetero Sacchara). Our main interest in the above experiments is to compare the effects of the two treatments (variations on varieties) and see if there is a difference between them.

The randomized complete block which has two experimental units per block is also an example of a paired design. Many biological and medical investigations often times involve repeated measurements made on the same plants, subjects, animals at different times. Some of these investigations could be studies relating to growth, development, and those involving before and after application of a certain stimulus. The case involving two measurements at only two times constitute paired observations and the procedure that will be discussed in this section might be appropriate for such data.

Case-control studies are often matched observational studies where the controls are matched with the experimental group. A study of cholesterol reduction in subjects for instance identified 50 patients with high blood cholesterol. These individuals are then matched by age, sex, and educational level to form a control group. The reduction in cholesterol level after 10 weeks of a particular treatment is then compared.

We shall employ the method of *student's paired comparison's* for the analysis of data obtained from such experiments.

Example 11.4.1

We illustrate this method with the following example which relates to the effect of exposure of flowers to different environmental conditions. He choose ten vigorous plants with freely exposed flowers at the top and flowers hidden as much as possible at the bottom. Finally, he determined the number of seeds set per two pods at each location. The data are given in Table 11.15.

Should we wish to test the hypothesis of no difference between population means against the alternative that top flowers get more seeds, i.e.,

Table 11.15 Number of seeds set/pod at two locations

Plant	1	2	3	4	5	6	7	8	9	10
Top flowers	4.0	5.2	5.7	4.2	4.8	3.9	4.1	3.0	4.6	6.8
Bottom flowers	4.4	3.7	4.7	2.8	4.2	4.3	3.5	3.7	3.1	1.9

$$H_0 : \mu_t \leq \mu_b \qquad (11.4)$$
$$H_a : \mu_t > \mu_b$$

11.4.1 Analysis

Let the observation on the top and the bottom flowers be designated as Y_{1i} and Y_{2i} $(i = 1, 2, \cdots, 10)$ respectively. Then we need to compute $D_i = Y_{1i} - Y_{2i}$ and the hypotheses in (11.4) then become in terms of D_i,

$$\begin{aligned} H_0 &: \mu_d \leq 0 \\ H_a &: \mu_d > 0 \end{aligned} \tag{11.5}$$

To implement the null hypothesis in (11.5) therefore, we need to calculate the mean of D_i, namely, $\bar{d}$ and the variance of D_i, again, namely, s_d^2.

Table 11.16 Summary statistics for our analysis

Plants	Top flowers Y_{1i}	Bottom flowers Y_{2i}	Difference $d_i = Y_{1i} - Y_{2i}$
1	4.0	4.4	−0.4
2	5.2	3.7	1.5
3	5.7	4.7	1.0
4	4.2	2.8	1.4
5	4.8	4.2	0.6
6	3.9	4.3	0.6
7	4.1	3.5	0.6
8	3.0	3.7	−0.7
9	4.6	3.7	0.9
10	6.8	1.9	4.9
$\sum d_i$			10.0
$\sum d_i^2$			33.0
$\bar{d}$			1.0

From Table 11.16, we have, $\sum d_i = 10$, and hence, $\bar{d} = 1.0$. Similarly, $\sum d^2 = 33$. Hence,

$$s_d^2 = \frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1} = \frac{33 - \frac{10^2}{10}}{10-1} = 2.5556$$

The equivalent one-sample Students' t-test becomes:

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}} = \frac{1.0}{\sqrt{0.2556}} = 1.977$$

which is significant at $\alpha = 0.05$, indicating that the top flowers set more seeds than the lower flowers. Note that this is a one-tailed test. The paired t-test is implemented in MINITAB with the following:

```
MTB > PRINT C1-C2

Data Display
 Row     Y1     Y2

   1    4.0    4.4
   2    5.2    3.7
   3    5.7    4.7
   4    4.2    2.8
   5    4.8    4.2
   6    3.9    4.3
   7    4.1    3.5
   8    3.0    3.7
   9    4.6    3.1
  10    6.8    1.9

MTB > Paired  'Y1' 'Y2';
SUBC>   Alternative 1.

Paired T-Test and CI: Y1, Y2

Paired T for Y1 - Y2

                  N      Mean     StDev   SE Mean
Y1               10     4.630     1.068     0.338
Y2               10     3.630     0.850     0.269
Difference       10     1.000     1.599     0.506

95% lower bound for mean difference: 0.073
T-Test of mean difference = 0 (vs > 0): T-Value = 1.98  P-Value = 0.040
```

11.5 Test for Non Additivity

The randomized complete block design model in (11.1) assumes that there is no interaction effect between blocks and treatments. In other words, the model implies that the block effect is the same for all treatments, and like wise that the treatments effect is the same for all blocks. This is often referred to as being that the terms are *additive* or simply as *additivity*. It is therefore of utmost interest to check whether the additivity assumption is violated when we run a RCBD analysis. Although for most practical purposes this is not often done, but we present in this section Tukey's single degree of freedom test for non additivity. We refer to the data in Table 11.3 to illustrate the procedure for conducting this test. The data and computations necessary are shown in Table 11.17.

Table 11.17 Tukey's one degree of freedom test for non-additivity

Strain	Blocks					Sum		
	1	2	3	4	5	Y_{i+}	Means	$C_i = \bar{Y}_{i+} - \bar{Y}$
A	32.3	34.0	34.3	35.0	36.5	172.1	34.42	1.6
B	33.3	33.0	36.3	36.8	34.5	173.9	34.78	1.96
C	30.8	34.3	35.3	32.3	35.8	168.5	33.70	0.88
D	29.3	26.0	29.8	28.0	28.8	141.9	28.38	− 4.44
Sum Y_{+j}	125.7	127.3	135.7	132.1	135.6	656.4		0
Means	31.425	31.825	33.925	33.025	33.90	$\bar{Y} = 32.82$		
$C_j = \bar{Y}_{+j} - \bar{Y}$	− 1.395	− 0.995	1.105	0.205	1.08			

From Table 11.17, we have,

$$\sum C_i^2 = 26.8896, \quad \sum C_j^2 = 5.3655$$

If we define $d_i = \sum_j C_j Y_{ij}$, then,

$$d_1 = (32.3 \times -1.395) + \cdots + (36.5 \times 1.08) = 5.608$$
$$d_2 = (33.3 \times -1.395) + \cdots + (34.5 \times 1.08) = 5.627$$
$$d_3 = (30.8 \times -1.395) + \cdots + (35.8 \times 1.08) = 7.1975$$
$$d_4 = (29.3 \times -1.395) + \cdots + (28.8 \times 1.08) = 3.0295$$

and for $i = 1, 2, 3, 4$, we have,

i	c_i	d_i
1	1.6	5.608
2	1.96	5.627
3	0.88	7.1975
4	− 4.44	3.0295

Hence,

$$\sum c_i d_i = (1.6)(5.608) + (1.96)(5.6270 + (0.88)(7.1975) + (-4.44)(3.0295)$$
$$= 12.8845$$

and therefore, the non-additivity SS (NASS) is calculated as,

$$\text{NASS} = \frac{(\sum_i c_i d_i)^2}{(\sum C_i^2 \times \sum C_j^2)} = \frac{(12.8845)^2}{5.3655 \times 26.8896} = 1.151$$

The revised analysis of variance table is presented in Table 11.18.

The computed F-value is not significant. Which indicates that our model is additive. However, when this is significant, and is not due to a few aberrant observations, a transformation would be required.

Table 11.18 Revised analysis of variance table

Source	d.f.	SS	MS	F
Blocks	4	21.46		
Treatments	3	134.45		
Error	12	26.26		
Additivity	1	1.151	1.151	0.50*
Residual	11	25.109	2.283	

*Not significant

The Tukey's test for additivity can be implemented in MINITAB with the following procedure which is not too difficult to follow.

```
MTB > let c4=c3-mean(c3)
MTB > anova c4=c2 c1;
SUBC> means c2 c1.

ANOVA: h.zero versus BLOCKS, STRAINS

Factor     Type Levels Values
BLOCKS    fixed      5     1     2     3     4     5
STRAINS   fixed      4     A     B     C     D

Analysis of Variance for h.zero

Source         DF         SS         MS        F       P
BLOCKS          4    21.46 2      5.365     2.45   0.103
STRAINS         3    134.448     44.816    20.48   0.000
Error          12     26.262      2.189
Total          19    182.172

Means

BLOCKS    N    h.zero
1         4   -1.3950
2         4   -0.9950
3         4    1.1050
4         4    0.2050
5         4    1.0800

STRAINS    N    h.zero
A          5    1.6000
B          5    1.9600
C          5    0.8800
D          5   -4.4400
```

```
MTB > set c5
DATA> 4(-1.3950 -0.9950 1.1050 0.2050 1.0800)
DATA> end
MTB > set c6
DATA> (1.6 1.96 0.88 -4.44)5
DATA> end
MTB > let c7=c5*c6

MTB > GLM 'YIELD' = BLOCKS STRAINS z;
SUBC>   Covariates 'z';
SUBC>   Brief 2 .

General Linear Model: YIELD versus BLOCKS, STRAINS


Factor     Type Levels Values
BLOCKS    fixed      5 1 2 3 4 5
STRAINS   fixed      4 A B C D

Analysis of Variance for YIELD, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
BLOCKS      4     21.462     21.462      5.365    2.35  0.118
STRAINS     3    134.448    134.448     44.816   19.63  0.000
z           1      1.151      1.151      1.151    0.50  0.493
Error      11     25.111     25.111      2.283
Total      19    182.172

Term          Coef   SE Coef        T      P
Constant   32.8200    0.3378    97.14  0.000
z           0.0893    0.1258     0.71  0.493

Unusual Observations for YIELD

Data Display

 Row  STRAINS  BLOCKS  YIELD  h.zero      b1     s1        z

   1        A       1   32.3   -0.52  -1.395   1.60  -2.2320
   2        A       2   34.0    1.18  -0.995   1.60  -1.5920
   3        A       3   34.3    1.48   1.105   1.60   1.7680
   4        A       4   35.0    2.18   0.205   1.60   0.3280
   5        A       5   36.5    3.68   1.080   1.60   1.7280
   6        B       1   33.3    0.48  -1.395   1.96  -2.7342
   7        B       2   33.0    0.18  -0.995   1.96  -1.9502
   8        B       3   36.3    3.48   1.105   1.96   2.1658
   9        B       4   36.8    3.98   0.205   1.96   0.4018
  10        B       5   34.5    1.68   1.080   1.96   2.1168
  11        C       1   30.8   -2.02  -1.395   0.88  -1.2276
  12        C       2   34.3    1.48  -0.995   0.88  -0.8756
  13        C       3   35.3    2.48   1.105   0.88   0.9724
  14        C       4   32.3   -0.52   0.205   0.88   0.1804
  15        C       5   35.8    2.98   1.080   0.88   0.9504
  16        D       1   29.3   -3.52  -1.395  -4.44   6.1938
  17        D       2   26.0   -6.82  -0.995  -4.44   4.4178
  18        D       3   29.8   -3.02   1.105  -4.44  -4.9062
  19        D       4   28.0   -4.82   0.205  -4.44  -0.9102
  20        D       5   28.8   -4.02   1.080  -4.44  -4.7952
```

The result indicate that the additive effect represented by z is not significant.

11.5.1 Alternative Implementation of Tukey's Additivity Test

In this approach we do the following:

(a) Run the RCB analysis and store the fitted values in say Column 9 (C9).
(b) Compute u = fitted values × fitted values and put the result in column 10 (C10).
(c) Now again run the RCB analysis with u in the defining model but declared as a covariate.
(d) The final results will give the additivity components etc.

```
SUBC> Abort.
MTB > Name c9 "FITS2"
MTB > GLM 'YIELD' = BLOCKS STRAINS;
SUBC>   Brief 2 ;
SUBC>   Fits 'FITS2'.

General Linear Model: YIELD versus BLOCKS, STRAINS

Factor   Type   Levels  Values
BLOCKS   fixed       5  1, 2, 3, 4, 5
STRAINS  fixed       4  A, B, C, D

Analysis of Variance for YIELD, using Adjusted SS for Tests

Source   DF   Seq SS   Adj SS  Adj MS      F      P
BLOCKS    4   21.462   21.462   5.365   2.45  0.103
STRAINS   3  134.448  134.448  44.816  20.48  0.000
Error    12   26.262   26.262   2.188
Total    19  182.172

MTB > let c10=c9*c9

MTB > Name c12 "FITS4"
MTB > GLM 'YIELD' = BLOCKS STRAINS u;
SUBC>   Covariates 'u';
SUBC>   Brief 2 ;
SUBC>   Fits 'FITS4'.

General Linear Model: YIELD versus BLOCKS, STRAINS

Factor   Type   Levels  Values
BLOCKS   fixed       5  1, 2, 3, 4, 5
STRAINS  fixed       4  A, B, C, D

Analysis of Variance for YIELD, using Adjusted SS for Tests

Source   DF   Seq SS  Adj SS  Adj MS     F      P
BLOCKS    4   21.462   0.496   0.124  0.05  0.994
STRAINS   3  134.448   0.501   0.167  0.07  0.973
u         1    1.151   1.151   1.151  0.50  0.493
Error    11   25.111  25.111   2.283
Total    19  182.172

S = 1.51091   R-Sq = 86.22%   R-Sq(adj) = 76.19%

Term         Coef  SE Coef      T      P
Constant   -15.63    68.24  -0.23  0.823
u         0.04465  0.06289   0.71  0.493
```

11.6 Departures from Assumptions in Analysis of Variance

Apart from those mentioned in Chap. 10, for the randomized complete block design, we also assume the following:

(i) That the treatment effects are additive.
(ii) That the blocks and treatments effects are additive.
(iii) That these errors all have a constant variance σ^2 (homogeneity).
(iv) That the treatment and block effects are fixed.

For data following non-normal distribution, as for instance, in the skewed distributions, the variance tends to be a function of the mean. The usual approach in combating this problem is to initially transform the data by using an appropriate transformation. Conclusions of the analysis of variance also apply to the transformed data. Several transformations have been suggested for various kinds of data. Some of these are mentioned below.

11.6.1 Square Root Transformation

This is suitable when the observations follow a Poisson distribution. Examples are counts or number of nematodes in a soil, number of stem borers in a corn plot after the application of pesticides, number of weeds found on a plot, number of seeds per plot, number of infected plants in a plot, or the number of 'yes' or '1' in a binary response variable. Most work on plant protection and particularly entomology and nematology always give rise to data that are of this kind. The appropriate transformation is therefore given by

$Y_{ij}^* = \sqrt{Y_{ij}}$ or

$Y_{ij}^* = \sqrt{(Y_{ij} + 0.5)}$ if data contains some zeros (Bartlett 1936)

$Y_{ij}^* = \sqrt{Y_{ij}} + \sqrt{(Y_{ij} + 1)}$ for a better result (Freeman and Tukey 1950).

The data for the square root transformation assume that the means are proportional to the variances for each treatment, i.e., when $\sigma_i^2 = k\mu_i$, then the square root transformation will be appropriate. The transformation ensures that the variance of the data is as nearly independent of the mean as much as possible. If there are zeros in the data, other suggested form of transformations are $Y_{ij}^* = \sqrt{(Y_{ij} + 1)}$ or $Y_{ij}^* = \sqrt{\left(Y_{ij} + \frac{3}{8}\right)}$.

11.6.2 The Logarithmic Transformation

If the variance of the data is proportional to the square of its mean, i.e, if,

$$\sigma_i^2 = k\mu_i^2$$

then, the logarithmic transformation will be appropriate in this case, i.e.,

$$Y_{ij}^* = \log(Y_{ij}).$$

The above transformations are variance stabilizing transformations and they do not necessarily induce normality. For example, while the square root transformation ($y^* \longrightarrow \sqrt{y}$) stabilizes the variance of the Poisson distributed data (Y), however, the normalizing transformation is ($y^* \longrightarrow y^{2/3}$). If there are zeros in the data, possible transformations are $\log(Y_{ij}+1)$, $\log(2Y_{ij}+1)$, or $\log\left(Y_{ij}+\frac{3}{8}\right)$.

11.6.3 Arc Sine Transformation

For data that came as counts or binomial proportions (p), the arc sine or angular transformation is most appropriate. Examples of the application of this is on data relating to percentages (derived from count data rather than % of protein content for example). For binomial data, expressed as fractions, the arc sine transformation is

$$\begin{aligned} Y_{ij}^* &= \text{arc sine } Y_{ij} \\ &= \sin^{-1} Y_{ij} \end{aligned}$$

However, for percentages 0 % and 100 %, the above transformation is not defined and in these cases, it has been suggested that these two values should be substituted with the following before transformation:

$$Y_{ij} = \begin{cases} \dfrac{1}{4n} & \text{if } Y_{ij} = 0 \\ (100 - \dfrac{1}{4n}) & \text{if } Y_{ij} = 100 \end{cases} \tag{11.6}$$

Examples 11.5.1

In an experiment on weed control in Red Beet, emergence counts were taken for an area of 3 square feet in each plot. The data and analysis are given below in Table 11.19.

Table 11.19 Emergence counts from a weed control experiment

Treatment	Blocks				Total
	I	II	III	IV	
1	2	7	12	18	39
2	4	12	23	22	61
3	29	61	56	64	210
4	44	61	85	94	284
5	30	62	71	93	256
Total	109	203	247	291	850

The usual analysis of variance on the above data would yield the analysis of variance in Table 11.20

Table 11.20 Analysis of variance table for the untransformed data

Source	d.f.	SS	MS	F
Blocks	3	3631	1210.3	
Treatments	4	1,2758	3189.6	34.0
Error	12	1126	93.8	
Total	19	1,7515		

The MINITAB implementation of the above analysis on the untransformed data is presented below. A plot of the residuals versus the fitted values $\hat{y}$ is also presented in Fig. 11.1. The plot clearly indicate that the errors are not randomly distributed. Thus a transformation of the data before analysis would be required. It should be noted here that a test for normality of the residuals reveal that the residuals satisfy the normality assumption, the Anderson-Darling test give a p-value of 0.247 which confirms normality.

```
MTB > set c1
DATA> (1:5)4
DATA> end
MTB > set c2
DATA> 5(1:4)
DATA> end
MTB > SET C3
DATA> 2 7 12 18 4 12 23 22
DATA> 29 61 56 64 44 61 85
DATA> 94 30 62 71 93
DATA> END

MTB > Name c4 = 'RESI1'
MTB > GLM 'COUNTS' = BLOCK TRT;
SUBC>   SSquares 1;
SUBC>   Brief 1 ;
SUBC>   Residuals 'RESI1';
SUBC>  GNormalplot;
SUBC>  GFits;
SUBC>  RType 1 .

General Linear Model: COUNTS versus BLOCK, TRT
```

```
Factor     Type Levels Values
BLOCK     fixed      4 1 2 3 4
TRT       fixed      5 1 2 3 4 5

Analysis of Variance for COUNTS, using Sequential SS for Tests

Source     DF     Seq SS     Adj SS     Seq MS        F       P
BLOCK       3     3631.0     3631.0     1210.3   12.90   0.000
TRT         4    12758.5    12758.5     3189.6   34.01   0.000
Error      12     1125.5     1125.5       93.8
Total      19    17515.0
```

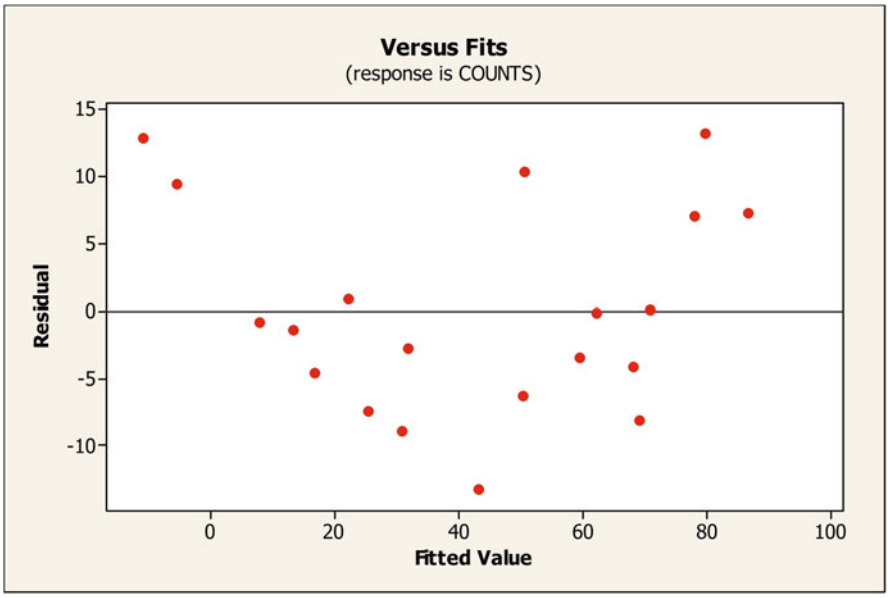


Fig. 11.1 Plot of residuals against $\hat{y}$

It is often recommended in situations of this kind (counts), that the data should be transformed to square root before analysis.

We may ask the question—why do we always use the square root transformation for this type of data? Why not the logarithmic transformation for instance? In other to make a quick check on the usefulness of either the square root or log transformations for these data, we find the range of values for each treatment and examine the variation of range with mean or square root of mean.

	Treatment means				
	9.75	15.25	52.50	71.0	64.0
Range	16	19	35	50	63
Range/Mean	1.64	1.25	0.67	0.70	0.98
Range/$\sqrt{\text{Mean}}$	5.1	4.9	4.8	5.9	7.9

Range increases steadily with the treatment mean but it is not proportional to the mean (which would suggest the log transformation); except for the last treatment, the value of Range/$\sqrt{\text{Mean}}$ are very consistent and the square root transformation should be used before analysis. Table 11.21 contains the transformed data for the data in Table 11.19.

Table 11.21 Square root of counts

	Blocks				
Treatment	I	II	III	IV	Total
1	1.4	2.6	3.5	4.2	11.7
2	2.0	3.5	4.8	4.7	15.0
3	5.4	7.8	7.5	8.0	28.7
4	6.6	7.8	9.2	9.7	33.3
5	5.5	7.9	8.4	9.6	31.4
Total	20.9	29.6	33.4	36.2	120.1

Note that the ranges of the transformed data, 2.8, 2.8, 2.6, 3.1 and 4.1 are very much consistent than those of the untransformed data.

Table 11.22 Analysis of variance table for the transformed data

Source	d.f.	SS	MS	F
Blocks	3	20.59	6.863	
Treatments	4	98.91	24.728	144.61***
Error	12	2.05	0.171	
Total	19	127.55		

*** Significant at $\alpha = 0.001$

As would be hoped, the transformation to a scale satisfying the assumptions of the analysis has produced more clear-cut results, the Error SS being reduced from 6.4 % of the total SS to 1.6 %, and the treatment/Error F ratio being correspondingly increased. Using the transformed values, differences between treatments 1 and 2 and between 3 and 5 are significant at the 5 % level, whereas previously they were not. Table 11.23 gives the means of the treatments for both the transformed and untransformed data, together with their accompanying standard errors.

Table 11.23 Treatment means and corresponding standard errors

Treatment	Untransformed	Transformed
1	9.8	2.92
2	15.2	3.75
3	52.5	7.18
4	71.0	8.32
5	64.0	7.85
S.E. of means	4.84	0.207
S.E. of difference between means	6.85	0.292

Below we present the MINITAB analysis of the transformed data.

```
MTB > LET C4=SQRT(C3)
MTB > PRINT C1-C4

Data Display

 Row   TRT  BLOCK  COUNTS        Y

   1     1      1       2  1.41421
   2     1      2       7  2.64575
   3     1      3      12  3.46410
   4     1      4      18  4.24264
   5     2      1       4  2.00000
   6     2      2      12  3.46410
   7     2      3      23  4.79583
   8     2      4      22  4.69042
   9     3      1      29  5.38516
  10     3      2      61  7.81025
  11     3      3      56  7.48331
  12     3      4      64  8.00000
  13     4      1      44  6.63325
  14     4      2      61  7.81025
  15     4      3      85  9.21954
  16     4      4      94  9.69536
  17     5      1      30  5.47723
  18     5      2      62  7.87401
  19     5      3      71  8.42615
  20     5      4      93  9.64365


MTB > GLM 'Y' = BLOCK TRT;
SUBC>   SSquares 1;
SUBC>   Brief 1 ;
SUBC>   Means TRT;
SUBC>   Pairwise TRT;
SUBC>     Tukey;
SUBC>     NoCI.

General Linear Model: Y versus BLOCK, TRT


Factor     Type Levels Values
BLOCK     fixed      4 1 2 3 4
TRT       fixed      5 1 2 3 4 5

Analysis of Variance for Y, using Sequential SS for Tests

Source     DF     Seq SS     Adj SS     Seq MS       F      P
BLOCK       3     26.721     26.721      8.907   49.63  0.000
TRT         4     99.021     99.021     24.755  137.93  0.000
Error      12      2.154      2.154      0.179
Total      19    127.896

Least Squares Means for Y
```

```
TRT       Mean    SE Mean
1        2.942     0.2118
2        3.738     0.2118
3        7.170     0.2118
4        8.340     0.2118
5        7.855     0.2118


Tukey Simultaneous Tests
Response Variable Y
All Pairwise Comparisons among Levels of TRT

TRT = 1 subtracted from:

Level   Difference      SE of              Adjusted
TRT      of Means  Difference    T-Value     P-Value
2          0.7959      0.2996      2.657      0.1207
3          4.2280      0.2996     14.114      0.0000
4          5.3979      0.2996     18.019      0.0000
5          4.9136      0.2996     16.402      0.0000

TRT = 2 subtracted from:

Level   Difference      SE of              Adjusted
TRT      of Means  Difference    T-Value     P-Value
3           3.432      0.2996      11.46      0.0000
4           4.602      0.2996      15.36      0.0000
5           4.118      0.2996      13.75      0.0000

TRT = 3 subtracted from:

Level   Difference      SE of              Adjusted
TRT      of Means  Difference    T-Value     P-Value
4          1.1699      0.2996      3.905      0.0146
5          0.6856      0.2996      2.289      0.2139

TRT = 4 subtracted from:
Level   Difference      SE of              Adjusted
TRT      of Means  Difference    T-Value     P-Value
5         -0.4843      0.2996     -1.617      0.5148
```

The results of Tukey's test on the transformed data is presented in the following:

2.942	3.738	7.170	7.855	8.340
1	2	3	5	4

While treatments 3, 4, and 5 might not be significantly different, clearly they are different from treatments 1 and 2, who themselves are not significantly different.

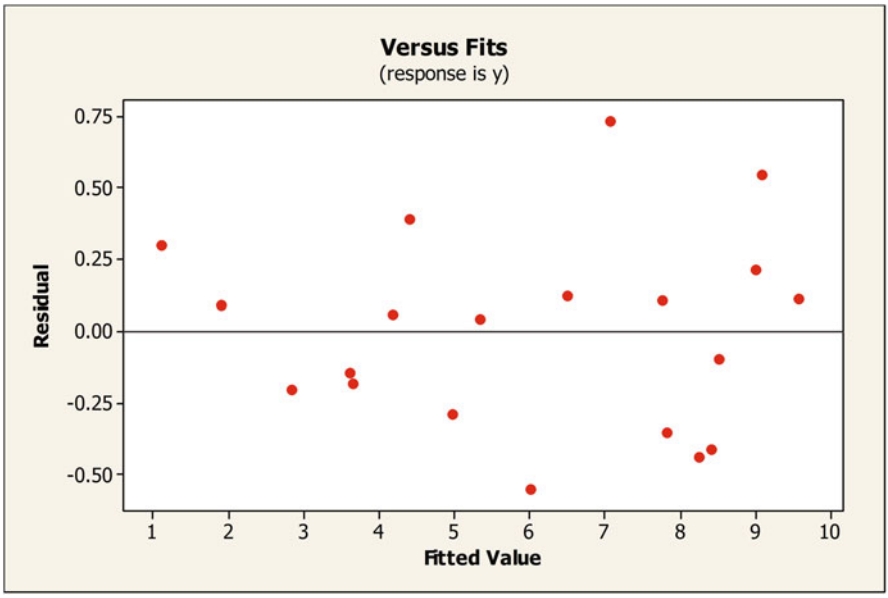


Fig. 11.2 Plot of residuals against $\hat{y}$

The plot of the residuals against the fitted values as displayed in Fig. 11.2 which now shows a random pattern for the residuals satisfying the assumption of random distribution of the error terms around zero. Further, the corresponding normality plot of the residuals in Fig. 11.3 indicates that the assumption of normality is strongly justified with the transformed data, with p-value from the Anderson-Darling test being in this case 0.863.

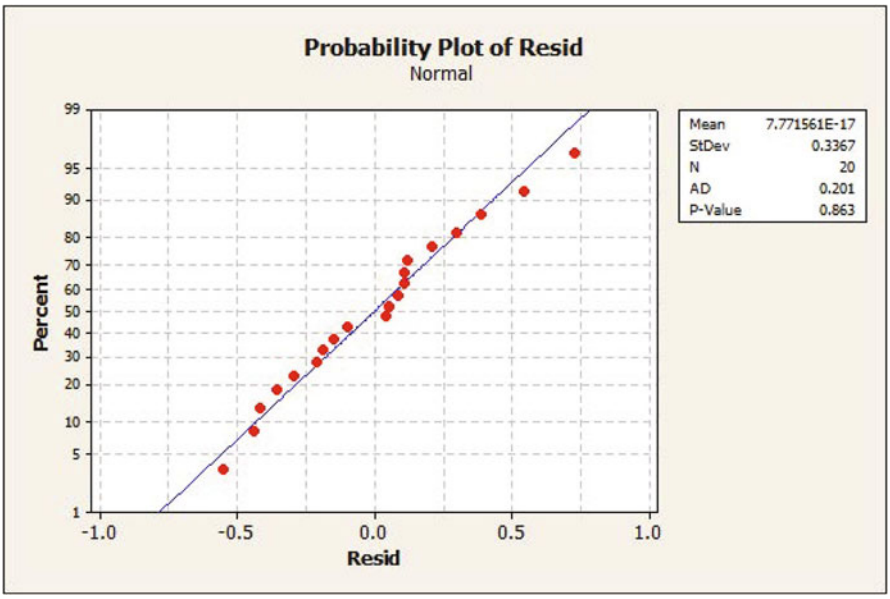


Fig. 11.3 Normality test and plot of the residuals

11.6.4 Box-Cox Transformation

The Box-Cox (1964) approach identifies a suitable transformation from the family of power transformations of the form

$$y* = y^{\lambda}, \tag{11.7}$$

where λ is to be estimated from the data. Possible values of λ are presented in the table below. For instance, a λ value of 0.5 corresponds to a transformation of $Y^* = \sqrt{Y}$ or if $\lambda = 0$ then $Y^* = \ln Y$.

λ	Y^*
2	Y^2
1	Y
0.5	$\sqrt{Y}$
0	$\ln Y$
-0.5	$\frac{1}{\sqrt{Y}}$
-1	$\frac{1}{Y}$

To determine λ, a regression of the following form is carried out.

$$y_i^{\lambda} = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{11.8}$$

where $\beta_0, \beta_1, \lambda$ and σ^2, the variance of ε are to be estimated from available data. One procedure to accomplish this is to use a numerical search for values

of λ ranging from -3 to 3. It has been suggested that we use the following:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{(y_i^{\lambda} - 1)}{\lambda K^{\lambda - 1}} & \lambda \neq 0 \\ K \ln(y_i), & \lambda = 0, \end{cases} \tag{11.9}$$

where Y is the original data, and,

$$K = \left(\prod_{i=1}^{n} y_i \right)^{1/n} = \exp \left[\frac{1}{n} \sum_{i=1}^{n} \ln(y_i) \right], \tag{11.10}$$

K here, is the geometric mean of the y_i observations. We assume that the transformed values $y_i^{(\lambda)}$ follow a normal linear model with parameters β and σ^2 for some values of λ.

MINITAB calculates the standard deviation and plots the graph of standard deviation (SD) versus λ. We are thus looking for the λ value that optimizes the log-likelihood profile. Figure 11.4 gives this plot as generated by MINITAB and unexpectedly, $\lambda = 0.5$ is selected and this further confirms our earlier choice of the square root transformation. MINITAB also computes the transformed variable automatically based on the choice of lambda and stores these values in a new column for further analysis.

```
MTB > BoxCox 'COUNTS' 2;
SUBC>   Store c4.
```

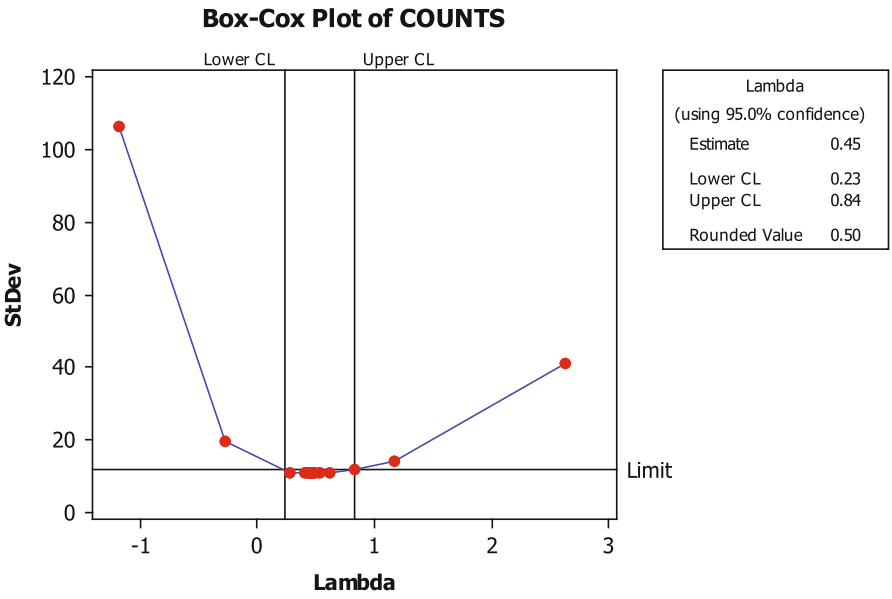


Fig. 11.4 Boxplot for the data in table 11.23

It must be emphasized again, however, that the justification for the transformation lies not in the 'better' results achieved but rather in the validity of the analysis in the transformed scale.

11.7 Relative Efficiency of RCBD

The relative efficiency of a randomized complete block design over that of the completely randomized design is defined as:

$$RE(RCBD, CRD) = \frac{(b-1)MSB + b(t-1)MSE}{(bt-1)MSE}$$

where MSE is the error mean square under RCBD and MSB is the mean square for blocks in the RCBD. For example, the relative efficiency for the data in example 11.1 (no 307) will be computed as: $b = 5, t = 4, MSE =$ 2.188, and MSB = 5.365.

$$\frac{4(5.365) + (5 \times 3)2.188}{19 \times 2.188} = 1.31$$

i.e., approximately 30 % times more observations of each treatment will be required in a completely randomized design (CRD) to obtain the same precision for the estimating treatment means as with the RCB design. Thus the RCDB is 30 % more efficient than the CRD.

11.8 Group Balanced Block Design

In variety or similar trials, the researcher knows that some varieties behave differently in terms of days to maturity, days to flowering or plant heights. For instance, plants with different heights grown adjacent to one another in experimental plots will be subjected to plant competition or plants with different maturity days planted adjacent to each other are bound to create harvesting problems in addition to various competition effects. To mitigate against these undesirable effects, researchers often group together certain varieties that have similar characteristics and other traits.

The group balanced design therefore ensures that varieties or treatments with the same characteristics are grouped together such that varieties in the same group are always arranged and tested in the same block, thus treatments or varieties belonging to different groups are never tested together in the same block. Suppose we have fifteen varieties of sorghum divided into three groups of five treatments each and designated as groups A, B, and C. The grouping is based on the maturity days of the plants based on less than 100 days,

between 100 and 120 days and more than 120 days. The experiment is to be laid out in four replicates. A typical layout is presented below:

	REP I	REP II	REP III	REP IV
Block 1	Group A	Group C	Group A	Group B
Block 2	Group B	Group A	Group C	Group A
Block 3	Group C	Group B	Group B	Group C

Here, we have grouped the $t = 15$ treatments into $g = 3$ with each group having $\frac{t}{g} = 5$ treatments. These groupings are A with treatments {1,2,3,4,5}, B with treatments {6,7,8,9,10}, and C with treatments {11, 12, 13, 14, 15}. The experimental area here has $rt = 4 \times 15 = 60$ plots such that each replication has a total of t plots. Each replication is then divided into $g = 3$ blocks each of $t/g = 5$ units.

11.8.1 Randomization

The randomization scheme could be as described here:

1. Randomize the g groups into the g blocks as in REP I above. Then for each of the remaining replications, independently assign at random the groups to the blocks. The layout should be as displayed above.
2. For each block in each replicate, we assign at random the treatments corresponding to the groupings as indicated in the layout above. The resulting randomization will be as displayed below. For instance in replicate I, group A treatments are randomized within block 1, while group B treatments are randomized within block 2. The group C treatments are also randomized within block 3 in replicate I. Similar randomization of treatments within blocks are carried out as led out in the earlier randomization scheme of groups to blocks. The resulting lay out is presented below.

Block 1	1	2	3	4	5
Block 2	8	6	7	9	10
Block 3	11	14	13	15	12

REP I

Block 1	12	14	11	13	15
Block 2	3	1	4	5	2
Block 3	9	7	10	8	6

REP II

Block 1	3	5	1	4	2
Block 2	14	11	12	15	13
Block 3	7	9	6	8	10

REP III

Block 1	8	10	7	9	6
Block 2	4	1	3	2	5
Block 3	15	11	13	12	14

REP IV

The structure of the analysis of variance for this design is presented in Table 11.24

Table 11.24 Structure of the ANOVA table

Source	d.f.
Reps	$r-1$
Treatment groups	$g-1$
Error (a)	$(r-1)(g-1)$
Treat within group 1	$t/g-1$
Treat within group 2	$t/g-1$
$\vdots$	$\vdots$
Treat within group g	$t/g-1$
Error (b)	$g(t-1)\left(\frac{t}{g}-1\right)$
Total	$rt-1$

11.8.2 An Example

An experiment to determine the yield per plot of 15 varieties of sorghum was conducted with four replications and three groupings of the varieties based on the commencement of their flowering characteristics. The yield from the experiment which consists of 60 plots of 15 replication per plot and three blocks of unit five per replicated is presented in Table 11.25.
The Total SS is computed as:

$$\text{Total SS} = [(3.544)^2 + (2.870)^2) + \cdots + (3.641)^2] - \frac{(199.658)^2}{60} = 4.41514$$

Table 11.25 Yield of sorghum in kg/plot of 15 varieties

Trt	Yield kg/plot				Total
	RepI	RepII	RepIII	RepIV	
1	3.544	2.870	3.318	3.171	12.903
2	3.215	2.935	3.128	3.611	12.889
3	3.628	3.078	3.200	3.326	13.232
4	3.152	3.102	2.875	3.222	12.351
5	3.550	3.286	2.786	3.343	12.965
6	2.878	3.054	3.461	3.244	12.637
7	3.171	3.026	3.283	3.641	13.121
8	3.471	3.220	3.865	3.202	13.758
9	3.126	3.051	3.126	3.183	12.486
10	3.260	3.119	3.333	3.388	13.100
11	3.383	2.866	3.551	3.495	13.295
12	3.232	3.198	3.162	3.403	12.995
13	3.774	3.461	3.596	3.723	14.554
14	3.782	3.867	3.589	3.539	14.777
15	3.538	3.473	3.943	3.641	14.595
Total	50.704	47.606	50.216	51.132	199.658

$$\text{REP SS} = \frac{(50.704)^2}{15} + \frac{(47.606)^2}{15} + \frac{(50.216)^2}{15} + \frac{(51.132)^2}{15} - \frac{(199.658)^2}{60}$$
$$= 0.50171$$

Group	Yield total				Group total
	REP I	REP II	REP III	REP IV	
A	17.089	15.271	15.307	16.673	64.34
B	15.906	15.470	17.068	16.658	65.102
C	17.709	16.865	17.841	17.801	70.216

Replication by group table of yields

With the above replication by group table of yields, we can now obtain the following:

$$\text{Groups SS} = \frac{(64.34)^2}{20} + \frac{(65.102)^2}{20} + \frac{(70.216)^2}{20} - \frac{(199.658)^2}{60} = 1.02102$$

$$\text{Error (a) SS} = \left[\frac{(17.089)^2}{5} + \frac{(15.271)^2}{5} + \cdots + \frac{(17.801)^2}{5}\right] - \frac{(199.658)^2}{60}$$
$$- \text{Groups SS} - \text{Reps SS}.$$
$$= 0.46298$$

We present below the total yield for each of the grouping treatments. Hence we can now calculate the sum of squares of treatments within each group.

Total yield for group A				
1	2	3	4	5
12.903	12.889	13.232	12.351	12.965
Total yield for group B				
6	7	8	9	10
12.637	13.121	13.758	12.486	13.100
Total yield for group C				
11	12	13	14	15
13.295	12.995	14.554	14.777	14.595

These SS are computed as:

$$\text{Trt SS within GP A} = \left[\frac{(12.903)^2}{4} + \frac{(12.889)^4}{4} + \cdots + \frac{(12.965)^2}{4}\right] - \frac{(64.340)^2}{20}$$
$$= 0.10272$$

$$\text{Trt SS within GP B} = \left[\frac{(12.637)^2}{4} + \frac{(13.121)^4}{4} + \cdots + \frac{(13.100)^2}{4}\right] - \frac{(65.102)^2}{20}$$
$$= 0.248272$$

$$\text{Trt SS within GP C} = \left[\frac{(13.295)^2}{4} + \frac{(12.995)^4}{4} + \cdots + \frac{(14.595)^2}{4}\right] - \frac{(70.216)^2}{20}$$
$$= 0.69060$$

The above computations therefore lead to the following analysis of variance table for the experiment:

Table 11.26 ANOVA for group balanced design example

Source	d.f.	SS	MS	F
Reps	3	0.5017	0.1672	
Groups	2	1.0210	0.5105	6.61
Error (a)	6	0.4630	0.0772	
Trt (within Groups)	12	1.0416	0.0868	2.25
Trt within GP A	4	0.1027	0.0257	0.67
Trt within GP B	4	0.2483	0.0621	1.61
Trt within GP C	4	0.6906	0.1727	4.47
Error (b)	36	1.13879	0.0386	
Total	59	4.4151		

Results from ANOVA Table 11.26 indicate significant differences among the group means and significant differences only among the treatments within group C. The others did not show any significant differences within groups A and B.

The Above analysis is better analyzed with MINITAB. We present the data structure within MINITAB and the subsequent ANOVA Table obtained. Based on the structure of the ANOVA Table in Table 11.24, the ANOVA Table for the above data in Table 11.25 using MINITAB are displayed below. The data layout presents the data for the first and fourth replicates I and IV. Column 3 gives the replication numbers, column 4 the grouping number and column 6 the treatment numbers.

```
Row    Y    REP  GP  TRT
  1  3.544    1   1    1
  2  3.215    1   1    2
  3  3.628    1   1    3
  4  3.152    1   1    4
  5  3.550    1   1    5
  6  2.878    1   2    6
  7  3.171    1   2    7
  8  3.471    1   2    8
  9  3.126    1   2    9
 10  3.260    1   2   10
 11  3.383    1   3   11
 12  3.232    1   3   12
 13  3.774    1   3   13
 14  3.782    1   3   14
 15  3.538    1   3   15
 ......................
 46  3.171    4   1    1
 47  3.611    4   1    2
 48  3.326    4   1    3
 49  3.222    4   1    4
 50  3.343    4   1    5
 51  3.244    4   2    6
 52  3.641    4   2    7
 53  3.202    4   2    8
 54  3.183    4   2    9
 55  3.388    4   2   10
 56  3.495    4   3   11
 57  3.403    4   3   12
 58  3.723    4   3   13
 59  3.539    4   3   14
 60  3.641    4   3   15
MTB > GLM 'Y' = REP GP REP* GP TRT( GP);
SUBC>   Brief 2 .

General Linear Model: Y versus REP, GP, TRT

Factor   Type   Levels  Values
REP      fixed       4  1, 2, 3, 4
GP       fixed       3  1, 2, 3
TRT(GP)  fixed      15  1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15


Analysis of Variance for Y, using Adjusted SS for Tests

Source   DF   Seq SS   Adj SS   Adj MS      F      P
REP       3  0.50171  0.50171  0.16724   4.34  0.010
GP        2  1.02102  1.02102  0.51051  13.24  0.000
REP*GP    6  0.46298  0.46298  0.07716   2.00  0.091
TRT(GP)  12  1.04158  1.04158  0.08680   2.25  0.030
Error    36  1.38785  1.38785  0.03855
Total    59  4.41514


S = 0.196345   R-Sq = 68.57%   R-Sq(adj) = 48.48%
```

We may note here that MINITAB does not use the appropriate Error (a) MS to test for the group means significance. However, once such an output is obtained, it is not too difficult to manually conduct such a test ourselves. The results from MINITAB agree with the ones we calculated.

11.9 Exercises

1. An experiment involving four treatments A, B, C, and D are applied to rats that are drawn from the same litter. Ten litters are employed. The data is presented below.

Litter number	Rat number & treatment number			
1	1-B	2-A	3-D	4-C
2	5-B	6-C	7-A	8-D
3	9-C	10-A	11-B	12-D
4	13-A	14-B	15-D	16-C
5	17-D	18-C	19-A	20-B
6	21-D	22-C	23-A	24-B
7	25-B	26-A	27-D	28-C
8	29-C	30-B	31-A	32-D
9	33-D	34-C	35-A	36-B
10	37-D	38-A	39-C	40-B

   Analyze the data and draw your conclusions. What is the relative efficiency of this design?
2. The data below is adapted from Gomez and Gomez (1983) and relate to yield in kg/ha of six treatments on rice.

Treatment level	Rep I	Rep II	Rep III	Rep IV
25	5113	5398	5307	4678
50	5346	5952	4719	4264
75	5272	5713	5483	4749
100	5164	4831	4986	4410
125	4804	4848	4432	4748
150	5254	4542	4919	4098

   (i) Analyze the data as an RCB design treating the replicates as blocks
   (ii) If the treatment effects are significant, consider fitting an appropriate polynomial to the response
   (iii) Suppose treatment corresponding to treatment level 100 in replicate II were missing, estimate this missing value and run a different ANOVA Table

3. The partially completed ANOVA table for an experiment is shown below:

ANOVA table				
Source	d.f.	SS	MS	F
Blocks	7	–	14.0714	–
Treatments	–	231.5054	115.7527	–
Error	–	573.7500	40.9821	
Total	23	–		

a What design was employed?
How many treatments are involved in the experiment? What is the total sample size?
b Conduct a test of the null hypothesis that the treatment means are equal. Use $\alpha = 0.05$.
c What assumptions must be satisfied before the analysis above can be valid? (state these only).
d Compute the standard error for comparing any two treatment means.

4. Complete the following ANOVA table and state which design was used.

Source	d.f.	SS	MS	F
Blocks	2	177.7	–	–
Treatments	5	–	–	–
Error	–	448.99	–	
Total	–	2338.69		

5. Complete the following ANOVA table and state which design was used.

Source	d.f	SS	MS	F
Blocks	7	40.40	–	–
Treatments	2	–	–	–
Error	–	9.29	–	
Total	23	65.95		

Chapter 12
Multiple Blocking Designs

12.1 The Latin or Euler Square Design

The Latin square is a plan of t rows and t columns of a square with t symbols arranged such that each letter appears once in each row and once in each column. If the symbols are Latin letters we could, as Sir Ronald A. Fisher did, call this a Latin square plan or design. If the symbols used were Greek letters, we could call the plan a Greek square. If the symbols used were Arabic symbols, we could call the plan an Arabic square, and so on. By common usage this plan is used with Latin letters and is called a Latin square design. Furthermore, this design is for the removal of variations from two sources usually referred to, though not necessarily, the row and column variations. The row and column designation merely refers to the two sources.

Example 12.1.1

To illustrate, suppose a study to compare the tolerances (specified response to a certain amount of dose stimulus) of cats to three substances (A, B, C) is to be conducted. If nine cats are available for the study, the substances can be applied to the nine cats as follows:

C	B	A
B	A	C
A	C	B

The above design controls variations in two directions. Now suppose that nine cats are not available but only three cats are available. We can then administer the substances in the following order:

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_12,

	Cats								
	1			2			3		
	order			order			order		
	1	2	3	1	2	3	1	2	3
Substances	C	B	A	B	A	C	A	C	B

which when rearranged "looks like a square".

	Cats		
Order	1	2	3
1	C	B	A
2	B	A	C
3	A	C	B

Each treatment appears once in each order and once each for the cats.

Example 12.1.2

Pathologists studying tobacco mosaic virus in order to compare the toxicity of different solutions would smear the virus solutions over the leaves of tobacco plants. In three or four days, little spots came out on the leaves, the stronger the virus, the more spots. To compare the solutions then, they would smear them on the leaves and count the spots.

For some reason or other, most of these tobacco plants were grown to the point where they had about five leaves. By smearing the same solution on all leaves, for several plants, it was revealed immediately that there were certain natural groupings. The leaves from the same plant, as might well be expected, have a common quality of susceptibility to the production of spots. Another plant would be resistant; all the leaves on that would give smaller counts. The total count on the five leaves of one plant might be one-fifth or one-third what it was on another plant. But even more striking was the fact that there was a positional effect. The top leaves tended to be alike (as did the second, the third, the fourth, and the bottom leaves) in the sense that all the top leaves might give about half the count of their corresponding bottom leaves from the same plant.

Here, then, you see the familiar rows and columns made to order. Nothing hopeful about this regularity, it is there. To compare five virus solutions, we will simply make sure, if we label them A, B, C, D, E, that they are allotted to the leaves in the same kind of pattern we had a moment ago for the cats' experiment

	Plant Number				
Leaf Position	1	2	3	4	5
Top	A	B	C	D	E
2nd	B	E	D	C	A
3rd	C	A	E	B	D
4th	D	C	A	E	B
5th	E	D	B	A	C

Note the arrangement of the five letters: all five on every plant, all five in each leaf position. The net result of this was to improve the accuracy of the comparison, that it is quite conservative to say it was like presenting the pathologist with an extra greenhouse. He did not need to test as many plants with each solution. This was a case where these strips really paid off. Cox and Cochran (1946) described an experiment similar to the one above for the comparison of five virus inoculations of plants. The plot was a single leaf, and the two blocking systems were plants and leaf sizes. Five plants were taken with five leaves on each plant. The design is similar to the one above, in which the columns were the plants and the rows were the five largest leaves, the second five largest leaves, and so on. The treatment is represented by letters, have been allocated in such a way that one leaf of each plant has each treatment and, of the five leaves receiving a particular treatment, one is the largest on its plant, and one is the second largest, and so on.

12.2 The Model for the Latin Square Designs

The Latin square differs from the randomized block design in that the treatments are arranged in complete groups in two directions, the two classifications being orthogonal to each other and to the treatments.

Thus, for a $t \times t$ Latin square, the LS design has the linear model given by

$$Y_{ijk} = \mu + r_i + c_j + t_k + e_{ijk} \quad \begin{matrix}(i,j) = 1, 2, \cdots, t \\ k \quad = 1, 2, \cdots, t\end{matrix} \tag{12.1}$$

where

- μ is the overall mean;
- r_i is the ith row effect;
- c_j is the jth column effect;
- t_k is the kth treatment effect; and
- ϵ_{ijk} is the random error term distributed normal with mean zero and constant variance σ^2.

All the above effects act additively, and there is no interaction between any two of the three factors (rows, columns, and treatments). We also know that in a Latin square arrangement, each row contains a unit from each column, and vice versa. Hence, rows and columns are orthogonal. Further, each treatment appears in each row once and each row contains each of the treatments once; hence, treatments and rows are orthogonal. Similarly,

columns and treatments are also orthogonal. In short, a Latin square is a set up in which the three factors are mutually orthogonal

If we denote by $R_1, R_2, \cdots, R_t$ the row totals, $C_1, C_2, \cdots, C_t$ the column totals and $T_1, T_2, \cdots, T_t$ the treatment totals in an $t \times t$ table, then, the Analysis of Variance is of the form:

$$\text{Correction factor} = \frac{Y_{++}^2}{t^2}$$

$$\text{Total SS} = y_{11}^2 + y_{12}^2 + \cdots + y_{tt}^2 - \text{CF} \quad \text{with } (t^2 - 1) \text{ d.f.}$$

$$\text{Rows SS} = \frac{R_1^2 + R_2^2 + \cdots + R_t^2}{t} - \text{CF} \quad \text{with } (t - 1) \text{ d.f.}$$

$$\text{Column SS} = \frac{C_1^2 + C_2^2 + \cdots + C_t^2}{t} - \text{CF} \quad \text{with } (t - 1) \text{ d.f.}$$

$$\text{Treatments SS} = \frac{T_1^2 + T_2^2 + \cdots + T_t^2}{t} - \text{CF} \quad \text{with } (t - 1) \text{ d.f.}$$

The error sums of squares is obtained by subtraction as

$$\text{Error SS} = \text{Total SS} - \text{Row SS} - \text{Column SS} - \text{Treatments SS}$$

$$\text{with } (t - 1)(t - 2) \text{ d.f.}$$

For the 2×2, 3×3 and 4×4 Latin squares, there are 0, 2 and 6 degrees of freedom associated with the error sum of squares, and with such few degrees of freedom, in the error term, it is recommended that the Latin square be repeated or another design used. Since the Latin square design requires as many replicates as treatments, the design is seldom used for more than 10–12 treatments.

The main advantages of the Latin square is that with two-way stratification or grouping, it controls more of the variation than the CRD or the RCBD. The two-way elimination of variation often results in a smaller error mean square, further the analysis is simple and remains so even with missing data as analytical procedures are available for omitting one or more treatments, rows or columns.

The structure of ANOVA Table in Latin Square design is displayed as:

Structure of ANOVA

Source	d.f.
Rows	$t - 1$
Cols	$t - 1$
Trt	$t - 1$
Error	$(t - 1)(t - 2)$
Total	$(t^2 - 1)$

Example 12.2.1

In a digestion trial carried out with six shorthorn steers, each animal received each of six rations in six successive periods, the experimental design being

a Latin square. Coefficient of digestibility of nitrogen were calculated for each animal for each period as follows in Table 12.1 (rations are indicated in brackets after each value).

12.2.1 Analysis

We notice that the treatments A, B, C, D, E, F (rations) are arranged such that the design is a Latin square because each treatment occurs once in each row (steer) and once in each column (period).Here, there are $6 \times 6 = 36$ plots.

$$\text{Total SS} = 61.1^2 + 69.3^2 + 67.6^2 + \cdots + 62.9^2 - \text{CF} = 704.75$$

Table 12.1 The 6×6 LS data for the example

	Period						
Steer	I	II	III	IV	V	VI	Totals
1	61.1(B)	69.3(D)	67.6(C)	61.9(F)	58.8(A)	65.2(E)	383.9
2	56.9(A)	59.1(F)	64.0(D)	61.0(C)	65.7(E)	56.6(B)	363.3
3	66.5(C)	62.2(A)	61.1(E)	66.2(E)	62.0(F)	62.2(D)	380.2
4	66.7(E)	67.4(B)	65.1(F)	65.1(D)	69.6(C)	52.7(A)	386.6
5	67.8(D)	64.7(C)	63.6(E)	53.2(A)	61.7(B)	62.0(F)	373.0
6	71.4(F)	67.5(E)	55.8(A)	63.2(B)	68.0(D)	62.9(C)	388.8
Totals	390.4	390.2	377.2	370.6	385.8	361.6	2275.8

The treatments (Rations) totals are presented in the following:

$$A = 58.8 + 56.9 + 62.2 + 52.7 + 53.2 + 55.8 = 339.6$$
$$B = 61.1 + 56.6 + 61.1 + 67.4 + 61.7 + 63.2 = 371.1$$
$$C = 67.6 + 61.0 + 66.5 + 69.6 + 64.7 + 62.9 = 392.3$$
$$D = 69.3 + 64.0 + 62.2 + 65.1 + 67.8 + 68.0 = 396.4$$
$$E = 65.2 + 65.7 + 66.2 + 66.7 + 63.6 + 67.5 = 394.9$$
$$F = 61.9 + 59.1 + 62.0 + 65.1 + 62.0 + 71.4 = 381.5$$

$$\text{Rows (Steer) SS} = \frac{383.9^2}{6} + \cdots + \frac{388.8^2}{6} - \text{CF} = 76.87$$

$$\text{Columns (Periods) SS} = \frac{390.4^2}{6} + \cdots + \frac{361.6^2}{6} - \text{CF} = 112.94$$

$$\text{Treatments (Rations) SS} = \frac{339.6^2 + \cdots + 381.5^2}{6} - \text{CF} = 392.16$$

$$\text{Error (by subtraction) SS} = \text{Total} - \text{Row SS} - \text{Column SS} - \text{Treatment SS}$$
$$= 122.78.$$

The analysis of variance table for the data in Table 12.1 is presented in Table 12.2.

Table 12.2 Analysis of variance table

Source	d.f.	SS	MS	F
Steers	5	76.87	15.37	2.50
Periods	5	112.94	22.59	3.68
Rations	5	392.16	78.43	12.78
Error	20	122.78	6.14	
Total	35	704.75		

F(5,20) at $\alpha = 0.05 = 2.71$. Since $12.78 > 2.71$, there are therefore significant differences between the six ration means. Here, $S^2 = 6.14$.

S.E. for comparing any two treatment means $= \sqrt{\dfrac{2S^2}{t}} = \sqrt{\dfrac{2 \times 6.14}{6}} = 1.431$.

Treatment means					
1	2	3	4	5	6
56.6	61.85	65.38	66.07	65.82	63.58

$t_{.05}$ with 20 d.f. $= 2.086$. Hence, LSD $= 2.086 \times 1.431 = 2.98$.

Certainly, Rations 3, 4, 5 and 6 are not significantly different. However, ration 1 is significantly different from ration 2 and consequently for rations 3–6. It should be noted that the LSD procedure is not encouraged when $t > 4$.

The above analysis is carried out in MINITAB with the following commands.

```
MTB > SET C1
DATA> (1:6)6
DATA> END
MTB > SET C2
DATA> 6(1:6)
DATA> END
MTB > SET C4
DATA> 61.1 69.3 67.6 61.9 58.8 65.2
DATA> 56.9 59.1 64 61 65.7 56.6
DATA> 66.5 62.2 61.1 66.2 62 62.2
DATA> 66.7 67.4 65.1 65.1 69.6 52.7
DATA> 67.8 64.7 63.6 53.2 61.7 62
DATA> 71.4 67.5 55.8 63.2 68 62.9
DATA> END
MTB > PRINT C1-C4

Data Display

 Row  STEER  PERIOD   TRT      Y

   1      1       1     B   61.1
   2      1       2     D   69.3
   3      1       3     C   67.6
   4      1       4     F   61.9
   5      1       5     A   58.8
   6      1       6     E   65.2
   7      2       1     A   56.9
   8      2       2     F   59.1
   9      2       3     D   64.0
  10      2       4     C   61.0
```

```
11      2       5     E   65.7
12      2       6     B   56.6
13      3       1     C   66.5
14      3       2     A   62.2
15      3       3     B   61.1
16      3       4     E   66.2
17      3       5     F   62.0
18      3       6     D   62.2
19      4       1     E   66.7
20      4       2     B   67.4
21      4       3     F   65.1
22      4       4     D   65.1
23      4       5     C   69.6
24      4       6     A   52.7
25      5       1     D   67.8
26      5       2     C   64.7
27      5       3     E   63.6
28      5       4     A   53.2
29      5       5     B   61.7
30      5       6     F   62.0
31      6       1     F   71.4
32      6       2     E   67.5
33      6       3     A   55.8
34      6       4     B   63.2
35      6       5     D   68.0
36      6       6     C   62.9

MTB > GLM 'Y' = STEER PERIOD TRT;
SUBC>   Brief 1 ;
SUBC>   Pairwise TRT;
SUBC>     Tukey;
SUBC>     NoCI.

General Linear Model: Y versus STEER, PERIOD, TRT

Factor     Type Levels Values
STEER     fixed      6 1 2 3 4 5 6
PERIOD    fixed      6 1 2 3 4 5 6
TRT       fixed      6 A B C D E F

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
STEER       5     76.867     76.867     15.373    2.50  0.065
PERIOD      5    112.943    112.943     22.589    3.68  0.016
TRT         5    392.157    392.157     78.431   12.78  0.000
Error      20    122.783    122.783      6.139
Total      35    704.750

Least Squares Means for Y

TRT         Mean     SE  Mean
A          56.60        1.012
B          61.85        1.012
C          65.38        1.012
D          66.07        1.012
E          65.82        1.012
F          63.58        1.012
```

The Tukey simultaneous tests (not displayed above) indicate that while treatment A is significantly different from the other treatments (B, C, D, E, F), the latter treatments are not significantly different from one another. This agrees with our earlier conclusion and the results from Tukey are presented as follows.

μ_D	μ_E	μ_C	μ_F	μ_B	μ_A
66.07	65.82	65.38	63.58	61.85	56.60

Example 12.2.2

The plan below is for a Latin square experiment to test the efficiency of the methods of dusting with sulphur in order to control stem rust of wheel.

B	D	E	A	C	A = dusted before rains
C	A	B	E	D	B = dusted after rain
D	C	A	B	E	C = dusted once each week
E	B	C	D	A	D = drifting once each week
A	E	D	C	B	E = not dusted

Drifting means that the dust was allowed to settle over the plan from above, as in airplane dusting. The plot yield in pounds per acre are given in Table 12.3.

Table 12.3 Yields in a Latin square experiment

Rows	1	2	3	4	5
1	4.9	6.4	3.3	9.5	11.8
2	9.3	4.0	6.2	5.1	5.4
3	7.6	15.4	6.5	6.0	4.6
4	5.3	7.6	13.2	8.6	4.9
5	9.3	6.3	11.8	15.9	7.6
Total	36.4	39.7	41.0	45.1	34.3

We can summarize the row and treatment totals in the following table.

Row totals	Treatment	Totals	Treatment mean
35.9	A	34.2	6.84
30.0	B	32.3	6.46
40.1	C	65.6	13.12
39.6	D	39.8	7.9
50.9	E	24.6	4.92
196.5		196.5	

Calculations of the sums of squares are as follows:

$$\text{Total CF} = \frac{G^2}{t^2} = \frac{196.5^2}{25} = 1544.49$$

$$\text{Total SS} = 1829.83 - \text{CF} = 285.34$$

$$\text{Rows SS} = \frac{35.9^2 + 30.0^2 + \cdots + 50.9^2}{5} - \text{CF} = 1591.16 - \text{CF} = 46.67$$

$$\text{Columns SS} = 36.4^2 + 39.7^2 + \cdots + 34.3^2 - \text{CF} = 1558.51 - \text{CF} = 14.02$$

$$\text{Error SS} = 122.78.$$

Hence, the analysis of variance Table is as presented in Table 12.4.

Table 12.4 Analysis of variance table for data in Table 12.3

Source	d.f.	SS	MS	F
Rows	4	44.67		
Columns	4	14.02		
Treatments	4	196.61	49.15	21.0**
Error	12	28.04	2.337	
Total	24	285.34		

** Significant at $\alpha = 0.01$

S.E. for comparing any two treatment means equals

$$\sqrt{\frac{2 \times 2.337}{5}} = 0.967.$$

With an LSD at $\alpha = 0.05 = 2.11$ and on examining the means, we see that the significance of the treatment mean square is due chiefly to treatment C which is much more effective than any of the others.
The MINITAB implementation of the analysis is presented as follows.

```
MTB > SET C1
DATA> (1:5)5
DATA> END
MTB > SET C2
DATA> 5(1:5)
DATA> END
MTB > SET C4
DATA> 4.9 6.4 3.3 9.5 11.8 9.3 4 6.2 5.1 5.4
DATA> 7.6 15.4 6.5 6 4.6 5.3 7.6 13.2 8.6 4.9
DATA> 9.3 6.3 11.8 15.9 7.6
DATA> END
MTB > GLM 'Y' = ROWS COLS TRT;
SUBC>   Brief 1 ;
SUBC>   Means TRT;
SUBC>   Pairwise TRT;
SUBC>     Tukey;
SUBC>     NoCI.

General Linear Model: Y versus ROWS, COLS, TRT


Factor     Type Levels Values
ROWS       fixed      5 1 2 3 4 5
COLS       fixed      5 1 2 3 4 5
TRT        fixed      5 A B C D E
```

Analysis of Variance for Y, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
ROWS	4	46.668	46.668	11.667	4.99	0.013
COLS	4	14.020	14.020	3.505	1.50	0.263
TRT	4	196.608	196.608	49.152	21.03	0.000
Error	12	28.044	28.044	2.337		
Total	24	285.340				

Least Squares Means for Y

TRT	Mean	SE Mean
A	6.840	0.6837
B	6.460	0.6837
C	13.120	0.6837
D	7.960	0.6837
E	4.920	0.6837

Tukey Simultaneous Tests
Response Variable Y
All Pairwise Comparisons among Levels of TRT

TRT = A subtracted from:

Level TRT	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
B	-0.380	0.9669	-0.393	0.9943
C	6.280	0.9669	6.495	0.0002
D	1.120	0.9669	1.158	0.7736
E	-1.920	0.9669	-1.986	0.3283

TRT = B subtracted from:

Level TRT	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
C	6.660	0.9669	6.888	0.0001
D	1.500	0.9669	1.551	0.5517
E	-1.540	0.9669	-1.593	0.5283

TRT = C subtracted from:

Level TRT	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
D	-5.160	0.9669	-5.337	0.0014
E	-8.200	0.9669	-8.481	0.0000

TRT = D subtracted from:

Level TRT	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
E	-3.040	0.9669	-3.144	0.0538

The results from Tukey's tests above can be succinctly summarized in Table 12.5.

Table 12.5 Summary of Tukey's tests for the means

μ_C	μ_D	μ_A	μ_B	μ_E
13.120	7.960	6.840	6.460	4.920

12.3 Missing Values in Latin Square Designs

With one missing value the analysis of the Latin square design is relatively not complicated. Suppose the yield on the plot in row i and column j is missing and this plot received treatment k, then for a $t \times t$ Latin square, an estimate of the missing yield (plot) is

$$Y^*_{ijk} = \frac{tR_i^{'} + tC_j^{'} + tT_k^{'} - 2G^{'}}{(t-1)(t-2)} \tag{12.2}$$

where the primes indicate totals for the row, column and treatment with the missing value and $G^{'}$ is the grand total with the missing value.

As in the randomized block design, we may replace the missing yield with this quantity and perform the analysis of variance exactly as above with the modifications that both error and total degrees of freedom are reduced by one each. That is, the error degree of freedom becomes $(t-1)(t-2)-1 = t^2 - 3t + 1$ while that for total becomes $t^2 - 2$.

Example 12.3.1

Suppose in Example 12.2.1, the yield in row (steer) 4, column (Period) V is missing. This corresponds to the yield for treatment C. Hence,

$$R_4^{'} = 386.6 - 69.6 = 317.0$$

$$C_5^{'} = 385.8 - 69.6 = 316.2$$

$$T_C^{'} = 3923.3 - 69.6 = 322.7$$

$$G^{'} = 2275.8 - 69.6 = 2206.2.$$

If we had not known this value before hand, and we have had to estimate it, we would have the above totals for the fourth row and fifth column in which the missing plot occurs, and the total as it affects the treatment concerned in $T_C^{'}$. Hence, from Eq. (12.2), we have

$$\begin{aligned} Y^*_{45C} &= \frac{6(317.0 + 316.2 + 322.7) - 2(2206.2)}{5 \times 4} \\ &= \frac{6(955.9) - 4412.4}{20} \\ &= 66.2. \end{aligned}$$

Hence, an estimate of the missing value is 66.2 (compare with the true value of 69.6)

$$R_4 = 317.0 + 66.2 = 383.2$$
$$C_5 = 316.2 + 66.2 = 382.4$$
$$T_C = 322.7 + 66.2 = 388.9$$
$$G' = 2206.2 + 66.2 = 2272.4$$
$$\text{CF} = \frac{2272.4^2}{36} = 143438.94.$$

Hence, the total, row, column, treatments and errors SS are computed giving the following results and the corresponding analysis of variance table in Table 12.6.

$$\text{Total SS} = 672.58$$
$$\text{Rows (Steer) SS} = 70.20$$
$$\text{Columns (Periods) SS} = 107.182$$
$$\text{Treatments (Rations) SS} = 379.03$$
$$\text{Error SS} = 116.17$$

Note the reduction in both the error and total degrees of freedom by 1. This is so because we have estimated one parameter $Y_{45\ C}$ from the data. If two missing values have been estimated, the reductions could have been by 2 degrees of freedom each and so on.

The estimated standard error of the difference between the corresponding treatment (C) mean and the mean of a treatment with no missing values is

Table 12.6 Analysis of variance table of a missing value

Source	d.f.	SS	MS	F
Steers	5	70.200		
Periods	5	107.182		
Rations	5	379.030	75.81	12.41
Error	19	116.17	$6.11 = S^2$	
Total	34	672.58		

$$= \sqrt{S^2\left[\frac{2}{t} + \frac{1}{(t-1)(t-2)}\right]}$$
$$= \sqrt{6.11\left[\frac{2}{6} + \frac{1}{20}\right]}$$
$$= 1.53.$$

The missing value analysis is carried out in MINITAB by declaring that the missing cell is blank (*), and then carry out the analysis as usual. MINITAB will adjust both the total and error degrees of freedom appropriately. The output from the implementation is presented as follows.

```
.................
20      4   67.4
21      4   65.1
22      4   65.1
23      4      *
24      4   52.7
25      5   67.8
26      5   64.7
27      5   63.6
.................

General Linear Model: Y versus STEER, PERIOD, TRT


Factor     Type Levels Values
STEER     fixed      6 1 2 3 4 5 6
PERIOD    fixed      6 1 2 3 4 5 6
TRT       fixed      6 A B C D E F

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
STEER       5     66.989     69.342     13.868    2.27  0.089
PERIOD      5    104.920    106.596     21.319    3.49  0.021
TRT         5    374.759    374.759     74.952   12.26  0.000
Error      19    116.171    116.171      6.114
Total      34    662.839

Least Squares Means for Y

TRT      Mean   SE Mean
A       56.60     1.009
B       61.85     1.009
C       64.81     1.151
D       66.07     1.009
E       65.82     1.009
F       63.58     1.009
```

12.4 Graeco–Latin Square Designs

Graeco–Latin squares are Latin squares in which for a given $t \times t$ Latin square, a second $t \times t$ Latin square of treatments which are denoted by Greek letters are superimposed.

The Graeco–Latin square properties are exemplified simply by the 3×3 Latin square below in Table 12.7.

Table 12.7 A 3×3 Graeco–Latin square design

A_α	B_β	C_γ
B_γ	C_α	A_β
C_β	A_γ	B_α

In the above arrangement, every Latin letter (A, B, C) occurs once in each row and once in each column, each Greek letter (α, β, γ) occurs once in each row and once in each column and each Greek letter occurs once and only once with each Latin letter. The squares (A, B, C) and (α, β, γ) are said to be orthogonal and the design is called a Graeco–Latin square.

Graeco–Latin squares of side t exist when t is a prime number (i.e., a number that can only be divided by 1 or itself) or a power of a prime. Example, 3 is a prime number, hence a 3×3 G–L square exists. 2 is also a prime number, so a 2^2 or 3^2 G–L squares exists, that is, a 4×4 or 9×9 G–L squares. Example of a 4×4 Graeco–Latin square is shown below in Table 12.8.

Table 12.8 A 4×4 Graeco–Latin (G-L) square design

A_α	B_β	C_γ	D_δ
B_δ	A_γ	D_β	C_α
C_β	D_α	A_δ	B_γ
D_γ	C_δ	B_α	A_β

As in the 3×3 case, the Latin and Greek letters each form a Latin square. Further each Greek letter occurs once and only once with each Latin letter. Hence, by property of Graeco–Latin, the above design is a Graeco–Latin Square.

Example 12.4.1

In an experiment involving maize of different varieties, it was considered that the spacing of the maize could influence yield. So five spacing methods were studied in addition to five varieties of maize and five locations, as well as five fertilizer treatments. A Graeco–Latin square design was chosen for the study and the yield of maize per hectare are presented in Table 12.9.

Table 12.9 Yield of varieties for the example

Variety	I		II		III		IV		V	
1	β C	5.65	δ D	7.68	α E	8.75	γ B	4.32	ϕ A	5.27
2	ϕ B	3.79	γ C	8.35	β D	4.98	α A	5.94	δ E	7.50
3	γ E	8.12	β A	6.27	δ B	4.22	ϕ D	7.29	α C	4.71
4	δ A	7.93	α B	4.77	ϕ C	6.92	β E	8.48	γ D	6.51
5	α D	4.85	ϕ E	8.88	γ A	8.45	δ C	4.49	β B	4.88

where

$$\text{Rows} = \text{the varieties}$$

$$\text{Columns} = \text{the locations}$$

$$\text{Latin letters} = \text{the fertilizer type}$$

$$\text{Greek letters} = \text{the spacings.}$$

We have decided to employ the MINITAB for this analysis.

```
MTB > SET C1
DATA> (1:5)5
DATA> END
MTB > SET C2
DATA> 5(1:5)
DATA> END
MTB > SET C5
DATA> 5.65 7.68 8.75 4.32 5.27
DATA> 3.79 8.35 4.98 5.94 7.50
DATA> 8.12 6.27 4.22 7.29 4.71
DATA> 7.93 4.77 6.92 8.48 6.51
DATA> 4.85 8.88 8.45 4.49 4.88
DATA> END
MTB > PRINT C1-C5

Data Display

 Row  VARIETY  LOCATION  SPACING   FERT      Y

   1        1         1    beta       C   5.65
   2        1         2    delta      D   7.68
   3        1         3    alpha      E   8.75
   4        1         4    gamma      B   4.32
   5        1         5    phi        A   5.27
   6        2         1    phi        B   3.79
   7        2         2    gamma      C   8.35
   8        2         3    beta       D   4.98
   9        2         4    alpha      A   5.94
  10        2         5    delta      E   7.50
  11        3         1    gamma      E   8.12
  12        3         2    beta       A   6.27
  13        3         3    delta      B   4.22
  14        3         4    phi        D   7.29
  15        3         5    alpha      C   4.71
  16        4         1    delta      A   7.93
  17        4         2    alpha      B   4.77
  18        4         3    phi        C   6.92
  19        4         4    beta       E   8.48
  20        4         5    gamma      D   6.51
  21        5         1    alpha      D   4.85
  22        5         2    phi        E   8.88
  23        5         3    gamma      A   8.45
  24        5         4    delta      C   4.49
  25        5         5    beta       B   4.88

MTB > GLM 'Y' = VARIETY LOCATION SPACING FERT;
SUBC>   Brief 1 ;
SUBC>   Means VARIETY LOCATION SPACING FERT.

General Linear Model: Y versus VARIETY, LOCATION, SPACING,FERT


Factor     Type Levels Values
VARIETY   fixed      5 1 2 3 4 5
LOCATION  fixed      5 1 2 3 4 5
SPACING   fixed      5 alpha beta  delta gamma phi
FERT      fixed      5 A B C D E
```

```
Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
VARIETY     4      2.186      2.186      0.546    0.34  0.843
LOCATION    4      6.378      6.378      1.594    0.99  0.464
SPACING     4      5.165      5.165      1.291    0.80  0.555
FERT        4     40.469     40.469     10.117    6.30  0.014
Error       8     12.839     12.839      1.605
Total      24     67.036

Least Squares Means for Y

VARIETY       Mean   SE Mean
1            6.334    0.5665
2            6.112    0.5665
3            6.122    0.5665
4            6.922    0.5665
5            6.310    0.5665

LOCATION
1            6.068    0.5665
2            7.190    0.5665
3            6.664    0.5665
4            6.104    0.5665
5            5.774    0.5665

SPACING
alpha        5.804    0.5665
beta         6.052    0.5665
delta        6.364    0.5665
gamma        7.150    0.5665
phi          6.430    0.5665

FERT
A            6.772    0.5665
B            4.396    0.5665
C            6.024    0.5665
D            6.262    0.5665
E            8.346    0.5665
```

The above results indicate that there are no significant differences between (i) the means of varieties, (ii) the means of locations and (iii) spacing. However, there are significant differences between the means of the fertilizers, the p value being 0.014. However, Tukey's test prove inconclusive (see below) and we may state here that there does not seem to be an overwhelming significant difference between the five variety means.

```
Tukey Simultaneous Tests
Response Variable Y
All Pairwise Comparisons among Levels of FERT

FERT = A subtracted from:

Level    Difference       SE of             Adjusted
FERT       of Means  Difference   T-Value    P-Value
B           -2.376      0.8012    -2.966     0.0976
C           -0.748      0.8012    -0.934     0.8762
D           -0.510      0.8012    -0.637     0.9645
E            1.574      0.8012     1.965     0.3598
```

```
FERT = B subtracted from:

Level    Difference       SE of             Adjusted
FERT       of Means  Difference   T-Value    P-Value
C             1.628      0.8012     2.032     0.3320
D             1.866      0.8012     2.329     0.2290
E             3.950      0.8012     4.930     0.0074

FERT = C subtracted from:

Level    Difference       SE of             Adjusted
FERT       of Means  Difference   T-Value    P-Value
D            0.2380      0.8012    0.2971     0.9979
E            2.3220      0.8012    2.8981     0.1070

FERT = D subtracted from:

Level    Difference       SE of             Adjusted
FERT       of Means  Difference   T-Value    P-Value
E             2.084      0.8012     2.601     0.1600
```

12.4.1 The Completely Orthogonalized Square

If t is a prime or a power of a prime, there exists a $t \times t$ squares with each cell containing a letter of each of $(t-1)$ languages, such that the letters of any two languages form a square with the Graeco–Latin square property. An example is displayed in Table 12.10.

Note that the Latin letters, the Greek letters and the numerals have the Latin square property, and also the Latin and Greek letters form a Graeco–Latin square, as do the Latin letters and numerals, and that the Greek letters and numerals have the Graeco–Latin square property. Corresponding to the existence of a completely orthogonalized square of side t, there exists a very useful partitioning of t^2 plots. There are in general, $(t+1)$ groups of partitioning, namely,

Table 12.10 A 4×4 Completely orthogonalized square

$A_{1\ \alpha}$	$B_{2\ \beta}$	$C_{3\ \gamma}$	$D_{4\ \delta}$
$B_{4\ \gamma}$	$A_{3\ \delta}$	$D_{2\ \alpha}$	$C_{1\ \beta}$
$C_{2\ \delta}$	$D_{1\ \gamma}$	$A_{4\ \beta}$	$B_{3\ \alpha}$
$D_{3\ \beta}$	$C_{4\ \alpha}$	$B_{1\ \delta}$	$A_{2\ \gamma}$

Group	d.f.
1 Rows	$t-1$
2 Columns	$t-1$
3 Latin Letters	$t-1$
4 Greek Letters	$t-1$
5 Numerals	$t-1$
Error	$(t-1)(t-6)$
Total	t^2-1

While for the 3×3 Graeco–Latin square in Table 12.7, there are $3-1=2$ language letters, i.e., the square in Table 12.7 is also a completely orthogonalized square with the following ANOVA structure

Group	d.f.
Rows	2
Columns	2
Latin Letters	2
Greek Letters	2
Error	0
Total	8

Note that the error d.f $= 0$ and hence, the error mean square can not be estimated for this design. For this reason, multiples of this set up or $t > 3$ Graeco–Latin squares are always used. We also note that with this, there are $(3+1)=4$ groupings.

Example 12.4.2

In Sir Ronald A. Fisher's book entitled "The Design of Experiments," the following puzzle is given.

> Sixteen passengers on a liner discover that they are an exceptionally representative body. Four are Englishmen, four are Scots, four are Irish, and four are Welsh. There are also four each of four different ages, 35, 45, 55 and 65 and no two of the same age are of the same nationality. By profession also four are lawyers, four soldiers, four doctors and four clergymen, and no two of the same profession are of the same age or of the same nationality". It appears, also that four are bachelors, four married, four widowed and four divorced, and that no two of the same marital status are of the same profession, or the same age, or the same nationality. Finally, four are conservatives, four Liberals, four Socialists and four fascists, and no two of the same political sympathies are of the same marital status, or the same profession or the same age, or the same nationality."
>
> (i) Three of the fascists are known to be an unmarried English lawyer of 65, a married Scots soldier of 55 and a widowed Irish doctor of 45. It is then easy to specify the remaining fascist.
> (ii) It is further given that the Irish socialist is 35, the conservative of 45 is a Scotsman, and the English man of 55 is a clergyman.

What do you know of the Welsh Lawyer?

Solution

Marital (Row)	Nationality (Column)	Profession	Political sympathy	Age
1. Divorced	1. Englishman	A. Soldier	α. Liberal	1. 45
2. Widowed	2. Scot	B. Clergyman	β. Socialist	2. 55
3. Married	3. Irish	C. Doctor	γ. Conservative	3. 35
4. Bachelor	4. Welsh	D. Lawyer	δ. Fascist	4. 65

The above puzzle can be solved by the use of a completely orthogonalized Latin square of side 4 we encourage readers to solve this as an exercise.

Marital status	Nationality 1	2	3	4
1	$A_{1\ \alpha}$	$C_{4\ \beta}$	$D_{2\ \gamma}$	$B_{3\ \delta}$
2	$B_{2\ \beta}$	$D_{3\ \alpha}$	$C_{1\ \delta}$	$A_{4\ \gamma}$
3	$C_{3\ \gamma}$	$A_{2\ \delta}$	$B_{4\ \alpha}$	$D_{1\ \beta}$
4	$D_{4\ \delta}$	$B_{1\ \gamma}$	$A_{3\ \beta}$	$C_{2\ \alpha}$

12.4.2 Relative Efficiencies of Latin Square Design

We present the following for computing the efficiencies of the Latin square design relative to the completely randomized design (CRD) and the randomized complete block design (RCBD) respectively.

(i) $RE(LS, CRD)$

$$RE = \frac{MSE_{CR}}{MSE_{LS}} = \frac{MSR + MSC + (t-1)MSE}{(t+1)MSE}. \tag{12.3}$$

(ii) $RE(LS, RCBD_{col})$

$$RE = \frac{MSE_{RCBD}}{MSE_{LS}} = \frac{MSR + (t-1)MSE}{t\,MSE}. \tag{12.4}$$

(iii) $RE(LS, RCBD_{row})$

$$RE = \frac{MSE_{RCBD}}{MSE_{LS}} = \frac{MSC + (t-1)MSE}{t\,MSE}. \tag{12.5}$$

The expression in (12.4) is the relative efficiency of the Latin square design to the RCBD with columns of the Latin square employed as blocks in the RCBD. Similarly, the expression in (12.5) is the corresponding relative efficiency of the Latin square design to the RCBD with rows of the Latin square employed as blocks in the RCBD. As an example, consider the data in Table 12.1,

here $t = 6$ and from the analysis of variance table presented in Table 12.2 displayed on p. 344, we have $MSE = 6.14$, $MSER = 15.37$, $MSEC = 22.59$ and therefore,

$$RE(LS, CRD) = \frac{15.37 + 22.59 + 5(6.14)}{7(6.14)} = \frac{68.66}{42.98} = 1.60$$

$$RE(LS, RBD_{col}) = \frac{15.57 + 5(6.14)}{6(6.14)} = \frac{46.27}{36.84} = 1.26$$

$$RE(LS, RBD_{row}) = \frac{22.59 + 5(6.14)}{6(6.14)} = \frac{53.29}{36.84} = 1.45.$$

Clearly, the Latin square design is much more efficient than the CRD and 26 % more efficient than the RCBD if the columns were used as blocks in a RCBD, while it is 45 % more efficient than the RCBD if the rows were used as blocks—ignoring columns.

12.5 Multiple Latin Squares

We recall that for Latin squares design with t treatments the error degrees of freedom is always $(t - 1)(t - 2)$ if there are no missing values. One disadvantage of the Latin square design, however, is that by eliminating two sources of variation, we inevitably reduce the degrees of freedom available for estimating the remaining unexplained random variations of the error and this leads to comparisons of treatment means to be less precise. That is, fewer degrees of freedom are available. For example, for the 3×3 design, we have 2×1 degrees of freedom for error which is quite small.

One way of overcoming this difficulty while retaining the advantage of two blocking factors, is to use more than one Latin square which will more than double the degrees of freedom available for error.

This is usually accomplished by adopting one of the two designs below.

(a) The two (or more) squares are treated entirely independent of each other; we shall illustrate this with a 3 basic design. This usually occurs in say a field experiment where two Latin squares are employed in two separate locations of the trial. The layout can be of the following form:

	Column					
Row	1	2	3	4	5	6
1	B	A	C			
2	C	B	A			
3	A	C	B			
4				C	B	A
5				B	A	C
6				A	C	B

In the above design, there are therefore completely separate squares and the rows (or columns) in the separate squares have no relation with each other.

B	A	C
C	B	A
A	C	B

(i)

C	B	A
B	A	C
A	C	B

(ii)

The analysis of variance of the above layout is summarized below. Let G_1 and G_2 be the total sum of the yields in squares (1) and (2), respectively. Let T_1, T_2 and T_3 be the treatment totals from six plots each (three from each square) corresponding, respectively, to treatments A, B, C. Then, the analysis of variance table takes the form

Source	d.f.	SS
Between squares	1	$\frac{G_1^2}{9} + \frac{G_2^2}{9} - \frac{(\sum x)^2}{18}$
Rows	4	$(\text{Row SS})_{(i)} + (\text{Row SS})_{(ii)}$
Columns	4	$(\text{Column SS})_{(i)} + (\text{Column SS})_{(ii)}$
Treatments	2	$\frac{T_1^2+T_2^2+T_3^2}{6} - \frac{(\sum x)^2}{18}$
Error	6	By subtraction
Total	17	$\sum x^2 - \frac{(\sum x)^2}{18}$

Of course, if there are more than two squares (say b), squares of size t, then the corresponding degrees of freedom become

	d.f.	$b=2$ $t=3$	$b=3$ $t=3$	$b=3$ $t=4$
Between squares	$b-1$	1	2	2
Rows (Sq)	$b(t-1)$	4	6	9
Columns (Sq)	$b(t-1)$	4	6	9
Treatments	$(t-1)$	2	2	3
Error	$(t-1)(bt-b-1)$	6	10	24
Total	bt^2-1	17	26	47

We give corresponding degrees of freedom for the cases when $b = 2$, and $t = 3$ in column 3 and $b = 3$ and $t = 3, 4$ in columns 4 and 5, respectively. The calculations of the sum of squares is also similar to that of the two 3×3 squares discussed earlier. If we assume that there is an interaction between Square and treatment as would be the case if we consider each 3×3 Latin square is laid out either at the same location, or different locations or different years. However, to have such a combined analysis, we must demonstrate that the error variances for each Latin square

are homogeneous (see Chap. 20). In this case, the square–treatment interaction degrees of freedom will be $(b-1)(t-1)$ and the error d.f. will be reduced accordingly.

(b) The second method considers the two (or more) squares being amalgamated to form a "rectangle" in which each treatment appears once in each column and twice (or more) in each row.

The two squares are written down initially and then randomization of rows and columns, and allocation of treatments proceeds as before except that now we have six columns (for our example) and not two distinct sets of three columns each. The layout is given below for our 3×3 example.

B	A	A	C	B	C
C	B	C	B	A	A
A	C	B	A	C	B

The analysis of variance table takes the following form for this design.

Source	d.f.	SS
Rows	2	$\dfrac{R_1^2+R_2^2+R_3^2}{6} - \dfrac{(\sum x)^2}{18}$
Columns	5	$\dfrac{C_1^2+C_2^2+\cdots+C_6^2}{3} - \dfrac{(\sum x)^2}{18}$
Treatments	2	$\dfrac{T_1^2+T_2^2+T_3^2}{6} - \dfrac{(\sum x)^2}{18}$
Error	8	by subtraction
Total	17	$\sum x^2 - \dfrac{(\sum x)^2}{18}$

Similarly if there are b such $(t \times t)$ squares, we give below the corresponding degrees of freedom.

		$b=2$ $t=3$	$b=3$ $t=4$
Rows	$(t-1)$	2	3
Columns	$(bt-1)$	5	11
Treatments	$(t-1)$	2	3
Error	$(t-1)(bt-2)$	8	30
Total	bt^2-1	17	47

We notice that there are more degrees of freedom for error in design (b) than in design (a). This superiority of (b) is true for all sizes of the Latin square. Not surprisingly of course, both designs are equivalent when $b=1$ that is, when we have a single $t \times t$ Latin square for the experiment. We give a practical example of a multiple Latin square experiment below.

12.5.1 Example

The example comes from Mead and Curnow (1983) and relates to an experiment to compare the effects of four light treatments, A, B, C, and D on the synthesis of mosaic virus in several tobacco leaves. The experiment was arranged in two Latin squares and leaves from four positions for eight tobacco plants formed a 4×8 rectangular Latin squares after appropriate randomization. the dependent variable is sap from the 32 leaves, which were assayed on leaves of test plants and the square root of the number of lesions appearing are taken as a measure of the treatment effects. The data is presented in Table 12.11.

Table 12.11 Sap from 32 leaves. (Source: Mead et al.)

Plant Leaf position	1	2	3	4	5	6	7	8
1	45.4 (A)	32.2 (D)	34.6 (B)	42.4 (C)	38.1 (C)	30.8 (A)	58.4 (B)	32.2 (D)
2	33.4 (B)	47.6 (B)	44.0 (D)	38.6 (D)	27.2 (A)	44.9 (C)	24.8 (A)	36.4 (C)
3	45.6 (C)	32.0 (A)	42.4 (C)	37.8 (A)	40.8 (B)	50.8 (D)	46.2 (D)	28.2 (B)
4	42.7 (D)	34.0 (C)	39.0 (A)	41.6 (B)	35.8 (D)	39.3 (B)	45.8 (C)	30.4 (A)

```
MTB > Set C1
DATA>   4( 1 : 8 / 1 )1
DATA>   End.
MTB > Set C2
DATA>   1( 1 : 4 / 1 )8
DATA>   End.
MTB > SET C3
DATA> 45.4 32.2 34.6 42.4 38.1 30.8 58.4 32.2
DATA> 33.4 47.6 44  38.6 27.2 44.9 24.8 36.4
DATA> 45.6 32 42.4 37.8 40.8 50.8 46.2 28.2
DATA> 42.7 34 39 41.6 35.8 39.3 45.8 30.4
DATA> END
MTB > SET C4
DATA> 1 4 2 3 3 1 2 4 2 2 4 4 1 3 1 3
DATA> 3 1 3 1 2 4 4 2 4 3 1 2 4 2 3 1
DATA> END
MTB > print c1-c4

Data Display

Row  PLANT  POSITION     Y  TRT
  1      1         1  45.4    1
  2      2         1  32.2    4
  3      3         1  34.6    2
  4      4         1  42.4    3
  5      5         1  38.1    3
  6      6         1  30.8    1
  7      7         1  58.4    2
  8      8         1  32.2    4
  9      1         2  33.4    2
 10      2         2  47.6    2
 11      3         2  44.0    4
 12      4         2  38.6    4
 13      5         2  27.2    1
 14      6         2  44.9    3
 15      7         2  24.8    1
 16      8         2  36.4    3
 17      1         3  45.6    3
 18      2         3  32.0    1
 19      3         3  42.4    3
 20      4         3  37.8    1
```

```
 21      5          3  40.8    2
 22      6          3  50.8    4
 23      7          3  46.2    4
 24      8          3  28.2    2
 25      1          4  42.7    4
 26      2          4  34.0    3
 27      3          4  39.0    1
 28      4          4  41.6    2
 29      5          4  35.8    4
 30      6          4  39.3    2
 31      7          4  45.8    3
 32      8          4  30.4    1

MTB > GLM 'Y' = PLANT POSITION TRT;
SUBC>   Brief 2 ;
SUBC>   Pairwise TRT;
SUBC>     Tukey;
SUBC>     NoCI.

General Linear Model: Y versus PLANT, POSITION, TRT

Factor    Type   Levels  Values
PLANT     fixed       8  1, 2, 3, 4, 5, 6, 7, 8
POSITION  fixed       4  1, 2, 3, 4
TRT       fixed       4  1, 2, 3, 4


Analysis of Variance for Y, using Adjusted SS for Tests

Source    DF   Seq SS  Adj SS  Adj MS     F      P
PLANT      7   438.22  438.22   62.60  1.24  0.333
POSITION   3    47.24   47.24   15.75  0.31  0.817
TRT        3   318.19  318.19  106.06  2.10  0.136
Error     18   909.89  909.89   50.55
Total     31  1713.54


S = 7.10980   R-Sq = 46.90%   R-Sq(adj) = 8.55%


Unusual Observations for Y

Obs        Y      Fit  SE Fit  Residual  St Resid
  7  58.4000  45.8375  4.7027   12.5625      2.36 R
 10  47.6000  36.3375  4.7027   11.2625      2.11 R
 15  24.8000  36.6250  4.7027  -11.8250     -2.22 R

R denotes an observation with a large standardized residual
```

The analysis of this rectangular Latin squares design indicates that there are no significant differences between the treatments at $\alpha = 0.05$ level of significance. Although there is no need for the Tukey's simultaneous tests, the results from these tests also indicate no significant differences between the means of the four treatments.
Suppose the design has been as displayed below.

Plant	1	7	4	8	6	3	5	2
	Square 1					Square 2		
Leaf position								
1	45.4 (A)	58.4 (B)	42.4 (C)	32.2 (D)	30.8 (A)	34.6 (B)	38.1 (C)	32.2 (D)
2	33.4 (B)	24.8 (A)	38.6 (D)	36.4 (C)	44.9 (C)	44.0 (D)	27.2 (A)	47.6 (B)
3	45.6 (C)	46.2 (D)	37.8 (A)	28.2 (B)	50.8 (D)	42.4 (C)	40.8 (B)	32.0 (A)
4	42.7 (D)	45.8 (C)	41.6 (B)	30.4 (A)	39.3 (B)	39.0 (A)	35.8 (D)	34.0 (C)

In the above data, the column sequence is now 1, 7, 4, 8 forming the first square and columns 6, 3, 5, 2 forming the second square. Now within each square, each treatment occurs once in each row and once in each column, thus forming two clearly defined two squares design. The analysis of this structure is presented with the MINITAB below.

```
MTB > print c1-c5

Data Display

Row  SQR  ROWS  COLS  TRT     Y
  1    1     1     1    1  45.4
  2    1     1     2    2  58.4
  3    1     1     3    3  42.4
  4    1     1     4    4  32.2
  5    2     1     5    1  30.8
  6    2     1     6    2  34.6
  7    2     1     7    3  38.1
  8    2     1     8    4  32.2
  9    1     2     1    2  33.4
 10    1     2     2    1  24.8
 11    1     2     3    4  38.6
 12    1     2     4    3  36.4
 13    2     2     5    3  44.9
 14    2     2     6    4  44.0
 15    2     2     7    1  27.2
 16    2     2     8    2  47.6
 17    1     3     1    3  45.6
 18    1     3     2    4  46.2
 19    1     3     3    1  37.8
 20    1     3     4    2  28.2
 21    2     3     5    4  50.8
 22    2     3     6    3  42.4
 23    2     3     7    2  40.8
 24    2     3     8    1  32.0
 25    1     4     1    4  42.7
 26    1     4     2    3  45.8
 27    1     4     3    2  41.6
 28    1     4     4    1  30.4
 29    2     4     5    2  39.3
 30    2     4     6    1  39.0
 31    2     4     7    4  35.8
 32    2     4     8    3  34.0

MTB > GLM 'Y' = SQR ROWS( SQR)  COLS( SQR)  TRT;
SUBC>   Brief 2 ;
SUBC>   Pairwise TRT;
SUBC>     Tukey;
SUBC>     NoCI.

General Linear Model: Y versus SQR, TRT, ROWS, COLS

Factor     Type   Levels  Values
SQR        fixed       2  1, 2
ROWS(SQR)  fixed       8  1, 2, 3, 4, 1, 2, 3, 4
COLS(SQR)  fixed       8  1, 2, 3, 4, 5, 6, 7, 8
TRT        fixed       4  1, 2, 3, 4

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF   Seq SS  Adj SS  Adj MS     F      P
SQR         1     8.40    8.41    8.41  0.23  0.638
ROWS(SQR)   6   410.65  410.65   68.44  1.88  0.151
COLS(SQR)   6   429.82  429.82   71.64  1.97  0.135
TRT         3   318.19  318.19  106.06  2.91  0.069
Error      15   546.47  546.47   36.43
Total      31  1713.54


S = 6.03587   R-Sq = 68.11%   R-Sq(adj) = 34.09%
```

```
Unusual Observations for Y

Obs        Y      Fit  SE Fit  Residual  St Resid
 12  36.4000  28.0750  4.3994    8.3250      2.01 R

R denotes an observation with a large standardized residual
Least Squares Means for Y


TRT   Mean  SE Mean
A    33.43    2.134
B    40.49    2.134
C    41.20    2.134
D    40.31    2.134
```

Here, as in the previous design, there does not seem to be any significant differences between the means of the four treatments. We note here the changes in the degrees of freedom for the second design particularly, for the error term. However, the second design has a smaller estimate for s^2 indicating that it will be more efficient than the former design.

12.6 Crossover Designs

One of the most common uses of crossover and Latin square design is in experiments in which the different treatments are applied in sequence to the same subject, animal or plot and the treatment effects are assumed to continue to the next period. The rows of the square represent the successive periods of application, while the columns represent the subjects, animals or plots. Previous analysis in this chapter assumes that there is no residual or carryover effect of any treatment into the succeeding period. Where there seems to be some risk of residual or carryover effects, a common practice is to separate any two periods of treatment by an interval of time long enough for the residual effects to have died out.

To allow a long enough "rest period" is some times not feasible or undesirable on other grounds. In agricultural field experiments, we should not want the rest period to be longer than the interval between the harvesting of 1 year's crop and the sowing of the next crop. In some dairy cow feeding experiments, the whole experiment must be completed in one lactation, so that the total time available for treatment periods plus rest periods is limited, and generally, the shorter the rest period, the sooner the experiment is finished.

We shall illustrate a simple design that enables the residual effects to be estimated.

Suppose that one wishes to utilize the following three merchandising treatments:

A = a display of 4-lb polythene bags

B = a display of 6-lb polythene bags

C = a display of 8-pound polythene bags

in determining the effect of size of bag on the sale of oranges in a supermarket. Six stores from a city were selected for the study. The period of observation on sales of oranges is for 1 week. If a person purchased two four-pound or one eight-pound bag of oranges, it is possible that this would affect their purchase of oranges during the following week, i.e., the effect of a treatment might last for more than the treatment period. This would be called a *residual effect* of the treatment. The sale of oranges for a given treatment above or below the mean during the treatment period (i.e., the week the treatment was in the super market) is the *direct effect* of the treatment. The following known as a double changeover design was used.

	Super markets					
Week	1	2	3	4	5	6
1	A	B	C	A	B	C
2	B	C	A	C	A	B
3	C	A	B	B	C	A

In the above, it will be noted that the treatment B follows A twice and treatment A follows B twice. The same balance is attained for pairs A and C and B and C.

The randomization procedure is to randomly allot the super markets to the columns and the letters to the treatments. The rows are not randomized. The statistical analysis is somewhat complicated because not all effects are orthogonal. The design is useful in many types of experiments. The dairy cow, the patient, the worker, the rate, the hospital etc. replace the super market category, and the period of treatment replaces the week.

Designs like the above in which each treatment is preceded equally often by each of the other treatments is said to be a *balanced design* with respect to residual effects, although the balance is incomplete because a treatment is never preceded by itself. The model here is:

$$Y_{ijkh} = \mu + w_i + s_{j(i)} + p_{h(i)} + t_k + \epsilon_{ijkh} \tag{12.6}$$

where

(i) w_i = week or period $i = 1, 2, 3$ (weeks)
(ii) $s_{j(i)}$ = super markets within week $j = 1, 2, \cdots, 6$ (supermarkets)
(iii) $p_{h(i)}$ = residual effect of treatment h in period or week i. $h = 1, 2, 3$ (carryover effects)
(iv) t_k = treatment $i = 1, 2, 3$ (treatments)
(v) ϵ_{ijkh} = random error term distributed $N(0, \sigma^2)$.

The design above with two orthogonal Latin squares, all ordered pairs of treatments occur twice and only twice throughout the design. It is thus balanced for residual effects. For designs with an even number of treatments, e.g. 4, 6, etc, this can be accomplished with a single Latin square and such

a design will be balanced. However, for odd number of treatments, e.g., 3, 5, 7, etc., a balanced design would require two orthogonal Latin squares. We present below a standard balanced crossover designs for even numbered treatments of 4 and 6 in Table 12.12 and a similar two orthogonal Latin squares design for a typical odd numbered treatments $t = 5$ in Table 12.13. That for $t = 3$ is earlier presented in our previous example.

Table 12.12 Balanced cross-over Latin squares designs for even numbered treatments

	Sequence					Sequence					
Period	1	2	3	4	Period	1	2	3	4	5	6
1	A	B	C	D	1	A	B	C	D	E	F
2	D	A	B	C	2	C	D	E	F	A	B
3	B	C	D	A	3	B	C	D	E	F	A
4	C	D	A	B	4	E	F	A	B	C	D
					5	F	A	B	C	D	E
					6	D	E	F	A	B	C

Table 12.13 Balanced cross-over design for five treatments

	Sequence group I					Sequence group 2				
Period	1	2	3	4	5	6	7	8	9	10
1	A	B	C	D	E	A	B	C	D	E
2	B	C	D	E	A	C	D	E	A	B
3	D	E	A	B	C	B	C	D	E	A
4	E	A	B	C	D	E	A	B	C	D
5	C	D	E	A	B	D	E	A	B	C

Example 12.6.1

The data in Table 12.14 come from an experiment on the feeding of dairy cows, the treatments being as follows: A = Roughage; B = Limited grain; C = Full grain. The results are the milk yields per period (3 weeks) (adapted from Cochran and Cox 1957).

Table 12.14 Plan and milk yields per period

		Period		
Sequence	Cow	1	2	3
1	1	A 38	B 25	C 15
	2	B 109	C 86	A 39
	3	C 124	A 72	B 27
2	1	A 86	C 76	B 46
	2	B 75	A 35	C 34
	3	C 101	B 63	A 1

In the above design, $3n = 3 \times 2$ cows are assigned to the $n = 2$ sequences, that is, 3 cows to each sequence. The periods correspond to the order in which the treatments are applied. The above cross-over design has the model formulation

$$Y_{ijkh} = \mu + s_i + c_{j(i)} + t_k + p_h + \epsilon_{ijkl} \tag{12.7}$$

where

- μ is the general mean
- s_i is the fixed effect of the i sequence, $i = 1, 2$
- $c_{j(i)}$ is random effect of the jth cow in sequence i.
- t_k is the fixed effect of the kth treatment, $k = 1, 2, 3$.
- p_h is the fixed effect of the hth period, $h = 1, 2, 3$.
- ϵ_{ijkl} is the random error term distributed normal with variance σ^2

We have assumed in the crossover design model in (12.7) that there is no interaction between the factor variables, treatments and period. If it is suspected that there might be interaction, this would need to be included in the above model. In the above example, since $h = 3$, we would refer the above crossover design as a *three-period crossover design.*

We also note that this design is balanced and is similar to our earlier design used for the experiment on the super market. We do note that sequence 1 forms a 3×3 Latin square while sequence 2 also forms another 3×3 Latin square. Because the design is balanced, we can further refine our model as in (12.8)

$$Y_{ijkh} = \mu + s_i + c_{j(i)} + p_{h(i)} + t_k + \epsilon_{ijkh} \tag{12.8}$$

where in this model $p_{h(i)}$ is random effect of the hth period in sequence i.

Analysis

The analysis of crossover design data is better performed with a computer software such as MINITAB or SAS. To implement the analysis in MINITAB of the data in Table 12.14, the analysis is implemented in MINITAB with the following statements. The data are read in C1–C5. The analysis of the data using the model in (12.7) is presented in the following MINITAB OUTPUT.

```
Data Display
 Row   SEQ   COW  PERIOD   TRT      Y

   1     1     1       1     A     38
   2     1     2       1     B    109
   3     1     3       1     C    124
   4     2     4       1     A     86
```

```
   5     2     5       1     B     75
   6     2     6       1     C    101
   7     1     1       2     B     25
   8     1     2       2     C     86
   9     1     3       2     A     72
  10     2     4       2     C     76
  11     2     5       2     A     35
  12     2     6       2     B     63
  13     1     1       3     C     15
  14     1     2       3     A     39
  15     1     3       3     B     27
  16     2     4       3     B     46
  17     2     5       3     C     34
  18     2     6       3     A      1
```

```
MTB > GLM 'Y' = SEQ COW(SEQ) PERIOD TRT;
SUBC>   Brief 1 ;
SUBC>   Means TRT SEQ;
SUBC>   Pairwise TRT;
SUBC>     Tukey;
SUBC>     NoCI.
```

General Linear Model: Y versus SEQ, PERIOD, TRT, COW

```
Factor     Type Levels Values
SEQ       fixed      2 1 2
COW(SEQ)  fixed      6 1 2 3 4 5 6
PERIOD    fixed      3 1 2 3
TRT       fixed      3 A B C
```

Analysis of Variance for Y, using Adjusted SS for Tests

```
Source     DF     Seq SS     Adj SS     Adj MS       F      P
SEQ         1       18.0       18.0       18.0    0.17  0.687
COW(SEQ)    4     5763.1     5763.1     1440.8   13.98  0.001
PERIOD      2    11480.1    11480.1     5740.1   55.70  0.000
TRT         2     2276.8     2276.8     1138.4   11.05  0.005
Error       8      824.4      824.4      103.1
Total      17    20362.4
```

Least Squares Means for Y

```
TRT      Mean   SE Mean
A       45.17     4.144
B       57.50     4.144
C       72.67     4.144
SEQ
1       59.44     3.384
2       57.44     3.384
```

Tukey Simultaneous Tests
Response Variable Y
All Pairwise Comparisons among Levels of TRT

TRT = A subtracted from:

```
Level   Difference       SE of             Adjusted
TRT       of Means  Difference   T-Value   P-Value
B            12.33       5.861     2.104    0.1503
C            27.50       5.861     4.692    0.0039
```

```
TRT = B subtracted from:

Level   Difference       SE of            Adjusted
TRT       of Means  Difference   T-Value   P-Value
C            15.17       5.861     2.588    0.0744
```

The sequences (cows) SS of 5781.1 on 5 d.f. obtained in our earlier ANOVA table is obtained from the MINITAB output as the sum of SEQ SS and the COW(SEQ) SS $= 18 + 5763.1 = 5781.1$ on 5 d.f. The appropriate F value computation for sequence effect is obtained from

$$F = \frac{\text{SEQ MS}}{\text{COW(SEQ) MS}} = \frac{18.0}{1440.8} = 0.01.$$

In this case, the corresponding p-value for the effect of sequence is *0.9164* rather than the value of 0.687 presented in the above analysis of variance table. The analysis indicates that, there are significant differences in the adjusted means of treatments A, B, and C. The Tukey pairwise comparison indicate that the only significant difference is between A and C. But B and C are also not significantly different, hence this analysis is not conclusive.

On the other hand, the MINITAB output for the analysis based on revised model (12.8) is again presented below. The error this time is based on 6 d.f.

```
MTB > GLM 'Y' = SEQ COW(SEQ) PERIOD(SEQ) TRT;
SUBC>   Brief 2 ;
SUBC>   Means trt.

General Linear Model: Y versus SEQ, TRT, COW, PERIOD


Factor        Type Levels Values
SEQ           fixed      2 1 2
COW(SEQ)      fixed      6 1 2 3 4 5 6
PERIOD(SEQ)   fixed      6 1 2 3 1 2 3
TRT           fixed      3 A B C

Analysis of Variance for Y, using Adjusted SS for Tests

Source         DF     Seq SS     Adj SS     Adj MS       F      P

SEQ             1       18.0       18.0       18.0    0.13  0.728
COW(SEQ)        4     5763.1     5763.1     1440.8   10.60  0.007
PERIOD(SEQ)     4    11489.1    11489.1     2872.3   21.13  0.001
TRT             2     2276.8     2276.8     1138.4    8.38  0.018
Error           6      815.4      815.4      135.9
Total          17    20362.4

Least Squares Means for Y

TRT       Mean   SE Mean
A        45.17     4.759
B        57.50     4.759
C        72.67     4.759
```

In both the analyses, the treatment effects are significant. We may note here that the above analyses do not incorporate the carryover effects of the treatments. We shall re-analyze the data to take account of the carryover effects in the next section.

Example 12.6.2

The following example is from Mead (1992). In the experiment, patients were given drugs on request. Three drugs were compared in the study and each patient received two different drugs. The allocation of drugs to patients at the first request was random but the allocation of drugs at the second request, was given to ensure equality of drug replication. The allocation of the second drug at the first request was also random. The results of the study are presented in Table 12.15. The response Y is hours of relief from pain.

Table 12.15 Hours of relief from pain for each patient after each drug application

			Patients							
Seq.	Period	Drug	1	2	3	4	5	6	7	8
1	1	T_1	2	6	4	13	5	8	4	
	2	T_2	10	8	4	0	5	12	4	
2	1	T_2	2	0	3	3	0			
	2	T_1	8	8	14	11	6			
3	1	T_1	6	7	6	8	12	4	4	
	2	T_3	6	3	0	11	13	13	14	
4	1	T_3	6	4	4	0	1	8	2	8
	2	T_1	14	4	13	9	6	12	6	12
5	1	T_3	12	1	5	2	1	4	6	5
	2	T_2	11	7	12	3	7	5	6	3
6	1	T_2	0	8	1	4	2	2	1	3
	2	T_3	8	7	10	3	12	0	12	5

The analysis is again presented as follows. We may note here that there were no sequences in the original data. The inclusion of the sequence variable is very important in the analysis of the data. Because this design is not balanced, we would use the model in (12.7) to analyze these data.

```
Data Display

 Row   SEQ  PERIOD   DRUG   SUBJ      Y

   1     1       1      1      1      2
   2     1       1      1      2      6
   3     1       1      1      3      4
   4     1       1      1      4     13
   5     1       1      1      5      5
   6     1       1      1      6      8
   7     1       1      1      7      4
   8     1       2      2      1     10
   9     1       2      2      2      8
  10     1       2      2      3      4
```

```
  ..............................
  ..............................
  76     6       1      2      6      2
  77     6       1      2      7      1
  78     6       1      2      8      3
  79     6       2      3      1      8
  80     6       2      3      2      7
  81     6       2      3      3     10
  82     6       2      3      4      3
  83     6       2      3      5     12
  84     6       2      3      6      0
  85     6       2      3      7     12
  86     6       2      3      8      5
```

```
 MTB > GLM 'Y' = SEQ SUBJ( SEQ) DRUG PERIOD;
SUBC>   Brief 2 ;
SUBC>   Means DRUG PERIOD;
SUBC>   Pairwise DRUG PERIOD;
SUBC>     Tukey.

General Linear Model: Y versus SEQ, DRUG, PERIOD, SUBJ

Factor     Type  Levels  Values
SEQ        fixed      6  1, 2, 3, 4, 5, 6
SUBJ(SEQ)  fixed     43  1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 6,
                         7, 1, 2, 3, 4, 5, 6, 7, 8, 1, 2, 3, 4, 5, 6, 7, 8, 1,
                         2, 3, 4, 5, 6, 7, 8
DRUG       fixed      3  1, 2, 3
PERIOD     fixed      2  1, 2


Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF   Seq SS  Adj SS  Adj MS      F      P
SEQ         5    72.68   16.33    3.27   0.30  0.909
SUBJ(SEQ)  37   534.58  534.58   14.45   1.34  0.184
DRUG        2   100.29  116.39   58.19   5.39  0.008
PERIOD      1   277.73  277.73  277.73  25.72  0.000
Error      40   431.99  431.99   10.80
Total      85  1417.26


S = 3.28628   R-Sq = 69.52%   R-Sq(adj) = 35.23%


Unusual Observations for Y

Obs        Y     Fit  SE Fit  Residual  St Resid
  4  13.0000  6.4104  2.4044    6.5896      2.94 R
 11   0.0000  6.5896  2.4044   -6.5896     -2.94 R

R denotes an observation with a large standardized residual.


Least Squares Means for Y

DRUG     Mean  SE Mean
1       7.907   0.6974
2       4.486   0.6895
```

```
3       5.870   0.6627
PERIOD
1       4.288   0.5051
2       7.888   0.5051


Tukey Simultaneous Tests
Response Variable Y
All Pairwise Comparisons among Levels of DRUG
DRUG = 1  subtracted from:

      Difference       SE of           Adjusted
DRUG    of Means  Difference  T-Value   P-Value
2         -3.421      1.0473   -3.266    0.0062
3         -2.037      0.9941   -2.049    0.1138


DRUG = 2  subtracted from:

      Difference       SE of           Adjusted
DRUG    of Means  Difference  T-Value   P-Value
3          1.384      0.9774    1.416    0.3423

Tukey Simultaneous Tests
Response Variable Y
All Pairwise Comparisons among Levels of PERIOD
PERIOD = 1  subtracted from:

        Difference       SE of           Adjusted
PERIOD    of Means  Difference  T-Value   P-Value
2            3.600      0.7099    5.071    0.0000
```

Results obtained in this analysis agree with those presented in Mead (1992) in terms of the interpretations. The advantage of drug 1 is quite clear in the analysis. Further the number of hours of relief is much more significantly higher in period 2 than in period 1. Clearly, drug 1 gives a significantly higher hours of pain relief than drug 2 but is not more significantly different from drug 3.

Example 12.6.3

The data in Table 12.16 come from a two-period crossover design and relates to responses for drug effects in a drug-testing study. Here, there are two drugs A and B. This is a Latin square design with the rows being represented by the experimental units (the subjects) and columns by the time periods, with the letters representing the treatments or drugs in this case. The letters can sometime also be factor-level combinations. Note that in Table 12.16, for instance, the subject numbers may not necessarily be the same. We have used 1, 2, and 3 in this case for the convenience of the analysis. These could in fact be subjects 3, 8, 10 in the first sequence, and subjects 11, 2, 20 in the second sequence.

Table 12.16 Clinical responses for drug-testing experiment using a 2-period crossover design

		Period	
Sequence	Subjects	1	2
1	1	A 7.2	B 9.0
	2	A 7.2	B 8.0
	3	A 16.4	B 20.9
2	1	B 10.2	A 9.2
	2	B 20.8	A 15.6
	3	B 11.2	A 9.0

The analysis of the data in Table 12.16 are carried out in MINITAB. The data are read into columns C1 to C5 as displayed below.

```
Data Display

 Row   SEQ   SUBJ  PERIOD   DRUG      Y

   1     1      1       1      1    7.2
   2     1      1       2      2    9.0
   3     1      2       1      1    7.2
   4     1      2       2      2    8.0
   5     1      3       1      1   16.4
   6     1      3       2      2   20.9
   7     2      1       1      2   10.2
   8     2      1       2      1    9.2
   9     2      2       1      2   20.8
  10     2      2       2      1   15.6
  11     2      3       1      2   11.2
  12     2      3       2      1    9.0

MTB > GLM 'Y' = SEQ SUBJ( SEQ) PERIOD DRUG;
SUBC>   Brief 1 ;
SUBC>   Means PERIOD DRUG SEQ;
SUBC>   Coefficients 'COEF1';
SUBC>   Pairwise SEQ PERIOD DRUG;
SUBC>     Tukey;
SUBC>     NoCI.
General Linear Model: Y versus SEQ, PERIOD, DRUG, SUBJ

Factor      Type Levels Values
SEQ        fixed      2 1 2
SUBJ(SEQ)  fixed      6 1 2 3 1 2 3
PERIOD     fixed      2 1 2
DRUG       fixed      2 1 2

Analysis of Variance for Y, using Adjusted SS for Tests

Source      DF     Seq SS     Adj SS     Adj MS       F      P
SEQ          1      4.441      4.441      4.441    2.13  0.218
SUBJ(SEQ)    4    247.783    247.783     61.946   29.70  0.003
PERIOD       1      0.141      0.141      0.141    0.07  0.808
DRUG         1     20.021     20.021     20.021    9.60  0.036
Error        4      8.343      8.343      2.086
Total       11    280.729

Least Squares Means for Y

PERIOD      Mean   SE Mean
1          12.17    0.5896
2          11.95    0.5896
```

```
DRUG
1          10.77    0.5896
2          13.35    0.5896

SEQ
1          11.45    0.5896
2          12.67    0.5896


Tukey Simultaneous Tests
Response Variable Y
All Pairwise Comparisons among Levels of DRUG

DRUG = 1 subtracted from:

Level    Difference       SE of             Adjusted
DRUG       of Means  Difference   T-Value    P-Value
2             2.583      0.8338     3.098     0.0363
```

The analysis indicates that there are no significant differences in the sequence and time-periods effects. However, there are significant differences in the effects of drugs and subjects. The drug parameter estimate is obtained as $\frac{1}{2}(10.77 - 13.35) = -1.29$.

12.6.1 Crossover Design Analysis With Carryover Effects

In examples 12.6.1, 12.6.2 and 12.6.3, we have assumed that the carryover effects are negligible. If this was not the case as often is not, then we would need to incorporate the carryover effect in our model. In this case, the treatment effects will be assessed from the new adjusted treatments SS (after adjusting for the carryover effect). Similarly, the carryover effect will be assessed after adjusting for the treatments effects.

Consider the data in example 12.6.1 again. Carry over effects would occur only in observations from the second and third periods. The model incorporating the carryover effects has model in (12.7) modified as in (12.9a), while the equivalent model for the balanced case is in (12.9b).

If we denote the direct effects by t_a, t_b, and t_c, these treatments also produce residual effects r_a, r_b, and r_c, respectively, in the period immediately following the one in which they are applied. Thus, for the third period in sequence I, the predicted total treatment effect is $(t_c + r_b)$, since treatment C is given in the third period and treatment B in the period immediately preceding. Then, the model becomes in this case,

$$Y_{ijkh} = \mu + c_j + p_{h(i)} + t_k + r_g + \epsilon_{ijkh} \tag{12.9a}$$

$$Y_{ijkh} = \mu + s_i + c_{j(i)} + p_{h(i)} + t_k + r_g + \epsilon_{ijkh} \tag{12.9b}$$

where the terms are as previously defined and r_g is the carryover effect of the treatment administered in period $h-1$ of sequence i. Thus, r_g would be zero in period 1 because no treatment precedes diet A, while the observations in

periods 2 and 3 contain the carryover effects of diet A and B, that is, r_a and r_b, respectively. The direct effects of the treatments as well as the accompanying carryover effects for the example data in Table 12.14 are presented in Table 12.17.

Table 12.17 Direct and carryover effects, t_i and r_g for (12.9)

		Period		
Sequence	Cow	1	2	3
1	1	t_a	$t_b + r_a$	$t_c + r_b$
	2	t_b	$t_c + r_b$	$t_a + r_c$
	3	t_c	$t_a + r_c$	$t_b + r_a$
2	1	t_a	$t_c + r_a$	$t_b + r_c$
	2	t_b	$t_a + r_b$	$t_c + r_a$
	3	t_c	$t_b + r_c$	$t_a + r_b$

While the model in (12.9a) can easily be implemented in SAS, this, however, cannot be easily implemented in MINITAB (model is not hierarchical). However, the equivalent model in (12.9b) is easily implementable in MINITAB. This is the one employed in this analysis.

We also note that this design is balanced and is similar to our earlier design used for the experiment on the supermarket. We do note that sequences I–III form a 3×3 Latin square while sequences IV–VI form another 3×3 Latin square.

Table 12.18 Marginal totals for this example

		Period			
Sequence	Cow	1	2	3	Totals
1	1	A 38	B 25	C 15	78
	2	B 109	C 86	A 39	234
	3	C 124	A 72	B 27	223
Totals		271	183	81	535
2	1	A 86	C 76	B 46	208
	2	B 75	A 35	C 34	144
	3	C 101	B 63	A 1	165
Totals		262	174	81	517

Analysis

$$\text{Number of treatments} = n = 3$$

$$\text{Number of squares} = m = 2.$$

For each treatment, we compute, T = treatment total. Thus

$$\mathrm{T}_A = 38 + 39 + 72 + 86 + 35 + 1 = 271$$

$$\mathrm{T}_B = 25 + 109 + 27 + 46 + 75 + 63 = 345$$

$$\mathrm{T}_C = 15 + 86 + 124 + 76 + 34 + 101 = 436.$$

If we define R to be equal to the total of the yields in periods immediately following the application of this treatment. Then, for treatment A, B and C, we have, respectively,

$$R_A = 25 + 27 + 76 + 34 = 162$$
$$R_B = 15 + 86 + 35 + 1 = 137$$
$$R_C = 39 + 72 + 46 + 63 = 220.$$

Similarly, if F = total of the sequences (columns or cows) in which this treatment is a final one, again we have for each of the treatments:

$$F_A = 234 + 165 = 399$$
$$F_B = 223 + 208 = 431$$
$$F_C = 78 + 144 = 222.$$

Let $P_1 =$ the total of all yields in the first period. Then,

$$P_1 = 271 + 262 = 533$$
$$P_2 = 183 + 174 = 357$$
$$P_3 = 81 + 81 = 162.$$

The grand total G = 533 + 357 + 162 = 1052. Hence,

$$P_1 - nG = P_1 - 3G = 533 - 3(1052) = 2623 \tag{12.10}$$
$$nP_1 - (n+2)G = 3P_1 - 5G = 3(533) - 5(1052) = 3661. \tag{12.11}$$

Table 12.19 gives for each treatment, the values of T, r and F and the subsequent computations in Eqs. (12.10) and (12.11).

Table 12.19 Computations of direct and residual effects

	T	R	F	$\hat{T} = 2\hat{t}$	Direct effect	$\hat{R} = 24\hat{r}$	$\hat{r}$	$\hat{R}$
A	271	162	399	−383	42.5	−193	−8.04	46
B	345	137	431	−56	56.1	−100	−4.17	−83
C	436	220	222	+439	76.7	+293	12.21	+37
	1052	519	1052	0		0	0	0

For the estimation of the direct effect $\hat{t}$ of a treatment, the general formula is for treatment i is:

$$mn(n^2 - n - 2)\hat{t}_i = (n^2 - n - 1)T_i + nR_i + F_i + (P_i - nG)$$
$$24\hat{t}_i = 5T_i + 3R_i + F_i + (P_i - 3G).$$

Then, for treatment A, we have for instance,

$$24\hat{T}_A = 5(271) + 3(162) + 399 + 533 - 3(1052-)$$

$$= 1355 + 486 + 399 - 2623$$
$$= -383.$$

Similarly, for treatment B, we have,

$$24\hat{t}_B = -56 \qquad \text{and} \qquad 25\hat{t}_C = 439.$$

The general mean $= \hat{\mu} = \dfrac{1052}{18} = 58.44$ since there are 2 (3×3) squares, that is, 18 rows in all. Hence, the direct effects for the treatments are calculated as follows:

$$\hat{t}_A = \frac{-383}{24} + 58.44 = 42.5$$
$$\hat{t}_B = \frac{-56}{24} + 58.44 = 56.1$$
$$\hat{t}_C = \frac{439}{24} + 58.44 = 76.7.$$

For the estimation of the residual effect $\hat{r}$ of a treatment, the general formula is for treatment i:

$$mn(n^2 - n - 2)\hat{r}_i = nT_i + n^2R_i + nF_i + nP_i - (n+2)G$$
$$24\hat{r}_i = 3T_i + 9R_i + 3F_i + 3P_i - 5G.$$

Thus for treatment A, we have:

$$24\hat{r}_A = 3(271) + 9(162) + 3(399) + 3(533) - 5(1052) = -193.$$

Similarly, $24\hat{r}_B = -100$ and $24\hat{r}_C = 293$.

Hence,

$$\hat{r}_A = \frac{-193}{24} = -8.04$$
$$\hat{r}_B = \frac{-100}{24} = -4.17$$
$$\hat{r}_C = \frac{293}{24} = +12.21.$$

12.6.2 Analysis of Variance for the Experiment

$$\text{Correction Factor, CF} = \frac{1052^2}{2 \times 3 \times 3} = 61483.6$$
$$\text{Total SS} = 38^2 + 109^2 + \cdots + 1^2 - \text{CF} = 20,362.4$$

$$\text{Sequences (cows) SS} = \frac{78^2 + 234^2 + \cdots + 165^2}{3} - \text{CF} = 5,781.1$$

$$\begin{aligned}\text{Periods SS} &= \frac{271^2}{3} + \frac{183^2}{3} + \cdots + \frac{81^2}{3} - \frac{535^2}{9} - \frac{517^2}{9} \\ &= \frac{271^2}{3} + \frac{183^2}{3} + \frac{81^2}{3} - \frac{535^2}{9} + \frac{262^2}{3} + \frac{174^2}{3} + \frac{81^2}{3} - \frac{517^2}{9} \\ &= 11,489.1\end{aligned}$$

that is, between periods within squares with $2x2 = 4$ d.f.
Direct effect (unadjusted) SS is computed as:

$$\frac{271^2}{6} + \frac{345^2}{6} + \frac{436^2}{6} - \text{CF} = 2,276.8$$

$$\begin{aligned}\text{Residual (adjusted) SS} &= \frac{\sum \hat{R}^2}{mn^3(n^2 - n - 2)} \\ &= \frac{1}{216}(193^2 + 100^2 + 293^2) \\ &= 616.2\end{aligned}$$

$$\begin{aligned}\text{Direct (adjusted) SS} &= \frac{\sum \hat{T}^2}{mn(n^2 - n - 1)(n^2 - n - 2)} \\ &= \frac{1}{120}(383^2 + 56^2 + 439^2) \\ &= 2854.6\end{aligned}$$

These computations are displayed in Table 12.20.

Table 12.20 Analysis of variance table

Source	d.f.	SS	MS	F
Sequences (cows)	5	5781.1		
Periods within squares	4	11,489.1		
Direct effects (adj)	2	2854.6	1,427.3	28.66
Residual effects (adj)	2	616.2	308.1	6.19
Error	4	199.2	49.8	
Total	17	20,362.4		

F(2,4) at $\alpha = 0.05 = 6.94$

In this example, the direct effects give a significant F-value, but residual effects do not attain significance.

MINITAB Analysis

To re-analyze the data in terms of the parameters of the model in (12.9a), we need to recode the data as follows to incorporate the carryover effects denoted by R for periods 1, 2 and 3.

Period 1						Period 2					
seq	cow	Period	trt	R	Y	seq	cow	Period	trt	R	Y
1	1	1	A	0	38	1	1	2	B	1	25
1	2	1	B	0	109	1	2	2	C	2	86
1	3	1	C	0	124	1	3	2	A	3	72
2	4	1	A	0	86	2	4	2	C	1	76
2	5	1	B	0	75	2	5	2	A	2	35
2	6	1	C	0	101	2	6	2	B	3	63

Period 3					
seq	cow	Period	trt	R	Y
1	1	3	C	2	15
1	2	3	A	3	39
1	3	3	B	1	27
2	4	3	B	2	46
2	5	3	C	3	34
2	6	3	A	1	1

The coding for the carryover is as presented in terms of the parameter λ_i in Table 12.17. To implement this model in MINITAB, we would need to create covariates X_1 and X_2 for the carryover effects. The covariates are created as follows:

$$X_1 = \begin{cases} 1 & \text{if R} = 1 \\ -1 & \text{if R} = 3 \\ 0 & \text{elsewhere} \end{cases} \qquad X_2 = \begin{cases} 1 & \text{if R} = 2 \\ -1 & \text{if R} = 3 \\ 0 & \text{elsewhere} \end{cases} \tag{12.12}$$

This leads to the display presented in the analysis from MINITAB below.

```
Data Display
Data Display

 Row   SEQ   COW  PERIOD   TRT      Y     R    X1    X2

   1     1     1       1     A     38     0     0     0
   2     1     2       1     B    109     0     0     0
   3     1     3       1     C    124     0     0     0
   4     2     4       1     A     86     0     0     0
   5     2     5       1     B     75     0     0     0
   6     2     6       1     C    101     0     0     0
   7     1     1       2     B     25     1     1     0
```

```
   8     1     2       2     C    86     2     0     1
   9     1     3       2     A    72     3    -1    -1
  10     2     4       2     C    76     1     1     0
  11     2     5       2     A    35     2     0     1
  12     2     6       2     B    63     3    -1    -1
  13     1     1       3     C    15     2     0     1
  14     1     2       3     A    39     3    -1    -1
  15     1     3       3     B    27     1     1     0
  16     2     4       3     B    46     3    -1    -1
  17     2     5       3     C    34     1     1     0
  18     2     6       3     A     1     2     0     1

MTB > GLM 'Y' = SEQ COW(SEQ) PERIOD ( SEQ ) TRT X1 X2;
SUBC>   Covariates 'X1' 'X2';
SUBC>   Brief 2 ;
SUBC>   Means TRT;
SUBC>   Pairwise TRT;
SUBC>     Tukey;
SUBC>     NoCI.

General Linear Model: Y versus SEQ, TRT, COW, PERIOD


Factor        Type Levels Values
SEQ          fixed      2 1 2
COW(SEQ)     fixed      6 1 2 3 4 5 6
PERIOD(SEQ)  fixed      6 1 2 3 1 2 3
TRT          fixed      3 A B C

Analysis of Variance for Y, using Adjusted SS for Tests

Source        DF     Seq SS     Adj SS     Adj MS       F      P
SEQ            1       18.0       18.0       18.0    0.36  0.580
COW(SEQ)       4     5763.1     3818.0      954.5   19.16  0.007
PERIOD(SEQ)    4    11489.1    11489.1     2872.3   57.66  0.001
TRT            2     2276.8     2854.6     1427.3   28.65  0.004
X1             1      546.8      258.7      258.7    5.19  0.085
X2             1       69.4       69.4       69.4    1.39  0.303
Error          4      199.3      199.3       49.8
Total         17    20362.4

Term            Coef   SE Coef        T      P
Constant      58.444     1.664    35.13  0.000
X1            -8.042     3.529    -2.28  0.085
X2            -4.167     3.529    -1.18  0.303

Least Squares Means for Y

TRT      Mean   SE Mean
A       42.49     3.112
B       56.11     3.112
C       76.74     3.112
```

```
Tukey Simultaneous Tests
Response Variable Y
All Pairwise Comparisons among Levels of TRT

TRT = A subtracted from:

Level   Difference       SE of             Adjusted
TRT       of Means  Difference   T-Value    P-Value
B            13.63       4.556     2.991     0.0842
C            34.25       4.556     7.518     0.0037

TRT = B subtracted from:

Level   Difference       SE of             Adjusted
TRT       of Means  Difference   T-Value    P-Value
C            20.63       4.556     4.527     0.0230
```

Notice that the addition of the SS for seq and cow(seq) equals $18.0 + 5763.1 = 5781.1$ on $1 + 4 = 5$ d.f. This gives a MS (unadjusted) of 1156.222 and a corresponding F value of 23.21. The combined SS for the carryover is $69.4 + 546.8 = 616.2$ on 2 d.f and a corresponding combined adjusted SS of $(258.7 + 69.4) = 328.1$ on 2 d.f. The corresponding MS $= 328.1/2 = 164.05$ with a computed F-value based on the residual MS of $\frac{164.05}{49.8} = 3.29$. This is clearly not significant. Thus, the carryover effects of the treatments are not significant in this example.

Similarly, the adjusted treatment effect has a computed F value of 28.65 with a p-value of 0.004, which is clearly significant. Hence, there are significant differences between the treatment means after adjusting for carryover effects. Adjusted treatment C is highly significantly different from both treatments A and B at $\alpha = .05$ level of significance.

From our analysis above, $\hat{r}_a = -8.042$, $\hat{r}_b = -4.167$. Hence, $\hat{r}_c = -(\hat{r}_a + \hat{r}_b)$ $= 8.042 + 4.167 = 12.209$. The above results and analysis of variance table agree with those obtained from hand calculations earlier.

Re-Analysis of Data in Example 12.6.2

The analysis of the data in example 11.6.2 provides the following results for adjusted drug effects and adjusted carryover effects. None of them was significant.

```
MTB > GLM 'Y' = SEQ SUBJ(seq) period drug x1 x2;
SUBC>   Covariates 'x1' 'x2';
SUBC>   Brief 2 .

General Linear Model: Y versus SEQ, PERIOD, DRUG, SUBJ


Factor      Type Levels Values
 SEQ        fixed      6 1 2 3 4 5 6
SUBJ(SEQ)   fixed     43 1 2 3 4 5 6 7 1 2 3 4 5 1 2 3 4 5 6 7 1 2 3 4 5 6 7 8 1
                       2 3 4 5 6 7 8 1 2 3 4 5 6 7 8
```

```
PERIOD     fixed      2 1 2
DRUG       fixed      3 1 2 3

Analysis of Variance for Y, using Adjusted SS for Tests
Source       DF     Seq SS     Adj SS     Adj MS       F      P
SEQ           5      73.95      11.29       2.26    0.20  0.961
SUBJ(SEQ)    37     535.98     535.98      14.49    1.28  0.226
PERIOD        1     265.13     281.85     281.85   24.90  0.000
DRUG          2     119.95      40.98      20.49    1.81  0.177
X1            1       1.30       0.89       0.89    0.08  0.780
X2            1       0.04       0.04       0.04    0.00  0.951
Error        38     430.08     430.08      11.32
Total        85    1426.43

Term           Coef   SE Coef        T      P
Constant     6.0711    0.3678    16.51  0.000
X1            0.564     2.009     0.28  0.780
X2            0.126     2.046     0.06  0.951

Least Squares Means for Y

DRUG      Mean   SE Mean
1        8.201     1.236
2        4.469     1.177
3        5.490     1.346
```

Example 12.6.4

A digestion trial with beef steers was conducted in an extra period Latin square crossover design to evaluate the effects of low-quality roughage on feed digestion. The low-quality roughages used in the trial were (a) cottonseed hull, (b) bermuda straw, (c) wheat straw and the high-quality roughage used as control was (d) alfalfa hay. One steer was randomly assigned to each sequence of four diets. The steer remained on each diet for 30 days and measurements on dry matter digestion were made during the last week of the trial allowing a 21-day adjustment to each diet. The rougage diet fed in the fourth period was repeated during the fifth period. The data on dry matter digestion for each steer in each sequence are shown in Table 12.21

Table 12.21 Dry matter digestion for this digestion trial

	Period				
Steer	I	II	III	IV	V
1	75(A)	76(B)	79(C)	81(D)	79(D)
2	79(C)	73(A)	79(D)	75(B)	77(B)
3	81(D)	79(C)	75(B)	72(A)	73(A)
4	76(B)	79(D)	72(A)	76(C)	73(C)

Solution

The linear model for this extra period crossover design is:

$$Y_{ij} = \mu + s_i + p_j + t_k + r_h + \epsilon_{ij} \tag{12.13}$$

where μ is the general mean, $s_i, i = 1, 2, 3, 4$ is the effect of the steer sequence, $p_j, j = 1, 2, \cdots, 5$ is the effect of the jth period, $t_k, k = 1, 2, 3, 4$ is the direct effect of the kth treatment, $r_h, h = 1, 2, 3, 4$ is the carryover effect of the hth treatment, and ϵ_{ij} is the random error term distributed normally with mean 0 and variance σ^2.

The design is balanced because every treatment follows all treatments including itself once. The carryover effect is obtained as factor variable R in the MINITAB output below. Dummy variables based on effect coding scheme is used to obtain dummy variables or covariates X1, X2 and X3, respectively, where,

$$X_1 = \begin{cases} 1 & \text{if R} = 1 \\ -1 & \text{if R} = 4 \\ 0 & \text{elsewhere} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if R} = 1 \\ -1 & \text{if R} = 4 \\ 0 & \text{elsewhere} \end{cases} \quad X_3 = \begin{cases} 1 & \text{if R} = 3 \\ -1 & \text{if R} = 4 \\ 0 & \text{elsewhere} \end{cases}$$

The above coding scheme results in the following values of X_1, X_2 and X_3 corresponding to the five levels of R.

R	X_1	X_2	X_3
0	0	0	0
1	1	0	0
2	0	1	0
3	0	0	1
4	−1	−1	−1

Thus, the linear model in (12.13) can be written in terms of the covariates X_1, X_2 and X_3 as in Eq. (12.14)

$$Y_{ij} = \mu + s_i + p_j + t_k + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \epsilon_{ij}. \tag{12.14}$$

Here, $\sum_{i}^{4} \alpha_i = 0$, which implies that $\hat{\alpha}_4 = -(\hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3)$. The implementation of the analysis in MINITAB is carried out as in the following output by first reading the data in in columns C1–C5, read in R and then create X_1, X_2 and X_3.

```
Data Display
 Row  STEERS  PERIOD   TRT      Y     R    X1    X2    X3

   1       1       1     A     75     0     0     0     0
   2       1       2     B     76     1     1     0     0
   3       1       3     C     79     2     0     1     0
   4       1       4     D     81     3     0     0     1
   5       1       5     D     79     4    -1    -1    -1
   6       2       1     C     79     0     0     0     0
   7       2       2     A     73     3     0     0     1
   8       2       3     D     79     1     1     0     0
   9       2       4     B     75     4    -1    -1    -1
  10       2       5     B     77     2     0     1     0
  11       3       1     D     81     0     0     0     0
  12       3       2     C     79     4    -1    -1    -1
  13       3       3     B     75     3     0     0     1
  14       3       4     A     72     2     0     1     0
  15       3       5     A     73     1     1     0     0
  16       4       1     B     76     0     0     0     0
  17       4       2     D     79     2     0     1     0
  18       4       3     A     72     4    -1    -1    -1
  19       4       4     C     76     1     1     0     0
  20       4       5     C     73     3     0     0     1


MTB > GLM 'Y' = STEERS  PERIOD TRT X1 X2 X3;
SUBC>   Covariates 'X1' 'X2' 'X3';
SUBC>   Brief 2 ;
SUBC>   Means TRT.

General Linear Model: Y versus STEERS, PERIOD, TRT


Factor     Type Levels Values
STEERS    fixed      4 1 2 3 4
PERIOD    fixed      5 1 2 3 4 5
TRT       fixed      4 A B C D

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
STEERS      3     20.950     14.950      4.983    2.39  0.167
PERIOD      4     11.700     11.700      2.925    1.40  0.338
TRT         3    114.550    114.550     38.183   18.33  0.002
X1          1      0.125      0.083      0.083    0.04  0.848
X2          1      1.042      2.083      2.083    1.00  0.356
X3          1      2.083      2.083      2.083    1.00  0.356
Error       6     12.500     12.500      2.083
Total      19    162.950

Term          Coef   SE Coef        T      P
Constant   76.4500    0.3227   236.87  0.000
X1         -0.1250    0.6250    -0.20  0.848
X2          0.6250    0.6250     1.00  0.356
X3         -0.6250    0.6250    -1.00  0.356

Least Squares Means for Y

TRT      Mean   SE Mean
A       72.95    0.6555
B       75.74    0.6555
C       77.49    0.6555
D       79.62    0.6555
```

The parameter estimates for the covariates are,

$$\hat{\alpha}_1 = -0.125, \quad \hat{\alpha}_2 = 0.625, \quad \hat{\alpha}_3 = -0.625. \quad \text{Hence,} \quad \hat{\alpha}_4 = 0.125$$

Alternatively, we could employ the dummy variable or indicator variable capability in MINITAB to create our dummy variables. However, MINITAB creates cell reference indicator variables only, where in this case, the variables are created as follows:

$$Z_1 = \begin{cases} 1 & \text{if R} = 1 \\ 0 & \text{elsewhere} \end{cases}; \quad Z_2 = \begin{cases} 1 & \text{if R} = 1 \\ 0 & \text{elsewhere} \end{cases}; \quad Z_3 = \begin{cases} 1 & \text{if R} = 3 \\ 0 & \text{elsewhere} \end{cases}$$

The above coding scheme results in the following values of Z_1, Z_2 and Z_3 corresponding to the five levels of R.

R	Z_1	Z_2	Z_3
0	0	0	0
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

Again, the linear model in (12.13) can be written in terms of the covariates Z_1, Z_2 and Z_3 as in Eq. (12.15).

$$Y_{ij} = \mu + s_i + p_j + t_k + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \epsilon_{ij}. \tag{12.15}$$

Here, apart from the case when $R = 0$, which takes values of zeros for the dummy variables, cell $R = 4$ also have values of the dummy variables being zeros. Hence, this category is being used as a reference for the other three categories (1,2,3). Here, therefore, the parameter estimates, converted to the effect coding scheme, become,

$$\begin{aligned} \hat{\beta}_4 &= -\frac{\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3}{4} = \hat{\alpha}_4 \\ \hat{\alpha}_1 &= \hat{\beta}_1 + \hat{\beta}_4 \\ \hat{\alpha}_2 &= \hat{\beta}_2 + \hat{\beta}_4 \\ \hat{\alpha}_3 &= \hat{\beta}_3 + \hat{\beta}_4. \end{aligned}$$

The implementation of the analysis in MINITAB is carried out again as in the following output by first reading the data into columns C1–C4, read in R into C5, and then create Z_1, Z_2 and Z_3 in MINITAB. MINITAB will actually create five dummy variables, but we will eliminate the ones relating to cases when $R = 0$ and $R = 4$, respectively, leading to the output below, including those from the analysis.

Row	STEERS	PERIOD	TRT	Y	R	Z1	Z2	Z3
1	1	1	A	75	0	0	0	0
2	1	2	B	76	1	1	0	0
3	1	3	C	79	2	0	1	0
4	1	4	D	81	3	0	0	1
5	1	5	D	79	4	0	0	0
6	2	1	C	79	0	0	0	0
7	2	2	A	73	3	0	0	1
8	2	3	D	79	1	1	0	0
9	2	4	B	75	4	0	0	0
10	2	5	B	77	2	0	1	0
11	3	1	D	81	0	0	0	0
12	3	2	C	79	4	0	0	0
13	3	3	B	75	3	0	0	1
14	3	4	A	72	2	0	1	0
15	3	5	A	73	1	1	0	0
16	4	1	B	76	0	0	0	0
17	4	2	D	79	2	0	1	0
18	4	3	A	72	4	0	0	0
19	4	4	C	76	1	1	0	0
20	4	5	C	73	3	0	0	1

```
MTB > GLM 'Y' = STEERS  PERIOD TRT Z1 Z2 Z3;
SUBC>   Covariates 'Z1' 'Z2' 'Z3';
SUBC>   Brief 2 ;
SUBC>   Means TRT.
```

General Linear Model: Y versus STEERS, PERIOD, TRT

Factor	Type	Levels	Values
STEERS	fixed	4	1 2 3 4
PERIOD	fixed	5	1 2 3 4 5
TRT	fixed	4	A B C D

Analysis of Variance for Y, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
STEERS	3	20.950	14.950	4.983	2.39	0.167
PERIOD	4	11.700	7.750	1.937	0.93	0.505
TRT	3	114.550	114.550	38.183	18.33	0.002
Z1	1	0.083	0.125	0.125	0.06	0.815
Z2	1	2.042	0.500	0.500	0.24	0.642
Z3	1	1.125	1.125	1.125	0.54	0.490
Error	6	12.500	12.500	2.083		
Total	19	162.950				

Term	Coef	SE Coef	T	P
Constant	76.5500	0.5951	128.63	0.000
Z1	-0.250	1.021	-0.24	0.815
Z2	0.500	1.021	0.49	0.642
Z3	-0.750	1.021	-0.73	0.490

Means for Covariates

Covariate	Mean	StDev
XX1	0.2000	0.4104
XX2	0.2000	0.4104
XX3	0.2000	0.4104

```
Least Squares Means for Y

TRT      Mean   SE Mean
A       72.95    0.6555
B       75.74    0.6555
C       77.49    0.6555
D       79.62    0.6555
```

The analysis of variance in both cases are the same. However, the parameter estimates as expected are different. In this case, the parameter estimates of the covariates are:

$$\hat{\beta}_1 = -0.250, \quad \hat{\beta}_2 = 0.500, \quad \hat{\beta}_3 = -0.750.$$

Hence,

$$\hat{\beta}_4 = -\frac{-0.250 + 0.500 - 0.750}{4} = 0.125.$$

In terms of the effect coding therefore, we have,

$$\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_4 = -0.250 + 0.125 = -0.125$$

$$\hat{\alpha}_2 = \hat{\beta}_2 + \hat{\beta}_4 = 0.500 + 0.125 = 0.625$$

$$\hat{\alpha}_3 = \hat{\beta}_3 + \hat{\beta}_4 = -0.750 + 0.125 = -0.625$$

$$\hat{\alpha}_4 = \hat{\beta}_4 = 0.125$$

For both coding schemes, the analysis of variance becomes:

Source	d.f.	SS	MS	F
STEERS	3	20.950	6.98	
PERIOD	4	11.700	2.93	
TRT	3	114.550	38.183	18.33
R	3	3.25	1.083	0.520
Error	6	12.500	12.500	2.083

Notice that F-value for R (carryover effects) = 0.520 equals the average of the F-values of each component of the covariates. That is, $F = 0.52 = (0.06 + 0.50 + 1.00)/3 = 0.52$. Thus, we can say that the covariates are pairwise orthogonal. The results indicate that there are no significant carryover effects. However, there is significant differences between the adjusted treatment means. To use Dunnet's test to compare treatments A, B and C with the control, first we recode the treatments from alphanumeric to numeric as in the MINITAB program below. Re-analyze the data and invoke Dunnett's comparison by stating that treatment 4, that is D, is the control treatment. The results suggest that while treatments A and B are significantly different from the control diet D, however, treatment C is not. The adjusted treatments means are given as:

$$\hat{\mu}_A = 72.95, \quad \hat{\mu}_B = 75.74, \quad \hat{\mu}_C = 77.49, \quad \hat{\mu}_D = 79.62.$$

```
MTB > Code ( "A" ) 1 ( "B" ) 2 ( "C" ) 3 ( "D" ) 4 'TRT' c12

MTB > GLM  'Y' = STEERS  PERIOD TRT1  XX1 XX2 XX3;
SUBC>   Covariates 'XX1' 'XX2' 'XX3';
SUBC>   SSquares 1;
SUBC>   Brief 2 ;
SUBC>   Means TRT1;
SUBC>   Control TRT1;
SUBC>     Levels 4;
SUBC>     Dunnett;
SUBC>     NoCI.

Dunnett Simultaneous Tests
Response Variable Y
Comparisons with Control Level
TRT1 = 4 subtracted from:

Level     Difference       SE of             Adjusted
TRT1        of Means  Difference   T-Value    P-Value
1             -6.667      0.9317    -7.155     0.0009
2             -3.875      0.9317    -4.159     0.0145
3             -2.125      0.9317    -2.281     0.1416
```

12.7 Exercises

1. A crossover study was conducted to evaluate four keyboard layouts. Twelve volunteers experienced in a common keyboard configuration were used in the study. Each subject used the four test layouts in sequence. Each subject was randomly assigned to a sequence of layouts. Each layout was used for 4 days in their ordinary data and text entry activities. On the fifth day, they were all given a common task to perform with their assigned layout and the number of errors on the task were recorded. None of the subjects knew they were being tested on the final day. The number of errors recorded on each layout are presented in the following table.

	Period			
Subject	I	II	III	IV
1	7(D)	2(B)	1(A)	5(C)
2	1(A)	4(C)	6(D)	3(B)
3	6(C)	1(A)	3(B)	7(D)
4	3(B)	6(D)	3(C)	1(A)
5	4(C)	5(D)	1(A)	2(B)
6	6(D)	4(C)	2(B)	0(A)
7	1(A)	3(B)	4(C)	5(D)
8	2(B)	2(A)	7(D)	4(C)
9	5(D)	0(A)	3(C)	3(B)
10	0(A)	4(D)	2(B)	3(C)
11	3(C)	2(B)	7(D)	0(A)
12	2(B)	4(C)	0(A)	6(D)

(a) Is the above design balanced for carryover effects? Explain.
(b) Compute the analysis of variance for the data and test the significance of the direct and carryover effects.
(c) Obtain a residual plots for the analysis and conduct a normality test on the residuals. Do you think that a transformation might be needed for these data? Which one would you suggest?

2. An animal scientist hypothesized that roughage source might influence utilization of mixed diets of beef steers by altering ruminant digestion of other diet ingredients. The mixed diet for a 65 % concentrate based on steam flaked milo and 35 % roughage, together with three roughage treatments (A) 35 % alfalfa hay as a control, (B) 17.5 % wheat straw and 17.5 % alfalfa and (C) 17.5 % cottonseed hulls and 17.5 % alfalfa. Twelve beef steers were available for the study. Each of the three roughage diets was fed to the steers in one of six possible sequences of the three diets. Each diet in each sequence was fed to two steers for 30 days. The steers were allowed a period of 21 days to adapt to a diet change before any data were collected.

   The Neutral Detergent Fiber (NDF) digestion coefficient calculated for each steer on each diet is presented in the following table.

	Sequence					
	1	2	3	4	5	6
Steer:	1 2	3 4	5 6	7 8	9 10	11 12
Period I	(A) 50 55	(B) 44 51	(C) 35 41	(A) 54 58	(B) 50 55	(C) 41 46
Period II	(B) 61 63	(C) 42 46	(A) 55 56	(C) 48 51	(A) 57 59	(B) 56 58
Period III	(C) 53 57	(A) 57 59	(B) 47 50	(B) 51 54	(C) 51 55	(A) 58 61

(a) Is this design balanced for crossover effects?
(b) Compute the analysis of variance for the data and test the significance of the direct and carryover effects.
(c) Use Dunnett test to compare the control diet, with each of the other diets and interpret your results.

3. The following are the plan and yields of grain (in lbs) from a Latin square fertilizer experiment on wheat conducted at Rothamstead Experimental Station. (Rothamstead Report 1932, p. 147):

	Columns				
Pows	1	2	3	4	5
1	72.2(D)	55.44 (SS)	36.6 (O)	67.9 (C)	73.0(S)
2	36.4(O)	46.9(C)	46.8(SS)	54.9 (S)	68.5 (D)
3	71.5 (SS)	55.6 (S)	71.6 (D)	67.5 (O)	78.4 (C)
4	68.9 (S)	53.2 (O)	69.8 (C)	79.6 (D)	77.2 (SS)
5	82.0 (C)	81.0 (D)	76.0 (S)	87.9 (SS)	70.9 (O)

where
- O = no fertilizer
- S = single dressing of Nitrogen (sulphate of Ammonia) in March
- SS = same as S, but applied six monthly dressings (November–April)
- C = Equivalent quantity of cyanamide in October (just before planting)
- D = 50:50 mixture of cyanamide and dicyanadiomide in October

Analyze the data and partition your treatment sum of squares accordingly.

4. In a study to compare the durations effects of three different formulations of a drug, 12 volunteered males were involved. A three-period crossover design was used, with four subjects assigned to each of the three treatment sequences(Sequence 1: T_1, T_2, T_3; Sequence 2: T_2, T_3, T_1; and Sequence 3: T_3, T_1, T_2). The sample data were originally presented in (Ott and Longnecker 2001).

		Period		
Sequence	Subject	1	2	3
1	1	A(1.5)	B (2.2)	C (3.4)
	2	A(2.0)	B (2.6)	C (3.1)
	3	A(1.6)	B (2.7)	C (3.2)
	4	A(1.1)	B (2.3)	C (2.9)
2	1	B(2.5)	C (3.5)	A (1.9)
	2	B(2.8)	C (3.1)	A (1.5)
	3	B(2.7)	C (2.9)	A (2.4)
	4	B(2.4)	C (2.6)	A (2.3)
3	1	C(3.3)	A (1.9)	B (2.7)
	2	C(3.1)	A (1.6)	B (2.5)
	3	C(3.6)	A (2.3)	B(2.2)
	4	C(3.0)	A (2.5)	B (2.0)

Analyze the data and test the significance of direct and carryover effects.

5. The following example is taken from Mead and Curnow (1983) and relates to an experiment to to compare the effects of four light treatments, A, B, C and D on the synthesis of mosaic virus in several tobacco leaves and was arranged in two Latin squares. Leaves from four positions for eight tobacco plants formed a 4×8 rectangle of units and two Latin squares were randomized together to produce a randomized design with each treatment appearing once for each plant and twice at each position. Sap from the 32 leaves were assayed on leaves of test plants and the square root of the number of lesions appearing are taken as a measure of the treatment effect. The data are displayed below.

Leaf	Plants							
Position	1	2	3	4	5	6	7	8
1	45.4 A	32.2D	34.6D	42.4 C	38.1 C	30.8 A	58.4B	32.2D
2	33.4B	47.6B	44.0D	38.6D	27.2 A	44.9 C	24.8 A	36.4 C
3	45.6 C	32.0 A	42.4 C	37.8 A	40.8B	50.8D	46.2D	28.2B
4	42.7D	34.0 C	39.0 A	41.6B	35.8D	39.3B	45.8 C	30.4 A

Analyze the results of this experiment and draw your conclusions.

6. The following data relate to a 4×4 Latin square design four four treatments. The results are the total milk yield in the third week of each period. We assume here that there are no carryover effects of the previous treatments. (Source: Mead and Curnow 1983)

	Cow			
Period	1	2	3	4
1	A192	B195	C292	D249
2	B190	D203	A218	C210
3	C214	A139	D245	B163
4	D221	C152	B204	A134

Carry out the analysis of variance and draw your conclusions. What is the standard error for comparing any two treatment means? Any two periods?

7. An experiment was designed to determine the effects of three diets on liver cholesterol in rats (A = control, B = control + vegetable fat, C = control + animal fat). Body weight classifications (H, M or L) of the rats and the litters from which they came were used to form a balanced set of Latin squares. The litter was nested in squares (i.e., different litters were used in each square), whereas the weight classifications were not nested. The data is presented as follows (Source: Lei Gao, Michigan State University).

	Weight	Litter		
Square	class	1	2	3
1	H	B(1.60)	A (1.97)	C (2.07)
	M	C(1.83)	B (1.71)	A (1.56)
	L	A(1.44)	C (1.84)	B (1.72)
	Weight	Litter		
	class	4	5	6
2	H	A(1.71)	C (2.02)	B (1.85)
	M	B(1.63)	A (1.75)	C (2.06)
	L	C(1.70)	B (1.59)	A (1.68)
	Weight	Litter		
	class	7	8	9
3	H	C(2.09)	B (1.83)	A (1.98)
	M	A(1.63)	C (1.91)	B (1.83)
	L	B(1.67)	A (1.63)	C (2.00)

Chapter 13
Analysis of Covariance

13.1 Introduction

The analysis of covariance (ANACOVA) is a statistical technique which is a combination of *Regression* and *Analysis of variance*. It is used in experiments where besides the observations of primary interests, (variates) one or more other observations are taken on each experimental unit, called CONCOMITANT variables or *Covariates*. Measurements on the covariates are made for the purpose of adjusting the measurements on the variate. These can be used to increase precision of the experimental comparisons or to throw further light on the treatment effects, or to remove environmental effects. It is assumed that the concomitant variable (X) cannot be controlled by the experimenter but can be observed along with the variable of interest (Y). Thus, analysis of covariance is a method of adjusting for the effects of an uncontrollable nuisance variable. We present examples of the use of covariance analysis.

Example 13.1.1

Suppose in an experiment to study the effects of various diets (treatments) for the increase in body weight (Y) of cows, it would be necessary to have a group of cows at a fixed age and record initial weight (X) of each cow. Then X is the concomitant variable, and in the analysis, we shall try to adjust the experimental results Y on the basis of their X values.

Example 13.1.2

In this example, consider an experiment to investigate drugs that are hypothesized to reduce blood pressure of adults. Since the blood pressure of adults before the administration of the drugs (treatments) varies considerably from one adult to another, a grouping of the adults according to their initial blood pressure is sometimes possible, albeit cumbersome in practice.

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_13,

Thus as in the earlier example, we may record the blood pressure X of each adult before the administration of the drugs and use this information to adjust the treatment effects Y in our analysis.

Example 13.1.3

It must be noted that not all concomitant observations are taken before treatments. On certain occasions, they are taken either during the experiment or at the end of the experiment. In this example for instance, suppose 40 plants (e.g., maize) had originally been planted in each plot. At harvest time, however, some of the plots have only 25 or even 20 plants left, the rest of them being eaten by wild animals. The yields from such plots are naturally lower than those from the original 40 plots. The number of plants may be recorded as X and later used to correct the yield Y. The assumption is that the number of plants left is not due to treatments but is due to an uncorrelated factor (animals) which introduces heterogeneity to the experimental plots.

Example 13.1.4

A further use of analysis of covariance is in the missing plot technique where dummy variables are used as concomitants. Several examples will be given in this chapter to further illustrate the type of experimental situations in which the analysis of covariance can be profitably employed.

Example 13.1.5 (Data Example)

The data in Table 13.1 were obtained in an experiment to compare three methods of applying a rust arrestor compound to steel coupons. These methods were brushing, spraying, and dipping. Fifteen steel coupons used were divided randomly into three groups. All of the steel coupons were in an initial state of rust (measured as X) and they were all exposed to a salt spray, the additional amount of rust due to this being measured as Y.

Table 13.1 Data for this example

Brushing			Spraying			Dipping		
Coupon	Y	X	Coupon	Y	X	Coupon	Y	X
1	63	16	6	81	48	11	72	40
2	77	45	7	73	40	12	54	31
3	81	50	8	59	24	13	57	40
4	60	19	9	74	33	14	59	33
5	63	18	10	77	41	15	52	20
Total	344	148		364	186		294	164
Mean	68.8	29.6		72.8	37.2		58.8	32.8

A question of primary interest is whether the method of application affects the additional amount of rust. The observation of X can be used to improve the precision of experimental comparisons here.

The analysis of covariance is a method of adjusting the treatment means to what might have been obtained if a common value of X had been used throughout. If in an experiment, there exists a relationship between the observation Y and a concomitant variable X, the observed (unadjusted) means of Y could indicate completely wrong results.

13.1.1 Model and Assumptions

A model of the following form is suggested by the discussion above:

$$Y_{ij} = \mu + t_i + \beta(X_{ij} - \bar{X}_{..}) + \epsilon_{ij}; \quad \begin{matrix} i = 1, 2, 3 \\ j = 1, 2, 3, 4, 5 \end{matrix} \tag{13.1}$$

where, it is assumed that

(i) The X's are fixed, measured without error, and independent of treatments.
(ii) The $\{\epsilon_{ij}\}$ are independently and normally distributed with mean 0 and variance σ^2
(iii) the regression coefficient β is the same for all treatments
(iv) the concomitant variable is unaffected by the particular assignment of treatments to units used. A further implicit assumption is that the concomitant variable does not itself contain errors.

Analysis

Firstly, the β of Eq. (13.1) must be estimated and its significance assessed. Then the treatment means must be adjusted for the mean values of the concomitant variable. Standard errors obtained, and assumptions checked.

The corrected SS (or cross-products) for total and treatments are first calculated in the usual way, for each of X, XY and Y, to give the following entries for the ANOVA Table in Table 13.2.
Where for the X's, we have,

Table 13.2 Analysis of variance table

Source	df	X	XY	Y
Treatment	2	145.6	100.0	520.0
Error	12	1686.8	1074.6	884.4
Total	14	1832.4	1174.6	1404.4

$$\text{Total} = 148 + 186 + 164 = 498$$

$$\text{Total line SS} = 16^2 + 45^2 + \cdots + 20^2 - \frac{490^2}{15} = 1832.4$$

$$\text{Treatment line SS} = \frac{148^2}{5} + \frac{186^2}{5} + \frac{164^2}{5} - \frac{482^2}{15} = 145.6$$

$$\text{Error Line SS} = \text{Total SS} - \text{Treatment SS} = 1686.3.$$

Similarly for the Y's, we have,

$$\text{Total Line SS} = 63^2 + 77^2 + \cdots + 52^2 - \frac{1002^2}{15} = 1404.4$$

$$\text{Treatment Line SS} = \frac{344^2}{5} + \frac{364^2}{5} + \frac{294^2}{5} - \frac{1002^2}{15} = 520.0$$

$$\text{Error Line SS} = 1404.4 - 520.0 = 884.4$$

and for the XY (Cross-products), we also have,

$$\begin{aligned}\text{Total Line} &= (63 \times 16) + (77 \times 45) + \cdots + (52 \times 20) - \frac{1002 \times 498}{15} \\ &= 1174.6 \\ \text{Treatment Line} &= \frac{(344 \times 148)}{5} + \frac{(364 \times 186)}{5} + \frac{(294 \times 164)}{5} - \frac{(1002 \times 498)}{15} \\ &= 100.0 \\ \text{Error Line} &= 1174.6 - 100.0 \\ &= 1074.6\end{aligned}$$

The results in Table 13.2 can be used to calculate the regression coefficients and regression sum of squares for both the Total SS and Error SS lines:

	Total line SS	Error line SS
S_{xy}	1174.6	1074.6
S_{xx}	1832.4	1686.8
$\hat{\beta}$	0.6410	0.6371
Reg.SS	752.94	684.6

where $\hat{\beta}$ is computed from both the total line and error line SS respectively as,

$$\frac{S_{xy}}{S_{xx}} = \frac{1174.6}{1832.4} = 0.6410; \qquad \frac{1074.6}{1686.8} = 0.6371$$

Similarly, the regression SS are again computed from both total and error SS lines as,

$$\frac{S_{xy}^2}{S_{xx}} = \frac{1174.6^2}{1832.4} = 752.94; \qquad \frac{1074.6^2}{1686.8} = 684.6$$

The reduced SS can now be formed, by subtracting the regression SS from the corrected SS for Y - this is done for both the Total SS and Error SS lines. These give

$$\text{For Total SS: Reduced SS} = 1404.4 - 752.9 = 651.5 \quad \text{on } 14 - 1 = 13 \text{ d.f.}$$

$$\text{For Error SS: Reduced SS} = 884.4 - 684.6 = 199.8 \quad \text{on } 12 - 1 = 11 \text{ d.f.}$$

This leads to the analysis of covariance in Table 13.3, where the adjusted treatment SS is calculated by subtraction.

Table 13.3 Analysis of covariance table

Source	df	SS	MS	F
Treatment (adj)	2	451.7	225	12.4
Residual (error)	11	199.8	18.16	
Total	13	651.5		

The value of F on 2 and 11 degrees of freedom is very highly significant, indicating evidence of differences between adjusted treatment means. The significance of the regression can now be tested from Table 13.4.

Table 13.4 Analysis of error variance

Source	df	SS	MS	F
Within groups unadjusted	12	884.4	73.7	
Reduction due to Reg.	1	684.6	864.6	37.7
Error for adjusted	11	199.8	18.16	

The F value is very highly significant, strongly indicating a strong *linear relationship between Y and X*. We note that the error mean square has been reduced from 73.7 to 18.16 by the use of the covariate.

The adjusted group means can now be calculated from Eq. (13.1), we have

$$\hat{\bar{t}}_i = \bar{Y}_i - \hat{\beta}(\bar{X}_i - \bar{X}_{..}) \tag{13.2}$$

Using (13.1), the summary statistics in Table 13.1, and the residual line estimate of slope $\hat{\beta} = 0.6371$, we have for the three treatments,

$$\hat{t}_1 = 68.6 - 0.6371\ (29.6 - 33.20) = 70.893$$

$$\hat{t}_2 = 72.8 - 0.6371\ (37.2 - 33.20) = 70.252$$

$$\hat{t}_3 = 58.8 - 0.6371\ (32.8 - 33.20) = 59.055$$

where $\bar{X}_{..} = 33.20$ is obtained from Table 13.1. The results are as follows:

Brushing : 70.893

Spraying : 70.252

Dipping : 59.055

We also note here that the use of the covariate has altered the order of the magnitude of two of the means.

13.2 Further Analysis

13.2.1 Standard Errors

Estimate of σ = s is $\sqrt{18.16} = 4.26$.

S.E of an adjusted mean is,

$$s\sqrt{\left\{\frac{1}{n} + \frac{(\bar{X}_{i.} - \bar{X}_{..})^2}{SSE_{xx}}\right\}} \tag{13.3}$$

S.E for comparing two adjusted means i and j equals,

$$s\sqrt{\left\{\frac{2}{n} + \frac{(\bar{X}_{i.} - \bar{X}_{j.})^2}{SSE_{xx}}\right\}} \tag{13.4}$$

Where $SSE_{xx} = 1686.80$. In the above example for instance, the standard error for comparing the adjusted means for Brushing and Dipping is given by

$$4.26\sqrt{\left\{\frac{2}{5} + \frac{(29.6 - 32.8)^2}{1686.8}\right\}} = 4.26\sqrt{0.4061} = 2.7146$$

S.E of an adjusted mean is,

$$s\sqrt{\left\{\frac{1}{n} + \frac{(\bar{X}_{i.} - \bar{X}_{..})^2}{SSE_{xx}}\right\}} \tag{13.5}$$

S.E for comparing two adjusted means i and j equals,

$$s\sqrt{\left\{\frac{2}{n} + \frac{(\bar{X}_{i.} - \bar{X}_{..})^2}{SSE_{xx}}\right\}} \tag{13.6}$$

In the above example for instance, the standard error for comparing the adjusted means for Brushing and Dipping is given by

$$4.26\sqrt{\left\{\frac{2}{5} + \frac{(29.6 - 32.8)^2}{1686.8}\right\}} = 4.26\sqrt{0.4061} = 2.7146$$

Similar calculations give the ses for comparing (i) Brushing and Spraying (ii) Dipping and Spraying to be 2.809 and 2.734 respectively.

13.3 Test for Parallelism of Regression Lines

The model for the covariance assumes that there is only a single regression coefficient β. This implies that the within-treatment regression coefficients are homogeneous, i.e., for t treatments, we have,

$$\beta_1 = \beta_2 = \beta_3 = \beta_t = \beta$$

Also implicit in the model is that the correct form of the relationship between the variate (Y) and the covariate (X) is linear. We have already shown that there is a significantly strong linear relationship between X and Y. The hypothesis of homogeneity is also equivalent to the hypothesis of no treatment effects in the analysis of covariance.

The hypothesis of homogeneity, we see from above is equivalent to testing whether the within treatment regression lines are parallel. We therefore consider below the procedure for testing for parallelism.

Procedure

(i) Calculate separately for each treatment; S_{xx}, S_{xy}, and S_{yy}.
(ii) Based on calculations in (i), calculate for each treatment, the regression coefficients, regression SS, and SS for deviations from regression.
(iii) Add together the SS for deviations from the treatment regressions, subtract this from the SS for deviations from average regression (error for adjusted observation in Table (13.4)). We will thus have a mean square based on 2 d.f for comparison between three treatments

Table 13.5 Regressions within treatments

	Brushing	Spraying	Dipping	d.f.
Reg. coeff.	0.5663	0.8531	0.6566	
Reg. SS	348.08	240.74	116.74	1
Residual	12.72	36.06	130.06	3

Total residual SS $= 12.72 + 36.06 + 130.06 = 178.84$ on 9 df.
Where for example for the Brushing treatment:

$$S_{yy} = 63^2 + 77^2 + \cdots + 63^2 - \frac{344^2}{5} = 360.8$$

$$S_{xx} = 16^2 + 45^2 + \cdots + 18^2 - \frac{148^2}{5} = 1085.2$$

$$S_{xy} = (63 \times 16) + \cdots + (63 \times 18) - \frac{344 \times 148}{5} = 614.6$$

Hence,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{614.60}{1085.20} = 0.5563$$

$$\text{Reg. SS} = \frac{S_{xy}^2}{S_{xx}} = \frac{614.60^2}{1085.20} = 348.0770$$

The residual SS is obtained by subtraction as: S_{yy} − Reg. SS = 360.80 − 348.08 = 12.72 on 3 d.f. These results and those of the other two treatments are presented in Table 13.5. The results in Table 13.5 are combined to give the results for the analysis of error in Table 13.6.

Table 13.6 Analysis of error variance

Source	df	SS	MS
Error for average Reg.	11	199.80	18.16
Error from treatments Reg	9	178.84	19.87
Difference	2	20.96	10.48

F - value for the test $= \dfrac{10.48}{\text{Adjust. Error MS}} = \dfrac{10.48}{18.16} = 0.577$ which when compared with $F(2,11)$ is not significant at $\alpha = 0.05$, which indicates that the lines are parallel, i.e.,

$$\beta_1 = \beta_2 = \beta_3 = \beta$$

The above test for parallelism is only applicable for the single-factor covariance analysis. The covariance analysis is implemented in MINITAB by reading TRT, Y and X into columns C1, C2 and C3 respectively. We immediately obtain the transformed variable XX which equals $X_{ij} - \bar{X}_{..}$, i.e., $X_{ij} - 33.20$. This variable is declared as a covariate in the MINITAB instructions. The results are presented in the following output.

```
MTB > LET C4=C3-MEAN(C3)

Data Display

 Row      TRT      Y      X      XX

   1     BRUS     63     16  -17.2
   2     BRUS     77     45   11.8
   3     BRUS     81     50   16.8
   4     BRUS     60     19  -14.2
   5     BRUS     63     18  -15.2
   6     SPRY     81     48   14.8
   7     SPRY     73     40    6.8
   8     SPRY     59     24   -9.2
   9     SPRY     74     33   -0.2
  10     SPRY     77     41    7.8
  11     DIPP     72     40    6.8
  12     DIPP     54     31   -2.2
  13     DIPP     57     40    6.8
  14     DIPP     59     33   -0.2
  15     DIPP     52     20  -13.2

MTB > GLM 'Y' = TRT;
SUBC>   Covariates 'XX';
SUBC>   Brief 1 ;
SUBC>   Means TRT;
SUBC>   Pairwise TRT;
SUBC>     Tukey;
SUBC>     NoCI.
```

```
General Linear Model: Y versus TRT

Factor     Type Levels Values
TRT       fixed      3 BRUS DIPP SPRY

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
XX          1     752.94     684.59     684.59   37.69  0.000
TRT         2     451.65     451.65     225.83   12.43  0.002
Error      11     199.81     199.81      18.16
Total      14    1404.40

Means for Covariates

Covariate      Mean     StDev
XX        -0.000000     11.44

Least Squares Means for Y

 TRT      Mean   SE Mean
BRUS     71.09     1.942
DIPP     59.05     1.906
SPRY     70.25     1.951

Tukey Simultaneous Tests
Response Variable Y
All Pairwise Comparisons among Levels of TRT

TRT = BRUS subtracted from:

Level    Difference       SE of             Adjusted
TRT       of Means  Difference   T-Value    P-Value
DIPP        -12.04       2.716    -4.433     0.0027
SPRY         -0.84       2.809    -0.300     0.9519

TRT = DIPP subtracted from:

Level    Difference       SE of             Adjusted
TRT       of Means  Difference   T-Value    P-Value
SPRY         11.20       2.734     4.096     0.0046
```

The results agree with our earlier results. The estimated adjusted treatment means also agree with our results. The F test indicate that there are significant differences between the adjusted treatment means (p-value $= 0.002$). A pairwise comparison test using Tukey's test indicate that while both the adjusted treatment means for brushing and spraying are not significantly different, both are however significantly different from the adjusted Dipping treatment mean. This conclusion is displayed in the table below.

Adjusted treatments		
$\hat{t}_{BR}$	$\hat{t}_{SP}$	$\hat{t}_{DP}$

13.3.1 Testing for Parallelism with MINITAB

To test for parallelism with MINITAB, we refit the model, but his time include the interaction term between treatments and covariate i.e., we fit the model,

$$Y_{ij} = \mu + t_i + \beta(X_{ij} - \bar{X}_{..}) + (tx)_{ij} + \epsilon_{ij}; \begin{array}{l} i = 1,2,3 \\ j = 1,2,3,4,5 \end{array} \tag{13.7}$$

where $(tx)_{ij}$ is the interaction term between treatments and centered covariate i.e., $x = (X_{ij} - \bar{X}_{..})$. The result of this fit in MINITAB is presented below:

```
MTB > GLM 'Y' = TRT TRT* XX;
SUBC>   Covariates 'XX';
SUBC>   Brief 1 .

General Linear Model: Y versus TRT


Factor     Type Levels Values
TRT       fixed      3 BRUS DIPP SPRY

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
XX          1     752.94     564.31     564.31   28.40  0.000
TRT         2     451.65     396.35     198.18    9.97  0.005
TRT*XX      2      20.97      20.97      10.48    0.53  0.607
Error       9     178.84     178.84      19.87
Total      14    1404.40
```

If the lines are parallel, we would expect that the interaction term would not be significant. In the above results, the hypotheses that:

$$H_0 : \text{Interaction term is zero}$$

$$H_a : \text{Interaction term is not zero}$$

is tested with the computed F value of 0.53 on 2 and 9 degrees of freedom. This gives a p-value of 0.607 which clearly indicates that we would fail to reject H_0 i.e., the interaction terms can be assumed to be zero. In other words the lines are parallel.

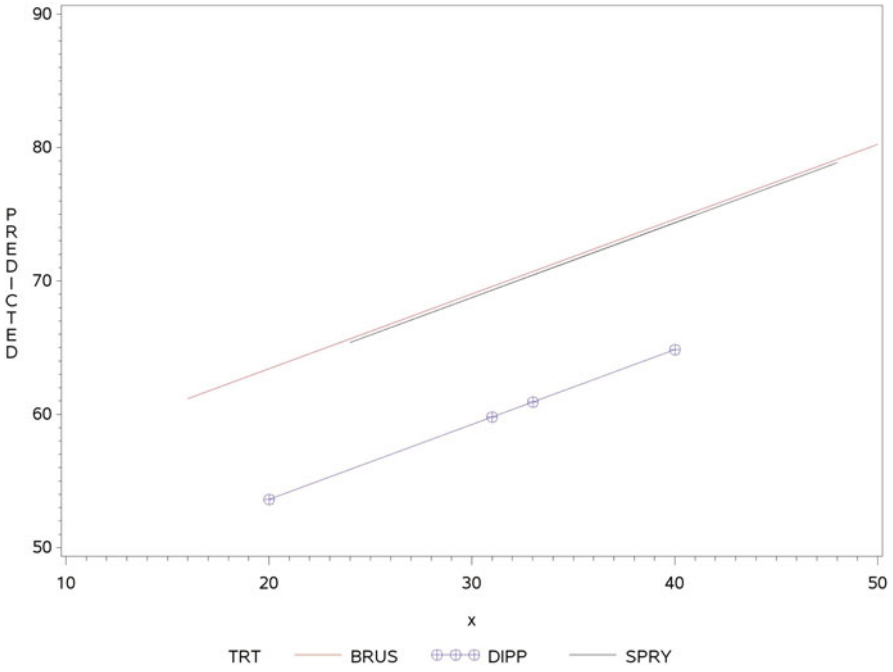


Fig. 13.1 Plot of adjusted treatment means against X

Example 13.3.1

The data below is taken from Ostle and Mensing (1975). It is an experiment in which the gain in weights of pigs for four different feeds were compared. the covariate X was the initial weight of the pig. Pigs were assigned to feeds completely at random. The data is recorded in Table 13.7.
Following the analysis procedure in the preceding section, we have the summarized results for the analysis of covariance for the data in Table 13.7 in Table 13.8.
$F = 536.53/276.58 = 1.94$ which is not significant at the 5 % level, and hence we are unable to reject the hypothesis of no differences among the true effects of the four treatments for the gain in weight of pigs after adjusting for varying initial weights of the experimental animals.

$$\bar{X}_{++} = 29.29, \quad \bar{Y}_{++} = 169.71, \quad \text{and} \quad \hat{\beta} = \frac{496.83}{361.50} = 1.374$$

Table 13.7 Gains in weight Y and individual weights X of Pigs

	Treatments							
	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	30	165	24	180	34	156	41	201
	27	170	31	169	32	189	32	173
	20	130	20	171	35	138	30	200
	21	156	26	161	35	190	35	193
	33	167	20	180	30	160	28	142
	29	151	25	170	29	172	36	189
Total	160	939	146	1031	195	1005	202	1098

Table 13.8 Analysis of covariance

					Deviations about regression		
Source	df	S_X	S_{XY}	S_Y	Error SS	df	MS
Trts.	3	365.46	451.21	21663			
Error	20	361.50	496.83	5937.83	5255.01	19	276.58
Total	23	726.96	948.04	8100.96	6864.61	22	
Difference for the adjusted treatments					1609.6	3	536.03

So that the adjusted treatment means from Eq. (13.2) and their corresponding standard errors from Eq. (13.5) are

Treatment	1	2	3	4
adj. $\hat{t}_i$	160.10	178.65	163.09	176.98
S.E	7.17	8.06	7.35	7.80

The analysis of covariance for the data in Table 13.9 is carried out in MINITAB with the following commands and output.

```
MTB > set c2
DATA> 165 170 130 156 167 151 180 169 171 161 180 170
DATA> 156 189 138 190 160 172 201 173 200 193 142 189
DATA> end
MTB > set c3
DATA> 30 27 20 21 33 29 24 31 20 26 20 25
DATA> 34 32 35 35 30 29 41 32 30 35 28 36
DATA> end
MTB > set c1
DATA> (1:4)6
DATA> end
LET C4=C3-MEAN(C3)

Data Display
 Row  TRT      Y      X          XX

   1    1    165     30      0.7083
   2    1    170     27     -2.2917
   3    1    130     20     -9.2917
   4    1    156     21     -8.2917
```

```
    5     1    167     33    3.7083
    6     1    151     29   -0.2917
    7     2    180     24   -5.2917
    8     2    169     31    1.7083
    9     2    171     20   -9.2917
   10     2    161     26   -3.2917
   11     2    180     20   -9.2917
   12     2    170     25   -4.2917
   13     3    156     34    4.7083
   14     3    189     32    2.7083
   15     3    138     35    5.7083
   16     3    190     35    5.7083
   17     3    160     30    0.7083
   18     3    172     29   -0.2917
   19     4    201     41   11.7083
   20     4    173     32    2.7083
   21     4    200     30    0.7083
   22     4    193     35    5.7083
   23     4    142     28   -1.2917
   24     4    189     36    6.7083

MTB > GLM 'Y' = TRT;
SUBC>   Covariates 'XX';
SUBC>   Brief 1 ;
SUBC>   Means TRT;
SUBC>   Pairwise TRT;
SUBC>     Tukey;
SUBC>     NoCI.

General Linear Model: Y versus TRT

Factor     Type Levels Values
TRT       fixed      4 1 2 3 4

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS      F      P
XX          1    1236.4      682.8      682.8   2.47  0.133
TRT         3    1609.6     1609.6      536.5   1.94  0.157
Error      19    5255.0     5255.0      276.6
Total      23    8101.0

Means for Covariates

Covariate       Mean    StDev
XX         -0.000000    5.622

Least Squares Means for Y

TRT      Mean   SE Mean
1       160.1     7.167
2       178.6     8.056
3       163.1     7.347
4       177.0     7.794
```

The results indicate that there are no significant differences among the adjusted treatment means (p-value $= 0.157$). A test to show if the covariate has been useful in adjusting the treatment means is performed by testing the covariate parameter, i.e.,

$$H_0 : \beta = 0 \tag{13.8}$$

$$H_a : \beta \neq 0 \tag{13.9}$$

Results from the MINITAB ANOVA table gives a p-value of 0.133 for this test, which indicates that we would fail to reject H_0, thus, the covariate has not improved our analysis, and we might just as well carry out the analysis based on the response variable Y alone.

Similarly, a test for the assumption of parallelism gives a SS of 1058.6 on 3 d.f and a calculated F value of 1.35 with a corresponding p-value of 0.295. Again, this indicates that the assumption of parallelism is tenable in this case.

```
Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F       P
XX          1     1236.4      158.7      158.7    0.61   0.448
TRT         3     1609.6      181.4       60.5    0.23   0.874
TRT*XX      3     1058.6     1058.6      352.9    1.35   0.295
Error      16     4196.4     4196.4      262.3
Total      23     8101.0
```

13.4 Covariance Analysis in a RCBD

We present in the next example, the covariance analysis in a randomized complete block design (RCBD).

Example 13.4.1

The data in Table 13.9 gives the yield of three varieties of a certain crop in a Randomized Complete Block Design in four blocks.

Table 13.9 Yields of three varieties of a crop

		Varieties			Block
Block		A	B	C	total
1	X	54	51	57	162
	Y	64	65	72	201
2	X	62	64	60	186
	Y	68	69	70	207
3	X	51	47	46	144
	Y	54	60	57	171
4	X	53	50	41	144
	Y	62	66	61	189
Total	X	220	212	204	636
	Y	248	260	260	768

Where,

$X =$ yield of a plot in a preliminary year under uniformity trial condition
$Y =$ yield on the same plot in the experimental year when the three varieties were used.

Analysis

S_{xx} : **Sum of squares of *X***

$$\text{Blocks SS} = \frac{162^2}{3} + \frac{186^2}{3} + \frac{144^2}{3} + \frac{144^2}{3} - \frac{636^2}{12} = 396$$
$$\text{Varieties SS} = \frac{220^2}{4} + \frac{212^2}{4} + \frac{204^2}{4} - \frac{636^2}{12} = 32$$
$$\text{Total SS} = 54^2 + 51^2 + \cdots + 41^2 - C.F = 514$$
$$\text{Error SS} = 514 - 396 - 32 = 86$$

S_{xy}: **Cross - Products**

$$\text{Blocks SS} = \frac{(162 \times 201)}{3} + \cdots + \frac{(144 \times 189)}{3} - \frac{636x768}{12} = 264$$
$$\text{Varieties SS} = \frac{(220 \times 248)}{4} + \cdots + \frac{(204 \times 260)}{4} - \frac{636 \times 768}{12} = 24$$
$$\text{Total SS} = (54 \times 64) + \cdots + (41 \times 61) - \frac{(636 \times 768)}{12} = 286$$
$$\text{Error SS}286 - 264 + 24 = 46$$

S_{yy}: **Sum of squares of *Y***

$$\text{Blocks SS} = \frac{201^2}{3} + \frac{207^2}{3} + \cdots + \frac{189^2}{3} - \frac{768^2}{12}(C.F) = 252$$
$$\text{Varieties SS} = \frac{240^2}{4} + \cdots + \frac{260^2}{4} - CF = 24$$
$$\text{Total SS} = 64^2 + 65^2 + \cdots + 61^2 - CF = 324$$
$$\text{Error SS} = 324 - 252 - 24 = 48$$

Hence, the analysis of covariance table is presented in Table 13.10.

Table 13.10 Analysis of covariance for data of 13.9

Source	df	S_{xx}	S_{xy}	S_{yy}	Adjusted SS: Error SS	Adjusted SS: df	MS
Blocks	3	396	264	252			
Varieties	2	32	−24	24			
Error	6	86	46	48	23.4	5	4.68
Total	11	514	286	324			
Varieties + Error	8	118	22	72	67.9	7	
Treatment adjusted for average error regression					44.5	2	22.25

The adjusted SS is obtained by noting that for the error line,

$$S_{xy} = 46, \quad S_{xx} = 86, \quad \text{and} \quad S_{yy} = 48.$$

Hence,

$$\text{Fitted SS} = \frac{S_{xy}^2}{S_{xx}} = 24.6 \quad \text{on 1 df.}$$

Therefore, the

$$\text{adjusted SS (reduced SS)} = \text{Total SS} - \text{Fitted SS} = 48 - 24.6 = 23.4$$

The adjusted or reduced SS is based on $(6-1) = 5$ degrees of freedom.

A similar argument leads to the varieties + Error line to give 67.9 on 7 df. Thus $F = 22.25/4.68 = 4.75$, which is significant at $\alpha = 0.10$, thus we can conclude that there are significant differences in the variety means after adjusting for X at $\alpha = 0.10$ but not at $\alpha = 0.05$.

To test the hypothesis $H_o : \beta = 0$ vs $H_1 : \beta \neq 0$. We note that for the Error line, we have the following table:

Source	df	SS	MS	F
Regression	1	24.6	24.6	5.26
Error adjusted	5	23.4	4.68	
Error for unadjusted	6	48	8	

$F(1,5) = 4.06$ at $\alpha = 0.10$, thus we could reject the hypothesis that $\beta = 0$.

The covariance analysis of the data in Table 13.9 is again implemented in MINITAB with the following commands and corresponding output.

```
MTB > LET C5=C2-MEAN(C2)
MTB > PRINT C1-C5

Data Display

Row  BLOCKS     X     Y    VART     XX
  1       1    54    64       A      1
  2       1    51    65       B     -2
  3       1    57    72       C      4
  4       2    62    68       A      9
  5       2    64    69       B     11
```

```
     6        2      60      70       C       7
     7        3      51      54       A      -2
     8        3      47      60       B      -6
     9        3      46      57       C      -7
    10        4      53      62       A       0
    11        4      50      66       B      -3
    12        4      41      61       C     -12
MTB > GLM 'Y' = BLOCKS VART;
SUBC>   Covariates 'XX';
SUBC>   Brief 1 ;
SUBC>   Means VART;
SUBC>   Pairwise VART;
SUBC>     Tukey;
SUBC>     NoCI.

General Linear Model: Y versus BLOCKS, VART

Factor     Type Levels Values
BLOCKS    fixed      4 1 2 3 4
VART      fixed      3 A B C

Analysis of Variance for Y, using Adjusted SS for Tests
Source     DF     Seq SS     Adj SS     Adj MS       F       P
XX          1    159.136     24.605     24.605    5.26   0.070
BLOCKS      3     96.966     77.227     25.742    5.50   0.048
VART        2     44.503     44.503     22.251    4.76   0.070
Error       5     23.395     23.395      4.679
Total      11    324.000

Means for Covariates

Covariate      Mean     StDev
XX         0.000000     6.836

Least Squares Means for Y

VART       Mean    SE Mean
A         60.93      1.178
B         65.00      1.082
C         66.07      1.178
```

The test of significance of the adjusted variety means from the analysis of variance table gives a p-value of 0.070. Thus, while there are no significant differences in adjusted variety means at $\alpha = .05$ level of significance, there is at $\alpha = 0.10$, since p-value in this case is less than 0.10. Further, a test of the covariate parameter, β indicates that the p-value is also 0.070, again we claim that the covariate is not effective at the 5 % significance level, but is at the 10 % significance level. The covariate in my opinion is not strongly effective in this analysis.

A test of parallelism is obtained again by fitting the interaction terms in the model. The results are presented in the following ANOVA Table from MINITAB. Clearly, the p-value here is 0.123, which clearly indicates that

interaction is not present and we would therefore expect the lines to have parallel profiles. In other words, the assumption of parallelism is satisfied in this analysis.

```
Analysis of Variance for Y, using Adjusted SS for Tests

Source      DF      Seq SS      Adj SS      Adj MS       F       P
XX           1     159.136      24.092      24.092   12.50   0.038
BLOCKS       3      96.966      75.263      25.088   13.01   0.032
VART         2      44.503      50.282      25.141   13.04   0.033
VART*XX      2      17.612      17.612       8.806    4.57   0.123
Error        3       5.783       5.783       1.928
Total       11     324.000
```

13.4.1 Factorial Treatment Case

We consider that the following example from Steal & Torrie relating to the case in which the dependent variable is affected by two independent variables (covariates) X_1 and X_2, which measures the initial weight of and forage consumed by guinea pigs, and a dependent variable Y which measured weight gained after 55 days. The data is presented in the following table.

Table 13.11 Initial weight X_1, forage consumed X_2, and gain in weight Y

		Soil treatment					
		Unfertilized			Fertilized		
Soil type	Block	X_1	X_2	Y	X_1	X_2	Y
Miami silt loam	1	220	1155	224	222	1326	237
	2	246	1423	289	268	1559	265
	3	262	1576	280	314	1528	256
Mean		242.7	1384.7	264.3	268.0	1471.0	252.7
Plainfield fine sand	1	198	1092	118	205	1154	82
	2	266	1703	191	236	1250	117
	3	335	1546	115	268	1667	117
Mean		266.3	1447.0	141.3	236.3	1357.0	105.3
Almena silt loam	1	213	1573	242	188	1381	184
	2	236	1730	270	259	1363	129
	3	288	1593	198	300	1564	212
Mean		245.7	1632.0	236.7	249.0	1436.0	175.0
Carlisle	1	256	1532	241	202	1375	239
	2	278	1220	185	216	1170	207
	3	283	1232	185	225	1273	227
Mean		272.3	1328.0	203.7	214.3	1272.7	224.3

The model in this case is:

$$Y_{ij} = \mu + t_i + \beta(X_{ij} - \bar{X}_{..}) + (tx)_{ij} + \epsilon_{ij}; \quad \begin{matrix} i = 1,2,3 \\ j = 1,2,3,4,5 \end{matrix} \tag{13.10}$$

where $(tx)_{ij}$ is the interaction term between treatments and centered covariate i.e., $x = (X_{ij} - \bar{X}_{..})$. The result of this fit in MINITAB is presented below:
The model in this case is:

$$\begin{aligned} Y_{ij} = &\mu + b_i + S_j + F_k + (SF)_{jk} + \beta_1(X_{1ijk} - \bar{X}_{1...}) + \beta_2(X_{2ijk} - \bar{X}_{2...}) \\ &+ (t\mathbf{X})_{ijk} + \epsilon_{ijk}; \quad \begin{matrix} i = 1,2,3 \\ j = 1,2,3,4 \\ k = 1,2 \end{matrix} \end{aligned} \tag{13.11}$$

The adjusted treatment means for $\hat{\bar{y}}_{ij}$ in this case are given by:

$$\hat{y}_{i.} = \bar{Y}_{i.} - \hat{\beta}_{y1.2}(\bar{x}_{1i.} - \bar{x}_{1..}) - \hat{\beta}_{y2.1}(\bar{x}_{2i.} - \bar{x}_{2..}) \tag{13.12}$$

where $\hat{\beta}_{y1.2}$ and $\hat{\beta}_{y1.2}$ are estimated partial regression coefficients and i refer to the $4(2) = 8$ treatment combinations. The MINITAB implementation of the above model, and corresponding output is presented below.

```
MTB > print c1-c6
Data Display

Row   BLK  FERT  SOIL    X1     X2     Y
  1     1     0     1   220   1155   224
  2     1     1     1   222   1326   237
  3     2     0     1   246   1423   289
  4     2     1     1   268   1559   265
  5     3     0     1   262   1576   280
  6     3     1     1   314   1528   256
  7     1     0     2   198   1092   118
  8     1     1     2   205   1154    82
  9     2     0     2   266   1703   191
 10     2     1     2   236   1250   117
 11     3     0     2   335   1546   115
 12     3     1     2   268   1667   117
 13     1     0     3   213   1573   242
 14     1     1     3   188   1381   184
 15     2     0     3   236   1730   270
 16     2     1     3   259   1363   129
 17     3     0     3   288   1593   198
 18     3     1     3   300   1564   212
 19     1     0     4   256   1532   241
 20     1     1     4   202   1375   239
```

```
 21    2     0     4  278  1220  185
 22    2     1     4  216  1170  207
 23    3     0     4  283  1232  185
 24    3     1     4  225  1273  227

MTB > Name c7 "COEF1"
MTB > GLM 'Y' = BLK SOIL| FERT;
SUBC>   Covariates 'X1' 'X2';
SUBC>   Brief 2 ;
SUBC>   Means SOIL FERT SOIL| FERT;
SUBC>   Coefficients 'COEF1'.

General Linear Model: Y versus BLK, SOIL, FERT

Factor  Type   Levels  Values
BLK     fixed       3  1, 2, 3
SOIL    fixed       4  1, 2, 3, 4
FERT    fixed       2  0, 1

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF   Seq SS   Adj SS   Adj MS      F      P
X1          1      4.0   1341.6   1341.6   3.13  0.102
X2          1  13219.3  10585.1  10585.1  24.73  0.000
BLK         2    332.0    395.4    197.7   0.46  0.641
SOIL        3  59945.3  59216.4  19738.8  46.12  0.000
FERT        1   2121.1   1850.6   1850.6   4.32  0.060
SOIL*FERT   3   1136.6   1136.6    378.9   0.89  0.476
Error      12   5135.5   5135.5    428.0
Total      23  81893.8


S = 20.6872   R-Sq = 93.73%   R-Sq(adj) = 87.98%


Term         Coef  SE Coef      T      P
Constant    99.41    69.49   1.43  0.178
X1        -0.4943   0.2792  -1.77  0.102
X2        0.15837  0.03184   4.97  0.000


Means for Covariates

Covariate    Mean   StDev
X1          249.3   38.68
X2         1416.0  191.92


Least Squares Means for Y
```

```
SOIL        Mean  SE Mean
1          259.6    8.594
2          126.5    8.485
3          186.2    9.334
4          229.4    9.146
FERT
0          210.1    6.293
1          190.7    6.293
SOIL*FERT
1    0     266.0   12.079
1    1     253.2   12.918
2    0     144.8   12.770
2    1     108.3   12.445
3    0     200.7   13.987
3    1     171.7   11.963
4    0     229.0   14.269
4    1     229.7   15.153
```

Only the main effect of soil is significant at the 5 % level. Neither the main effect of soil treatment nor the interactions terms are significant. However, for instance, the interaction adjusted treatment means, the adjusted mean for soil 2, and fertilizer 1 (i.e., plainfield fine sand in combination with fertilized soil treatment), for instance is obtained as:

$$\hat{\bar{y}}_{21} = 105.5 - (-0.4943)(236.3 - 249.3) - (0.1584)(1357 - 1416) = 108.4197$$

Since, the partial regression coefficients are estimated as -0.4943 and 0.1584 for X_1 and X_2 respectively. Similarly, the adjusted soil means are computed as:

$$\hat{y}_{.1} = \bar{Y}_{.j.} - \hat{\beta}_{y1.2}(\bar{x}_{1.j} - \bar{x}_{1.j}) - \hat{\beta}_{y2.1}(\bar{x}_{2.j} - \bar{x}_{2.j}) \tag{13.13}$$

For soil 1 for instance this becomes:

$$\hat{\bar{y}}_{.1} = 258.5 - (-0.4943)(255.3 - 249.3) - (0.1584)(1427.83 - 1416) = 259.6$$

Here, $255.3 = \dfrac{(728 + 804)}{6}$ and $1427.83 = \dfrac{(4154 + 4413)}{6}$.

13.5 Estimation of Missing Observations

To illustrate the use of covariance analysis to estimate a missing observation, let us consider the data in Table 11.3 for the Randomized Complete Blocks design in Chap. 10. Suppose as before the value for strain D in block five is missing. We reproduce for clarity below the data for Table 11.3 with the resulting table presented in Table 13.12.

Table 13.12 Yields of wheat in (lb/plot) with one missing observation

Strains		Blocks					Strain total
		1	2	3	4	5	
A	Y	32.3	34.0	34.3	35.0	36.5	172.1
	X	0	0	0	0	0	0
B	Y	33.3	33.0	36.3	36.8	34.5	173.9
	X	0	0	0	0	0	0
C	Y	30.8	34.3	35.3	32.3	35.8	168.5
	X	0	0	0	0	0	0
D	Y	29.3	26.0	29.8	28.0	0	113.1
	X	0	0	0	0	-1	-1
Total	Y	125.7	127.3	135.7	132.1	106.8	627.6
	X	0	0	0	0	-1	-1

For the missing value, we set the observation $Y = 0$. If we define a covariate X such that $X = 0$ for an observed Y and $-1\,(or + 1)$ for $Y = 0$, i.e.,

$$X = \begin{cases} -1 & \text{if observation is missing} \\ 0 & \text{otherwise} \end{cases}$$

To estimate the missing value, we would calculate the following:

S_{yy}: **Sum of Square**

$$\text{Total SS} = 32.3^2 + 33.3^2 + \cdots + 35.8^2 + 0^2 - \frac{627.6^2}{20} = 1201.692$$

$$\text{Strains SS} = \frac{172.1^2}{5} + \cdots + \frac{113.1^2}{5} - C.F = 514.608$$

$$\text{Blocks SS} = \frac{125.7^2}{4} + \cdots + \frac{106.8^2}{4} - C.F = 125.142$$

$$\text{Error SS} = \text{Total} - \text{Blocks SS} - \text{Strain SS} = 561.942$$

S_{xy}: **Sum Squares of Cross-Products (SSCP)**

$$\text{Total SSCP} = \frac{(0 \times 323)}{5} + \cdots + \frac{(-1 \times 0)}{5} - \frac{(-1 \times 627.6)}{20} = 31.38$$

$$\text{Strains SSCP} = \frac{(0 \times 172.1)}{5} + \cdots + \frac{(-1 \times 113.1)}{5} - \frac{(-1 \times 627.6)}{20} = 8.76$$

$$\text{Blocks SSCP} = \frac{(0 \times 125.7)}{4} + \cdots + \frac{(-1 \times 106.8)}{4} - \frac{(-1 \times 627.6)}{20} = 4.68$$

$$\text{Error SSCP} = \text{Total} - \text{Strains} - \text{Blocks} = 17.94$$

S_{xx}: **Sum of Squares of X**

$$\text{Total SS} = 0^2 + 0^2 + \cdots + (-1)^2 - \frac{(-1)^2}{20} = 0.95$$

$$\text{Strains SS} = \frac{0^2}{5} + \cdots + \frac{(-1)^2}{5} - \frac{(-1)}{20} = 0.15$$

$$\text{Blocks SS} = \frac{0^2}{4} + \cdots + \frac{(-1)^2}{4} - \frac{(-1)}{20} = 0.20$$

$$\text{Error SS} = \text{Total} - \text{Strains} - \text{Block} = 0.60$$

The missing value is given by:

$$\hat{x}_0 = \frac{\text{Error } S_{xy}}{\text{Error } S_{xx}} = \frac{17.94}{0.60} = 29.9 \tag{13.14}$$

which is the same value as that obtained in example 11.4.

We can now perform the Analysis of Covariance.

Source	df	S_{xx}	S_{xy}	S_{yy}	df	Adjusted SS(y)	MS
Total	19	0.95	31.38	1201.692			
Blocks	4	0.20	4.68	125.142			
Strains	3	0.15	8.76	514.608			
Error	12	0.60	17.94	561.942	11	25.536	2.32
Treatments + Error	15	0.75	26.7	1076.55	14	126.03	
Treatments adjusted					3	100.494	33.50

A test of the hypotheses:

$$H_0: \quad \mu_A = \mu_B = \mu_C = \mu_D$$

$$H_a: \quad \text{at least two of the adjusted means are unequal}$$

is tested by computing

$$F = \frac{33.50}{2.32} = 14.44$$

This is compared with an $F_{3,11}(0.95) = 3.59$, indicating that there are significant differences between the adjusted strain means. The MINITAB implementation is carried out with the following:

```
Data Display

   Row   STRAINS  BLOCKS      Y     X
     1         A       1   32.3     0
     2         A       2   34.0     0
     3         A       3   34.3     0
     4         A       4   35.0     0
     5         A       5   36.5     0
     6         B       1   33.3     0
     7         B       2   33.0     0
     8         B       3   36.3     0
     9         B       4   36.8     0
    10         B       5   34.5     0
    11         C       1   30.8     0
    12         C       2   34.3     0
    13         C       3   35.3     0
    14         C       4   32.3     0
    15         C       5   35.8     0
    16         D       1   29.3     0
    17         D       2   26.0     0
    18         D       3   29.8     0
    19         D       4   28.0     0
    20         D       5    0.0    -1

MTB > GLM 'Y' = BLOCKS STRAINS;
SUBC>   Covariates 'X';
SUBC>   Brief 1 ;
SUBC>   Means STRAINS.


General Linear Model: Y versus BLOCKS, STRAINS

Factor     Type Levels Values
BLOCKS    fixed      5 1 2 3 4 5
STRAINS   fixed      4 A B C D

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS        F      P
X           1    1036.53     536.41     536.41   231.06  0.000
BLOCKS      4      39.13      21.97       5.49     2.37  0.117
STRAINS     3     100.49     100.49      33.50    14.43  0.000
Error      11      25.54      25.54       2.32
Total      19    1201.69

Means for Covariates

Covariate      Mean     StDev

X          -0.05000    0.2236

Least Squares Means for Y


STRAINS     Mean   SE Mean
A          32.92    0.6885
B          33.28    0.6885
C          32.21    0.6885
D          27.11    0.7425
```

The F-value obtained from the covariance analysis in MINITAB is 14.43 with a p-value of 0.000. This value compares well with the value obtained from hand calculations. Further, the p-value for the effect of the covariate is 0.000 which indicates that the covariate is very effective in this analysis.

13.5.1 Another Missing Value Example

The missing value can be estimated as the negative of the slope parameter where, slope is estimated by $\hat{\beta} = E_{xy}/E_{xx}$, if we just apply covariance analysis the structure of the data in Table 13.12.

```
MTB > GLM 'Y' = BLOCKS STRAINS;
SUBC>   Covariates 'X';
SUBC>   Brief 3 ;
SUBC>   Means STRAINS.

General Linear Model: Y versus BLOCKS, STRAINS

Factor   Type   Levels  Values
BLOCKS   fixed       5  1, 2, 3, 4, 5
STRAINS  fixed       4  A, B, C, D


Analysis of Variance for Y, using Adjusted SS for Tests

Source   DF   Seq SS  Adj SS  Adj MS       F      P
X         1  1036.53  536.41  536.41  231.06  0.000
BLOCKS    4    39.13   21.97    5.49    2.37  0.117
STRAINS   3   100.49  100.49   33.50   14.43  0.000
Error    11    25.54   25.54    2.32
Total    19  1201.69


S = 1.52363   R-Sq = 97.87%   R-Sq(adj) = 96.33%


Term          Coef  SE Coef      T      P
Constant   32.8750   0.3546  92.71  0.000
X           29.900    1.967  15.20  0.000
BLOCKS
1          -1.4500   0.6885  -2.11  0.059
2          -1.0500   0.6885  -1.53  0.155
3           1.0500   0.6885   1.53  0.155
4           0.1500   0.6885   0.22  0.832
STRAINS
A           1.5450   0.5982   2.58  0.025
B           1.9050   0.5982   3.18  0.009
C           0.8250   0.5982   1.38  0.195


Means for Covariates

Covariate      Mean   StDev
X          -0.05000  0.2236
```

```
Least Squares Means for Y

STRAINS    Mean  SE Mean
A         32.92   0.6885
B         33.28   0.6885
C         32.20   0.6885
D         27.11   0.7425
```

The estimated missing value is estimated as 29.9, the estimate of the covariate parameter. We also note that MINITAB automatically adjusts the error degrees of freedom accordingly.

13.6 Exercises

1. Given the following data:

Source of variation	df	S_{xx}	S_{xy}	S_{yy}
Replicates	4	100	140	400
Treatments	10	100	100	900
Error	40	400	900	2500

   (a) With the above summary statistics, carry out a full analysis of covariance. What conclusions may be drawn about the effect of treatments.
   (b) Test the regression coefficient based on the experimental error at $= \alpha = 0.05$.

2. Analyze the following 5×5 Latin Square experiment on the yield in bags per acre of No.1 Irish potatoes (Y), adjusted for the percentage of No. 1's (X). The treatments were different amounts (pounds) of P_2O_5 per acre: $a = 0, b = 40, c = 80, d = 120, e = 160$. Partition the adjusted treatment SS into relevant components.

	Columns														
	1			2			3			4			5		
Rows	t	Y	X	t	Y	X	t	Y	X	t	Y	X	t	Y	X
1	a	134.0	91	b	149.1	88	c	141.3	87	d	161.3	91	e	149.2	91
2	b	148.5	90	d	148.5	91	e	199.3	94	a	148.5	90	c	152.7	93
3	c	145.2	93	e	149.5	95	a	119.9	90	b	149.2	94	d	145.8	90
4	d	171.1	91	c	169.0	94	b	144.9	89	e	170.8	95	a	130.4	88
5	e	175.8	91	a	153.4	94	d	168.9	92	c	167.6	96	b	141.5	93
Total		774.6	456		769.5	462		774.3	452		797.4	466		719.6	455

3. A rehabilitation center researcher was interested in examining the relationship between physical fitness prior to surgery of people undergoing corrective knee surgery, and time required in physical therapy until successful rehabilitation. Patients records at the center for 24 males whose age range from 18 to 30 years, and who had undergone corrective surgery

were selected. The number of days required for successful completion of physical therapy y, and the prior physical status (low, average, high) as factor levels for each patient as well as their ages (x) are presented below (Neter et al. 1990).

Factor		Observations									
level	Var	1	2	3	4	5	6	7	8	9	10
Low	Y	29	42	38	40	43	40	30	32		
	X	18.3	30.0	26.5	28.1	29.7	27.8	19.8	29.3		
Average	Y	30	35	39	28	31	31	29	35	29	33
	X	20.8	25.2	29.2	20.0	21.5	22.1	19.7	24.7	20.2	22.9
High	Y	26	32	21	20	23	22				
	X	22.7	28.7	18.9	18.0	21.7	20.0				

(a) Carry out a covariance analysis of the above data employing age as a concomitant variable.
(b) Test for treatment and adjusted treatment effects, use $\alpha = 0.05$.
(c) Conduct a multiple comparison tests on the adjusted treatment means.
(d) Is the covariate important for this study?

4. A horticulturist conducted an experiment to study the effects of flower variety (factor A: varities LP, WB), and moisture level (factor B: low, high) on yields of salable flower's (Y). Because the plots were not of the same size, the horticulturist wished to use plot size (X) as the concomitant variable. Six replications were made for each treatment, and the data is presented below (Neter et al. 1990).

	Factor B			
	B_1 (low)		B_2 (high)	
Factor A	Y	X	Y	X
A_1	98	15	71	10
	60	4	80	12
(*variety LP*)	77	7	86	14
	80	9	82	13
	95	14	46	2
	64	5	55	3
A_2	55	4	76	11
	60	5	68	10
(*variety WB*)	75	8	43	2
	65	7	47	3
	87	13	62	7
	78	11	70	9

(a) Carry out a covariance analysis for the above data and test for adjusted main and interaction effects.
(b) What are the estimated regression coefficients.

Chapter 14
Factorial Treatments Designs

14.1 Definitions

Factor-This is an independent variable for study in an experiment, and the first thing to consider in an experiment is to determine the factor or variable to be investigated. For example, let us suppose that we are interested in investigating the effect of inorganic chemicals on plant growth over the life period of the plant. Since many inorganic chemicals are involved we select one of them, say, nitrogen, in the nitrate form, NO_3. Quite a different result may be obtained with a different form, say, NH_3. The same amount of nitrogen could be used for the two forms, but the effect on plant response or growth could be quite different for the two forms of nitrogen.

Nitrogen in this example is the factor to be investigated, while the amounts of NO_3 to be used are referred to as the levels of the factor.

14.1.1 Factorial Design

The experimenter may wish to study two or more factors jointly to observe the manner in which the response varies with the changing levels of the factors under study. For example, suppose an agronomist is investigating the yield of maize. He decided to use various amounts of nitrogen (N) and phosphorous (P). Suppose that it is decided to use the following levels of N and P:

Levels of N – 10, 20, 40, 80 g.

Levels of P – 13, 39, 65, 91 g.

Now the question arises as to what combinations of the various levels of the two factors to use in the investigation. Well, why not use all the treatment combinations in Table 14.1?

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_14,

Table 14.1 The 16 treatment combinations

Levels of N	Levels of P (g)			
	$13=0$	$39=1$	$65=2$	$91=3$
$10=0$	$n_0p_0 = 00$	$n_0p_1 = 01$	$n_0p_2 = 02$	$n_0p_3 = 03$
$20=1$	$n_1p_0 = 10$	$n_1p_1 = 11$	$n_1p_2 = 12$	$n_1p_3 = 13$
$40=2$	$n_2p_0 = 20$	$n_2p_1 = 21$	$n_2p_2 = 22$	$n_2p_3 = 23$
$80=3$	$n_3p_0 = 30$	$n_3p_1 = 31$	$n_3p_2 = 32$	$n_3p_3 = 33$

The combination of the ith level of N and of the jth level of P is denoted as n_ip_j. Once the order of the subscript has been defined, the subscript ij is sufficient to define the treatment corresponding to the combination n_ip_j of the two factors N and P. The treatment design for the two factors described above is known as a factorial experiment.

A factorial treatment design is one which contains all combinations of levels of the various factors. That is, if there are X levels of factor A and Y levels of factor B, then each replicate contains all $X \times Y$ treatment combinations. It has been shown (Kempthorne 1952) that if the purpose of an experiment utilizing factorial treatment design is to estimate main effects and interactions, then the factorial design is optimum, that is, no other selection of treatments does this more effectively than a factorial with equal numbers of observations on each treatment.

In most agricultural experiments, factors tend to interact with one another, and the factorial experiments are most appropriate for this kind of situations.

14.2 Experiments at Two Levels: The 2^n Series

A 2^n factorial experiment is an experiment involving n factors each at two levels designated 0, 1. The simplest of the design is when $n = 2$, that is, two factors each at two levels—2^2. For this situation, suppose the two factors are A and B, with the levels designated 0, 1, respectively. Then, there will be four (2×2) treatment combinations (00), (10), (01), and (11).

If a_0 and b_0 denote the zero level for both factors and a_1 and b_1 also denote the upper levels of the two factors, then the four treatment combinations can be put in a table as follows:

Factor A	Factor B	
	b_0	b_1
a_0	a_0b_0	a_0b_1
a_1	a_1b_0	a_1b_1

These are sometimes written as follows.

$$(1) = a_0 b_0$$
$$\mathrm{a} = a_1 b_0$$
$$\mathrm{b} = a_0 b_1$$
$$\mathrm{ab} = a_1 b_1$$

Since there are four treatment combinations, it follows that there are three degrees of freedom for treatments made up as shown in Table 14.2.

Table 14.2 Structure of ANOVA table

Source	d.f.	
A	1	Main effect of A
B	1	Main effect of B
AB	1	Interaction effect of A and B.

(a) The *main effect* of a factor is a measure of the change in the level of the factor averaged over all levels of the other factors.
(b) The *interaction* is the differential response to one factor in combination with varying levels of a second factor applied simultaneously. That is, interaction is an additional effect due to the combined influence of two (or more) factors.

14.2.1 Factorial Effects in the 2^2 Factorial

Consider the 2×2 table of means in Table 14.3. If we let $\mu_{ij}, i = 0, 1; j = 0, 1$ be the expected response from treatment combination ij, then $\mu_{ij} = \{\mu_{00}, \mu_{01}, \mu_{10}, \mu_{11}\}$.

Simple Effects

The *simple effect* of A at level b_0 of B is defined as:

$$\mu[A\, B_0] = \mu_{10} - \mu_{00}. \tag{14.1}$$

That is, the simple effect of A at level b_0 of B is the amount of change in the expected response when the level of A is changed from a_1 to a_0, with the level of B held constant at b_0. Similarly, the simple effect of A at level b_1 of B is defined as:

$$\mu[A\, B_1] = \mu_{11} - \mu_{01} \tag{14.2}$$

which again can be interpreted as the amount of change in the response when level of A is changed from a_1 to a_0, while the level of B is kept constant at b_1.

Table 14.3 Population means and simple effects in a 2^2 factorial

Level of A	Level of B		Simple effects of B
	b_0	b_1	$\mu[A_i\ B]$
a_0	μ_{00}	μ_{01}	$\mu[A_0\ B] = \mu_{01} - \mu_{00}$
a_1	μ_{10}	μ_{11}	$\mu[A_1\ B] = \mu_{11} - \mu_{10}$
Simple effect of A $\mu[A\ B_j]$	$\mu[A\ B_0] = \mu_{10} - \mu_{00}$	$\mu[A\ B_1] = \mu_{11} - \mu_{01}$	

The main effect of A, therefore, denoted as $\mu[A]$ is defined as the average of the simple effects $\mu[A\ B_0]$ and $\mu[A\ B_1]$ in (14.1) and (14.2), respectively.

$$\begin{aligned}\mu[A] &= \frac{1}{2}\{\mu[A\ B_1] + \mu[A\ B_0]\}. \\ &= \frac{1}{2}(\mu_{11} - \mu_{01} + \mu_{10} - \mu_{00})\end{aligned} \tag{14.3}$$

Similarly, the main effect of B is defined as the average of the simple effects of B at a_0 and a_1, respectively. That is,

$$\begin{aligned}\mu[B] &= \frac{1}{2}\{\mu[A_1\ B] + \mu[A_0\ B]\} \\ &= \frac{1}{2}(\mu_{11} - \mu_{10} + \mu_{01} - \mu_{00}).\end{aligned} \tag{14.4}$$

Both main effects can be estimated from table of observed means $\bar{Y}_{ij}$ as,

$$\hat{\mu}[A] = \frac{1}{2}(\bar{Y}_{11} - \bar{Y}_{01} + \bar{Y}_{10} - \bar{Y}_{00}) \tag{14.5a}$$

$$\hat{\mu}[B] = \frac{1}{2}(\bar{Y}_{11} - \bar{Y}_{10} + \bar{Y}_{01} - \bar{Y}_{00}). \tag{14.5b}$$

Example 14.1.1

Consider the following table of means from a 2×2 factorial experiment with factors A and B.

	B	
A	b_0	b_1
a_0	33	63
a_1	22	52

The simple effects of A at b_0 and b_1 are, respectively, $22 - 33 = -11$ and $52 - 63 = -11$. Hence, main effect of A is $\frac{1}{2}(-11 - 11) = -11$ or it could have been computed as $\frac{1}{2}(52 - 63 + 22 - 33) = -11$. Similarly the main effect of B is $\frac{1}{2}(52 - 22 + 63 - 33) = 30$.

Interaction Effects

Refer again to 2×2 table of population means in Table 14.3. Factors A and B are said to have interaction or to *interact* if the simple effect of A changes with the level of B. Thus, the interaction term $\mu[AB]$ is defined as:

$$\mu[AB] = \frac{1}{2}(\mu[A\ B_1] - \mu[A\ B_0]) = \frac{1}{2}(\mu[A_1\ B] - \mu[A_0\ B]). \tag{14.6}$$

If there is no interaction, then $\mu[AB] = 0$. Thus, a nonzero value for $\mu[AB]$ is an indication of the presence of interaction between factors A and B.

Example 14.1.2

Consider again the table of means from a 2×2 factorial experiment with factors A and B used in the previous example. The simple effects from four different table of means (a)–(d) are presented in Table 14.4.

Table 14.4 Simple and interaction effects for four different tables of means

A	b_0	b_1	$\mu[A_iB]$
a_0	33	63	30
a_1	22	52	30
$\mu[AB_j]$	−11	−11	

(a)

A	b_0	b_1	$\mu[A_iB]$
a_0	12	32	20
a_1	4	10	6
$\mu[AB_j]$	−8	−22	

(b)

A	b_0	b_1	$\mu[A_iB]$
a_0	33	13	−20
a_1	22	42	20
$\mu[AB_j]$	−11	29	

(c)

A	b_0	b_1	$\mu[A_iB]$
a_0	16	23	7
a_1	34	23	−11
$\mu[AB_j]$	18	0	

(d)

In Fig. 14.1, which corresponds to Table (a) in Table 14.4, $\mu[AB] = 0$ because the simple effects of B are the same at both levels of A. Thus, the figure depicts the case when no interaction is present, which leads to parallel lines.

(a) Identical simple effects with $\mu[AB] = 0$
(b) Unequal simple effects with the same signs with $\mu[AB] = -7$
(c) Unequal simple effects with opposite signs. Here $\mu[AB] = 40$
(d) Unequal simple effects with the same signs and has $\mu[AB] = -18$

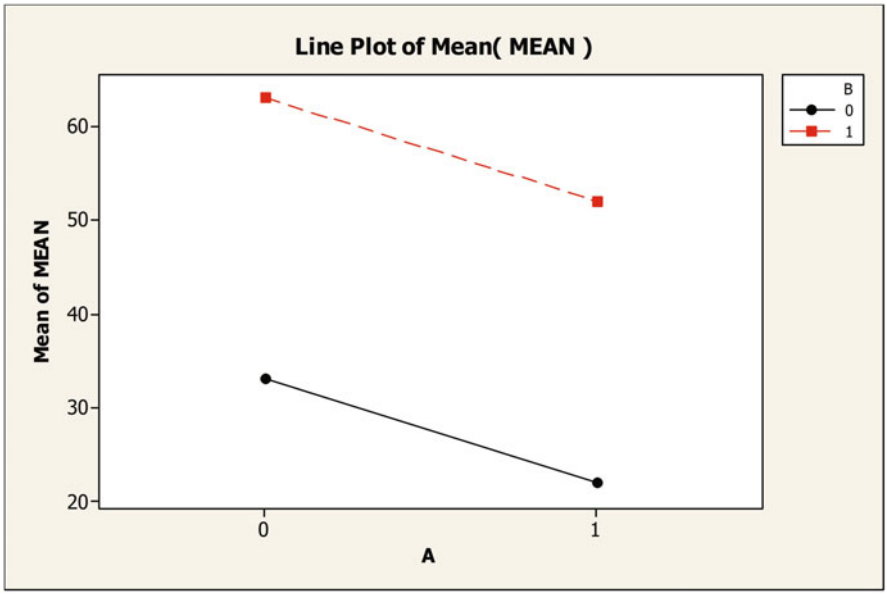


Fig. 14.1 (a) Identical simple effects

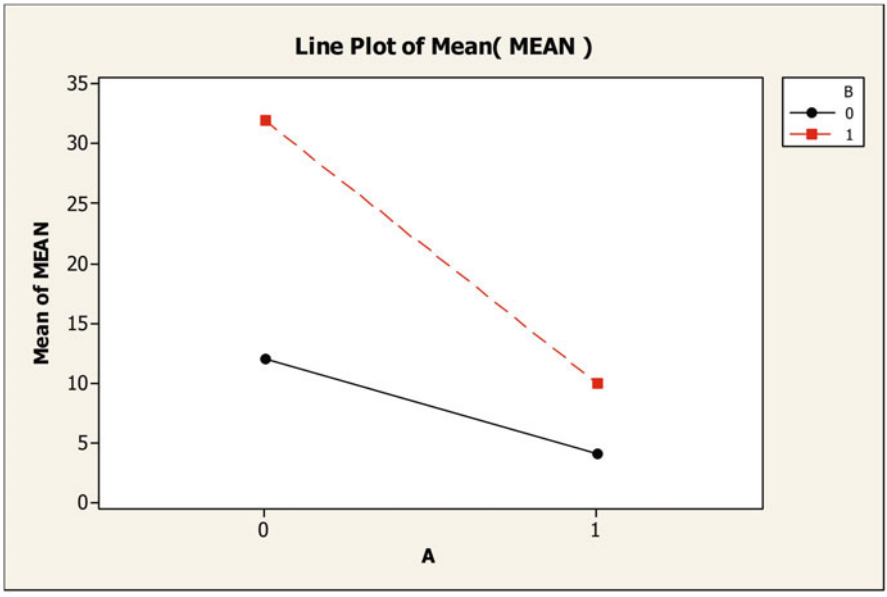


Fig. 14.2 (b) Unequal simple effects with the same signs

In Fig. 14.2 which corresponds to Table (b) in the table of means in Table 14.4, we have the simple effects of B changing with the level of A (and vice versa), indicating the presence of interaction. The interaction effect here is $\mu[AB] = -7$. The simple effects of A are both negative, while those of B are both positive. For A, the expected response decreases from a_0 to a_1 at both levels of B. The interaction presented in Fig. 14.2 therefore represents a *quantitative interaction*, because changing the levels of any one factor results in a change in the magnitude of the simple effects (but not the direction) of the other factor. Further, both lines in Fig. 14.2 have downward slopes and quantitative interaction has a pattern of having lines not being parallel but have the same direction for their slopes.

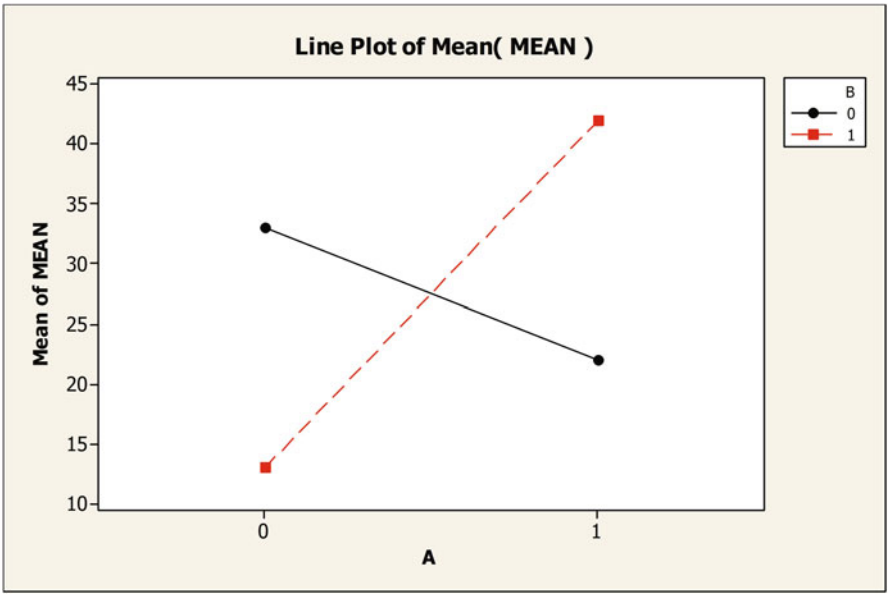


Fig. 14.3 Unequal simple effects with opposite signs

In Fig. 14.3 which also corresponds to Table (c) in the table of means in Table 14.4, we have the simple effect of B at a_0 is negative, while its simple effect at a_1 is positive. That is, the expected responses of Factor B decreases at a_0 and increases at a_1. The interaction therefore, is due to the difference in the signs of the simple effects. The interaction plotted in Fig. 14.3 represents therefore a *qualitative interaction* because changing the level of any one factor results in a change in the direction (sign, $-$ to $+$) of the simple effect of the other factor. Again, the pattern of the plot in this figure is nonparallel but the slopes have different directions.

The interaction plot in Fig. 14.4 is similar to that in Fig. 14.3 except that it is a variation of the former. Here too, the pattern is that the lines are not parallel and have different slopes except that there is no increase in the simple effect of A at b_1 and, therefore, is flat.

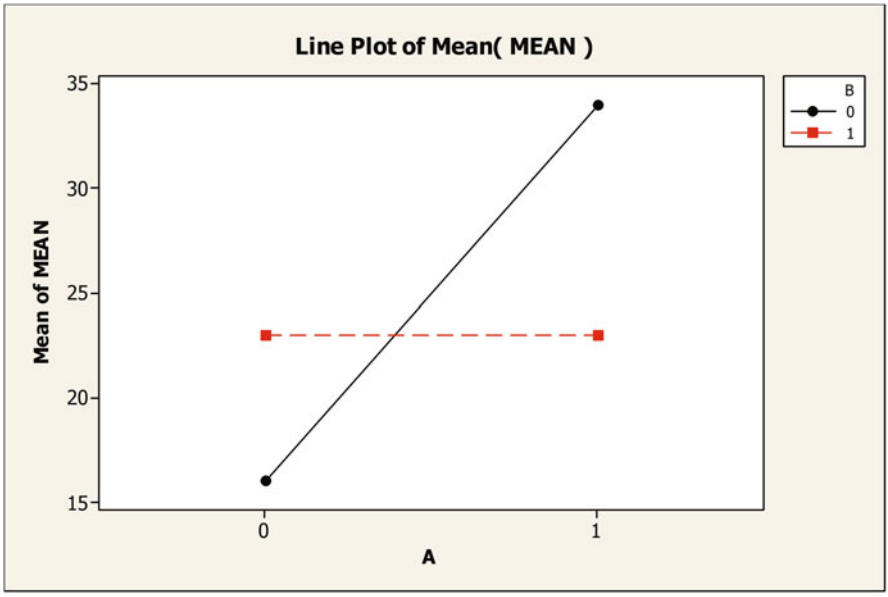


Fig. 14.4 Unequal simple effects with same signs

14.2.2 The 2^3 factorial system

For the 2^3 experiment in which we have three factors each at two levels, there will be $2 \times 2 \times 2 = 8$ treatment combinations namely,

$$
\begin{aligned}
(1) &= a_0b_0c_0 \\
\mathrm{a} &= a_1b_0c_0 \\
\mathrm{b} &= a_0b_1c_0 \\
\mathrm{ab} &= a_1b_1c_0 \\
\mathrm{c} &= a_0b_0c_1 \\
\mathrm{ac} &= a_1b_0c_1 \\
\mathrm{bc} &= a_0b_1c_1 \\
\mathrm{abc} &= a_1b_1c_1
\end{aligned}
$$

and seven main effects and interactions given by A, B, AB, C, AC, BC and ABC corresponding to the breakdown of seven treatments degrees of freedom. The order in which both the treatment combinations, factor effects and interactions are written is very important.

Example 14.2.2

A study was conducted to determine the influence of plant density and varieties on corn (*Zea mays* L.) yield. The experiment was a $2 \times 2 \times 2$ factorial replicated four times in a randomized complete design. The data are displayed in Table 14.5.

In Table 14.6, the eight treatment combinations totals are presented. These are extracted from Table 14.5.

$$\text{Total SS} = 140^2 + 138^2 + \cdots + 132^2 - \text{CF} = 605163 - \frac{4397^2}{32}$$

$$= 987.7188$$

$$\text{Treatment SS} = \frac{550^2}{4} + \cdots + \frac{532^2}{4} - \frac{4397^2}{32} = 605.9688$$

The analysis of variance Table (ignoring blocks for now) for the data is displayed in Table 14.7.

The seven treatment degrees of freedom can be broken into the following, each with 1 d.f.

A	1
B	1
AB	1
C	1
AC	1
BC	1
ABC	1

Table 14.5 Yield of corn in this 2^3 factorial experiment

Variety	Spacing (in)	Density (plants/acre)	Replications I	II	III	IV
A	12	12,000	140	138	130	142
		16,000	145	146	150	147
			435	433	426	439
	25	12,000	136	132	134	138
		16,000	140	134	136	140
			421	404	408	420
B	12	12,000	142	132	128	140
		16,000	146	136	140	141
			436	408	410	421
	25	12,000	132	130	136	134
		16,000	138	132	130	132
			410	396	396	402

Table 14.6 Treatment combination totals

(1)	= 550
a	= 542
b	= 540
ab	= 532
c	= 588
ac	= 563
bc	= 550
abc	= 532

Table 14.7 Analysis of variance table

Source	d.f.	SS	MS	F
Treatments	7	605.9688	86.5670	5.44
Error	24	381.7500	15.9063	
Total	31	987.7188		

In order to obtain these sums of squares (SSs), we draw up a series of two-way tables as shown in Table 14.8.

Table 14.8 Series of Two-way tables of observed yields

	C_0	C_1		A_0	A_1		B_0	B_1
B_0	1092	1151	C_0	1090	1074	A_0	1138	1090
B_1	1072	1082	C_1	1138	1095	A_1	1105	1064
	2164	2233		2228	2169		2243	2154

From Table 14.8, we calculate the following.

The SSs for C, that is SS(C) is given by

$$\begin{aligned} \text{SS(C)} &= \frac{2164^2}{16} + \frac{2233^2}{16} - \text{CF} \quad \text{or as,} \\ &= \frac{(2164 - 2233)^2}{32} \\ &= 148.7813 \end{aligned}$$

as the sums of 2164 and 2233 each came from 16 observations. Similarly for A and B, we have,

$$\begin{aligned} \text{SS(A)} &= \frac{2228^2}{16} + \frac{2169^2}{16} - \text{CF} \quad \text{or as,} \\ &= \frac{(2228 - 2169)^2}{32} \\ &= 108.7813 \end{aligned}$$

$$\text{SS(B)} = \frac{2243^2}{16} + \frac{2154^2}{16} - \text{CF} \quad \text{or as,}$$

$$= \frac{(2243 - 2154)^2}{32}$$
$$= 247.5313.$$

For the two-factor interactions, we have,

$$\text{SS(AB)} = \frac{1138^2 + 1090^2 + 1105^2 + 1064^2}{8} - \text{CF} - \text{SS(A)} - \text{SS(B)}$$
$$\text{or as} \quad \frac{(1138 + 1064 - 1090 - 1105)^2}{32}$$
$$= 1.5312.$$

Similarly,

$$\text{SS(AC)} = \frac{1090^2 + 1138^2 + 1074^2 + 1095^2}{8} - \text{CF} - \text{SS(A)} - \text{SS(C)} = 22.7813$$

and

$$\text{SS(BC)} = 75.0313.$$

For the three-factor interaction, the SS(ABC) is computed as,

$$\text{SS(ABC)} = \text{Total treatment SS} - \text{SS(A)} - \text{SS(B)} - \text{SS(C)} - \text{SS(AB)}$$
$$- \text{SS(AC)} - \text{SS(BC)}$$
$$= 1.5312500.$$

The revised analysis of variance (again, ignoring blocks) is given in Table 14.9

We present in Fig. 14.5 the main and interaction plot matrix from the above analysis as generated in MINITAB. This is a 3×3 display. The diagonals give the main effects of A, B, and C, respectively. Cell (1,2) gives the interaction plot of AB, while cell (3,2) similarly gives the interaction plot of BC. From the matrix, possible interaction effects exist for AC and BC. Of the two, only the BC interaction is highly significant at the 5 % point since $F_{(1,24)}(0.95) = 4.26$.

Table 14.9 The revised ANOVA table

Source	d.f.	SS	MS	F
A	1	108.7813	108.7813	6.84
B	1	247.5313	247.5313	15.56
AB	1	1.5313	1.5313	0.10
C	1	148.7813	148.7813	9.35
AC	1	22.7813	22.7813	1.43
BC	1	75.0313	75.0313	4.72
ABC	1	1.5313	1.5313	0.10
Treatments	7	605.9688	86.5670	5.44
Error	24	381.7500	15.9063	
Total	31	987.7188		

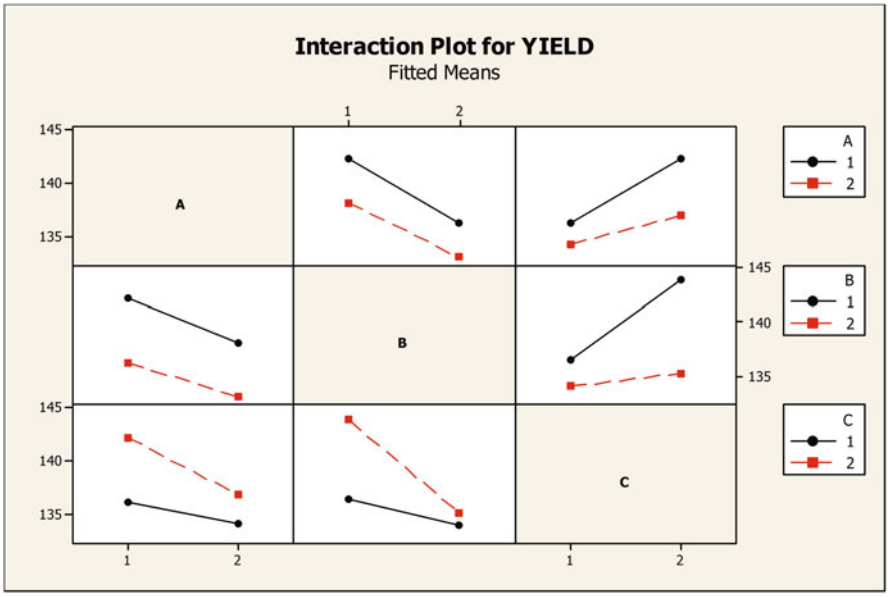


Fig. 14.5 Main and interaction plots matrix

14.2.3 Yates Algorithm

We may also employ Yates algorithm for the calculation of the sum of squares for the data in Table 14.6, which contains the treatment totals. To do this, first we arrange the treatment combination in the order earlier suggested in the last section, thus:

Table 14.10 Successive calculations based on Yates algorithm

Treatment combinations	Response	(1)	(2)	(3)	(4)	Effects and interactions
(1)	550	1092	2164	4397		
a	542	1072	2233	−59	108.7813	A
b	540	1151	−16	−89	247.5313	B
ab	532	1082	−43	7	1.5313	AB
c	588	−8	−20	69	148.7813	C
ac	563	−8	−69	−27	22.7813	AC
bc	550	−25	0	−49	75.0313	BC
abc	532	−18	7	7	1.5313	ABC

To use Yates method, we first list all treatment combinations in the first columns as in Table 14.6. We next place the total response to each of these treatment combinations in the second column. For the third column, labeled (1), add the responses in pairs, e.g., 550 + 542 = 1092, 540 + 532 = 1072, 550

$+ 532 = 1082$. This completes the first half of column (1). For the second half, subtract the responses in pairs, always subtracting the first from the second e.g. $542 - 550 = -8$, $532 - 540 = -8$, $532 - 550 = -18$. This complete column (1). Proceed in the same manner until the nth column is reached: (1), (2), $\cdots$. In our case $n = 3$, so there are just three columns (1), (2) and (3). The values in column n are the constants, where the first entry is $r.2^{n-1}$ times the grand total. The last column gives the corresponding contrasts to the treatment combinations in column one.

To obtain corresponding SS, square the total for each contrast and divide by $2^n r$. In our example, $n = 3$ and $r = 4$. Thus, $\text{SS(A)} = \dfrac{(-59)^2}{32} = 108.7813$. Similarly $\text{SS(ABC)} = \dfrac{(7)^2}{32} = 1.5313$. We can therefore proceed as before and conduct the usual F tests.

$$\text{Average effect of A} = \frac{2169 - 2228}{16} = -3.6875$$

$$\text{Average effect of B} = \frac{2154 - 2243}{16} = -5.5625$$

$$\text{Average effect of C} = \frac{2233 - 2164}{16} = 4.3125$$

The standard error (s.e.) for any of these differences $= \sqrt{\frac{2S^2}{16}} = 1.4107$. Since the BC interaction was significant, we can write this in the form:

	C_0	C_1	$C_1 - C_0$
B_0	1092	1151	59
B_1	1072	1082	10
$C_1 - C_0$	−20	−69	

$$\text{Average effect of B in absence of C} = \frac{1072 - 1092}{8} = -2.50$$

$$\text{Average effect of B in presence of C} = \frac{1082 - 1151}{8} = -8.625$$

The standard error of either of these two average effects

$$\sqrt{\frac{2S^2}{8}} = 1.7836.$$

Difference in effects depending on absence or presence of C

$$= -2.5 + 8.625 = 6.125$$

with an s.e. of

$$\sqrt{\frac{S^2}{8} + \frac{S^2}{8} + \frac{S^2}{8} + \frac{S^2}{8}} = \sqrt{\frac{4S^2}{8}} = 2.8201.$$

The yield of A on average is −3.6875, that is, level 2 of A produces a lower average yield than level 1 of A. This difference has an S.E. (24 d.f.) of 1.4107 and is therefore significant at 5 % level. There was little sign of the effects of B and C being different for the two levels of A (AB and AC not significant). The interaction between B and C was significant at 1 %, the four treatment means being:

	C_0	C_1	$C_1 - C_0$
B_0	136.500	143.875	7.375
B_1	134.000	135.250	1.250
$B_1 - B_0$	−2.500	−8.625	

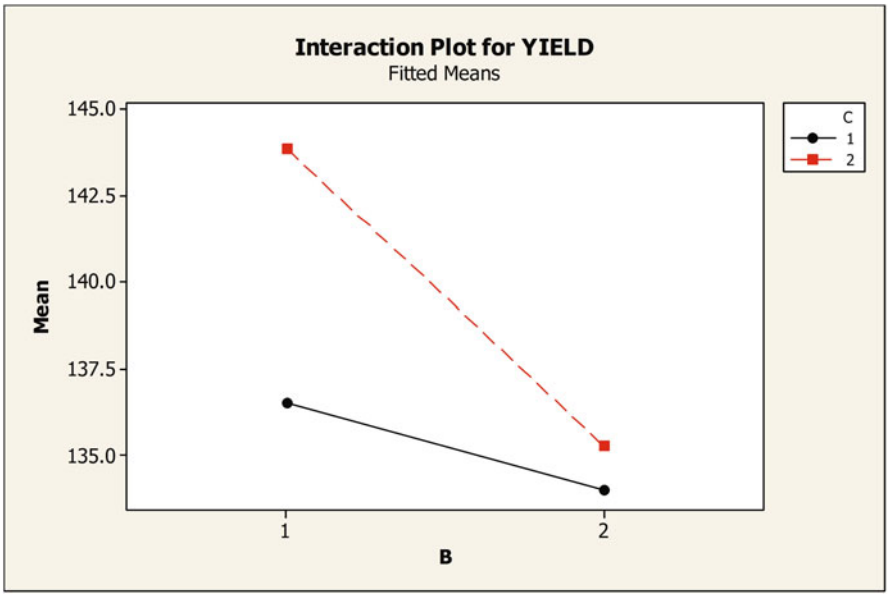


Fig. 14.6 Plot of the significant BC interaction

S.E. for comparing any two of the four treatment means = 1.7836.

Factor B has a significant larger effect at the lower level of C than at the higher level of C. Factor B significantly reduces the average yield at the higher level of C. The S.E. for comparing the responses to one factor at the two levels of the other =2.8201.

The implementation of the analysis of the above data in MINITAB is implemented as follows:

```
MTB > SET C1
DATA> (1:2)16
DATA> END
MTB > SET C2
DATA> 2(1:2)8
DATA> END
MTB > SET C3
DATA> 4(1:2)4
DATA> END
MTB > SET C4
DATA> 8(1:4)
DATA> END
MTB > SET C5
DATA> 140 138 130 142 145 146 150 147
DATA> 136 132 134 138 140 134 136 140
DATA> 142 132 128 140 146 136 140 141
DATA> 132 130 136 134 138 132 130 132
DATA> END
MTB > PRINT C1-C5

Data Display

 Row     A     B     C   REP  YIELD

   1     1     1     1     1    140
   2     1     1     1     2    138
   3     1     1     1     3    130
   4     1     1     1     4    142
   5     1     1     2     1    145
   6     1     1     2     2    146
   7     1     1     2     3    150
   8     1     1     2     4    147
   9     1     2     1     1    136
  10     1     2     1     2    132
  11     1     2     1     3    134
  12     1     2     1     4    138
  13     1     2     2     1    140
  14     1     2     2     2    134
  15     1     2     2     3    136
  16     1     2     2     4    140
  17     2     1     1     1    142
  18     2     1     1     2    132
  19     2     1     1     3    128
  20     2     1     1     4    140
  21     2     1     2     1    146
  22     2     1     2     2    136
  23     2     1     2     3    140
  24     2     1     2     4    141
  25     2     2     1     1    132
```

```
  26     2     2     1     2    130
  27     2     2     1     3    136
  28     2     2     1     4    134
  29     2     2     2     1    138
  30     2     2     2     2    132
  31     2     2     2     3    130
  32     2     2     2     4    132

MTB > ANOVA 'YIELD' = A B C A*B A*C B*C A*B*C.

ANOVA: YIELD versus A, B, C


Factor     Type Levels Values
A         fixed      2     1    2
B         fixed      2     1    2
C         fixed      2     1    2


Analysis of Variance for YIELD

Source      DF         SS         MS       F      P
A            1     108.78     108.78    6.84  0.015
B            1     247.53     247.53   15.56  0.001
C            1     148.78     148.78    9.35  0.005
A*B          1       1.53       1.53    0.10  0.759
A*C          1      22.78      22.78    1.43  0.243
B*C          1      75.03      75.03    4.72  0.040
A*B*C        1       1.53       1.53    0.10  0.759
Error       24     381.75      15.91
Total       31     987.72
```

14.2.4 Factorial in Complete Blocks

Now suppose the experiment leading to Table 14.5 had been conducted as a randomized complete block design with the replicates being blocks and each of the eight treatment combinations has been appropriately randomized within each of the four blocks (replicates are being used here as blocks). Let the data collected be tabulated as shown in Table 14.5 (note that the actual layout of the experiment is different from that shown in Table 14.5). The structure of the analysis of variance would now be:

Source	d.f.
Blocks	3
Treatments	7
Error	21
Total	31

Where the blocks SS is computed as:

$$\text{Blocks SS} = \frac{1119^2}{8} + \frac{1080^2}{8} + \frac{1084^2}{8} + \frac{1114^2}{8} - \frac{4397^2}{32}$$
$$= 151.34.$$

Again, this can be analyzed in MINITAB with the following statements and output.

```
MTB > GLM 'YIELD' = REP A B C A*B A* C B* C A*B*C;
SUBC>   Brief 2 .

General Linear Model: YIELD versus REP, A, B, C


Factor     Type Levels Values
REP       fixed      4 1 2 3 4
A         fixed      2 1 2
B         fixed      2 1 2
C         fixed      2 1 2
Analysis of Variance for YIELD, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
BLOCKS      3     151.34     151.34      50.45    4.60  0.013
A           1     108.78     108.78     108.78    9.91  0.005
B           1     247.53     247.53     247.53   22.56  0.000
C           1     148.78     148.78     148.78   13.56  0.001
A*B         1       1.53       1.53       1.53    0.14  0.712
A*C         1      22.78      22.78      22.78    2.08  0.164
B*C         1      75.03      75.03      75.03    6.84  0.016
A*B*C       1       1.53       1.53       1.53    0.14  0.712
Error      21     230.41     230.41      10.97
Total      31     987.72
```

The above analysis assumes that there are no treatments block interactions. The result shows that the interaction between B and C is significant. No other interaction term is significant at $\alpha = 0.05$ level of significance.

14.3 The 3^n Factorial Designs

Sometimes, the full effect of a treatment may be missed in a 2^n factorial design especially if treatment has a curvilinear effect. In such a situation, there would be a need to run an experiment in which the number of levels of a factor has to be at least three, so that at least, we can examine both the linear and quadratic effect of the treatment.

14.3.1 The 3^2 Factorial

The 3^2 design has two factors each at three levels. There are therefore $3^2 = 9$ treatment combinations usually written as:

A	B		
	0	1	2
0	00	01	02
1	10	11	12
2	20	21	22

The three levels are often described as (LOW, MEDIUM, HIGH) levels corresponding respectively to (0, 1, 2) levels. The treatment combinations are sometimes written as:

a_0b_0	a_1b_0	a_2b_0
a_0b_1	a_1b_1	a_2b_1
a_0b_2	a_1b_2	a_2b_2

The structure of the analysis of variance for a single replicate of the 3^2 design as well as for a replicated one is presented in the table below.

Source	d.f.
A	2
B	2
AB	4
Error	0
Total	8

Source	d.f.
Reps	$r - 1$
A	2
B	2
AB	4
Error	$8(r - 1)$
Total	$9r - 1$

In a single replicate, we see there are no degrees of freedom left for error. However, replicating this design r times either as a CRD or laid the nine treatment combinations in r blocks of size 9 with appropriate randomizations of treatment combinations within blocks will lead to the second structure of ANOVA Table. Even for $r = 2$, we would have 8 df for the error term and we would therefore be a able to estimate the unit plot variance σ^2. The interaction term AB has four degrees of freedom in a 3^2 design. These interaction effects have been described as the AB and AB^2 each having two df. If the levels of A and B are designated as x_1 and x_2, then the AB interaction component represents the response values whose x_1 and x_2 satisfy:

$$x_1 + x_2 = 0, 1, 2 \,(\text{mod } 3) \tag{14.7}$$

On the other hand, the AB^2 component represents the response values whose x_1 and x_2 satisfy:

$$x_1 + 2x_2 = 0, 1, 2 \,(\text{mod } 3) \tag{14.8}$$

Thus for (14.7), we have, the treatment combinations corresponding to:

```
=0 (mod 3):  (0,0), (1,2),  2,1)
=1 (mod 3):  (0,1), (1,0), (2,2)
=2 (mod 3):  (0,2), (2,0), (1,1)
```

Similarly for (14.8), we have the corresponding treatment combinations:

```
=0 (mod 3):  (0,0), (1,1), (2,2)
=1 (mod 3):  (0,2), (1,0), (2,1)
=2 (mod 3):  (0,1), (1,2), (2,0)
```

14.3.2 An Example of a 3^2 Design

The following dataset relate to a 3^2 experiment replicated $r = 3$ times.

Table 14.11 Synthetic data example for the 3^2 design

Treatment combinations	Rep 1	Rep 2	Rep 3
a_0b_0	10	12	14
a_0b_1	16	19	21
a_0b_2	24	27	32
a_1b_0	12	15	17
a_1b_1	19	23	29
a_1b_2	33	34	37
a_2b_0	24	27	29
a_2b_1	39	41	43
a_2b_2	45	47	52

The data is read into MINITAB. The partial output of the data is presented below together with the analysis as presented in the ANOVA Table.

```
MTB > print c1-c4

Data Display

Row  A  B  REP   Y
  1  0  0    1  10
  2  0  1    1  16
  3  0  2    1  24
  4  1  0    1  12
  5  1  1    1  19
  6  1  2    1  33
  7  2  0    1  24
  8  2  1    1  39
  9  2  2    1  45
```

```
.................
22  1  0    3  17
23  1  1    3  29
24  1  2    3  37
25  2  0    3  29
26  2  1    3  43
27  2  2    3  52

MTB > GLM 'Y' = REP A B A* B;
SUBC>   Brief 2 ;
SUBC>   GHistogram;
SUBC>   GNormalplot;
SUBC>   NoDGraphs;
SUBC>   RType 1 .

General Linear Model: Y versus REP, A, B

Factor  Type   Levels  Values
REP     fixed       3  1, 2, 3
A       fixed       3  0, 1, 2
B       fixed       3  0, 1, 2
Analysis of Variance for Y, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS  Adj MS       F      P
REP      2   150.89   150.89   75.44   57.18  0.000
A        2  1774.22  1774.22  887.11  672.34  0.000
B        2  1626.00  1626.00  813.00  616.17  0.000
A*B      4    56.44    56.44   14.11   10.69  0.000
Error   16    21.11    21.11    1.32
Total   26  3628.67


S = 1.14867   R-Sq = 99.42%   R-Sq(adj) = 99.05%
```

Clearly, both main effects A and B and their interactions are all highly significant. However, to break down the A, B, and AB into single degree of freedom components, we utilize the following coding commands based on orthogonal polynomials coefficients. Thus, we have the linear effect of factor A denoted as A_L and the quadratic effect of factor A also denoted as A_Q. Similar construction was made for the linear and quadratic components for B. For the interaction A, we have the linear-by-linear component $A_L B_L$, the linear-by-quadratic $A_L B_Q$, the quadratic–linear component $A_Q B_L$ and the quadratic–quadratic component, $A_Q B_Q$. The resulting data is presented below with the accompanying analysis using the GLM procedure in MINITAB. Notice that we declare all the components as covariates.

```
MTB > code (0) -1 (1) 0 (2) 1 c1 c5
MTB >  code (0) 1 (1) -2 (2) 1 c1 c6
MTB > code (0) -1 (1) 0 (2) 1 c2 c7
MTB > code (0) 1 (1) -2 (2) 1 c2 c8
MTB > let c9=c5*c7
MTB > let c10=c5*c8
MTB > let c11=c6*c7
MTB > let c12=c6*c8
MTB > print c1-c12
```

Data Display

Row	A	B	REP	Y	AL	AQ	BL	BQ	ALBL	ALBQ	AQBL	AQBQ
1	0	0	1	10	-1	1	-1	1	1	-1	-1	1
2	0	1	1	16	-1	1	0	-2	0	2	0	-2
3	0	2	1	24	-1	1	1	1	-1	-1	1	1
4	1	0	1	12	0	-2	-1	1	0	0	2	-2
5	1	1	1	19	0	-2	0	-2	0	0	0	4
6	1	2	1	33	0	-2	1	1	0	0	-2	-2
7	2	0	1	24	1	1	-1	1	-1	1	-1	1
8	2	1	1	39	1	1	0	-2	0	-2	0	-2
9	2	2	1	45	1	1	1	1	1	1	1	1
10	0	0	2	12	-1	1	-1	1	1	-1	-1	1
11	0	1	2	19	-1	1	0	-2	0	2	0	-2
12	0	2	2	27	-1	1	1	1	-1	-1	1	1
13	1	0	2	15	0	-2	-1	1	0	0	2	-2
14	1	1	2	23	0	-2	0	-2	0	0	0	4
15	1	2	2	34	0	-2	1	1	0	0	-2	-2
16	2	0	2	27	1	1	-1	1	-1	1	-1	1
17	2	1	2	41	1	1	0	-2	0	-2	0	-2
18	2	2	2	47	1	1	1	1	1	1	1	1
19	0	0	3	14	-1	1	-1	1	1	-1	-1	1
20	0	1	3	21	-1	1	0	-2	0	2	0	-2
21	0	2	3	32	-1	1	1	1	-1	-1	1	1
22	1	0	3	17	0	-2	-1	1	0	0	2	-2
23	1	1	3	29	0	-2	0	-2	0	0	0	4
24	1	2	3	37	0	-2	1	1	0	0	-2	-2
25	2	0	3	29	1	1	-1	1	-1	1	-1	1
26	2	1	3	43	1	1	0	-2	0	-2	0	-2
27	2	2	3	52	1	1	1	1	1	1	1	1

```
MTB > GLM 'Y' = REP AL AQ BL BQ ALBL ALBQ AQBL AQBQ;
SUBC>   Covariates 'AL' 'AQ' 'BL' 'BQ' 'ALBL' 'ALBQ' 'AQBL' 'AQBQ';
SUBC>   Brief 2 ;
SUBC>   GHistogram;
SUBC>   GNormalplot;
SUBC>   NoDGraphs;
SUBC>   RType 1 .
```

General Linear Model: Y versus REP

Factor	Type	Levels	Values
REP	fixed	3	1, 2, 3

Analysis of Variance for Y, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
REP	2	150.89	150.89	75.44	57.18	0.000
AL	1	1643.56	1643.56	1643.56	1245.64	0.000
AQ	1	130.67	130.67	130.67	99.03	0.000
BL	1	1624.50	1624.50	1624.50	1231.20	0.000
BQ	1	1.50	1.50	1.50	1.14	0.302
ALBL	1	24.08	24.08	24.08	18.25	0.001
ALBQ	1	23.36	23.36	23.36	17.71	0.001
AQBL	1	2.25	2.25	2.25	1.71	0.210
AQBQ	1	6.75	6.75	6.75	5.12	0.038
Error	16	21.11	21.11	1.32		
Total	26	3628.67				

```
S = 1.14867   R-Sq = 99.42%   R-Sq(adj) = 99.05%

Term          Coef  SE Coef        T      P
Constant   27.4444   0.2211   124.15  0.000
AL          9.5556   0.2707    35.29  0.000
AQ          1.5556   0.1563     9.95  0.000
BL          9.5000   0.2707    35.09  0.000
BQ         -0.1667   0.1563    -1.07  0.302
ALBL        1.4167   0.3316     4.27  0.001
ALBQ       -0.8056   0.1914    -4.21  0.001
AQBL       -0.2500   0.1914    -1.31  0.210
AQBQ       -0.2500   0.1105    -2.26  0.038
```

Our results indicate that $A_L, A_Q, B_L, A_LB_L, A_LB_Q$, and A_QB_Q all contributed significantly to the treatment variation in the data. Clearly, the components A_l, A_Q, B_L contributed most to this variation and they are the ones we should probably focus on in further analysis. The interaction plots for this example are presented in Fig. 14.7.

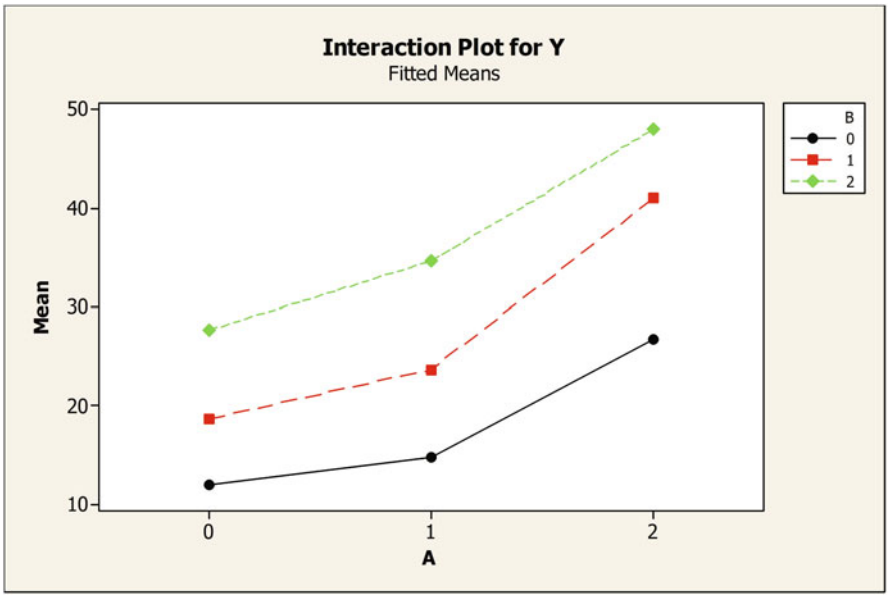


Fig. 14.7 Interaction plots in the 3^2 example

14.3.3 The 3^3 factorial Design

Here, we have three factors each at three levels resulting in a total of $3^2 = 27$ treatment combinations for a single replicate. We present these treatment combinations in the Table below with the pictorial representation of the design in Fig. 14.8.

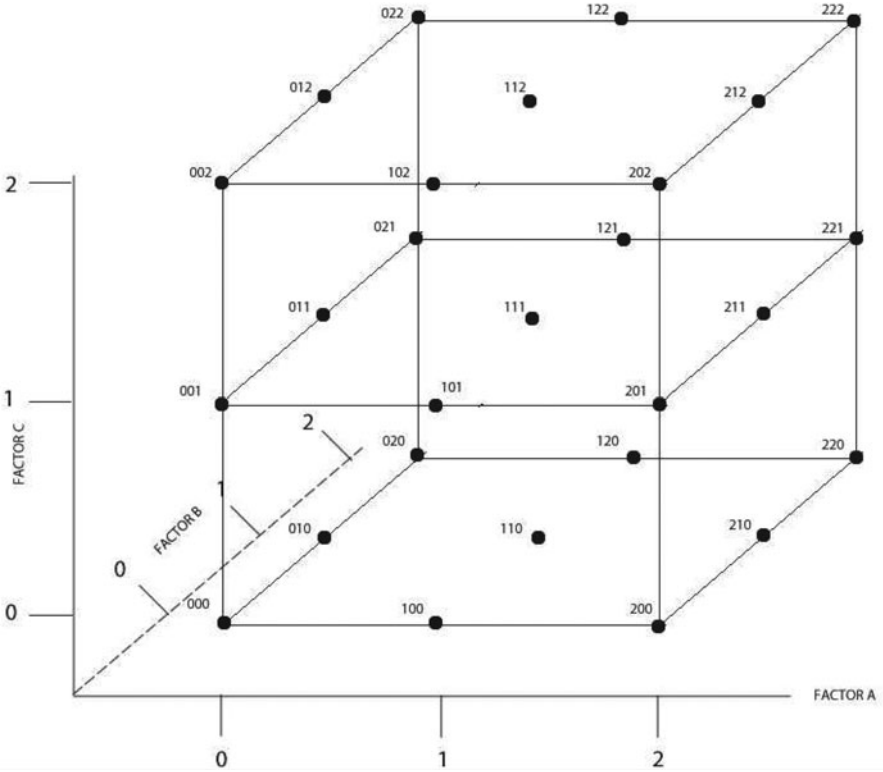


Fig. 14.8 Pictorial representation of a 3^3 Design

		Factor C		
Factor A	Factor B	0	1	2
0	0	000	001	002
0	1	010	011	012
0	2	020	021	022
1	0	100	101	102
1	1	110	111	112
1	2	120	121	122
2	0	200	201	202
2	1	210	211	212
2	2	220	221	222

The model for the 3^3 factorial design is:

$$Y(ijk) = \mu + A_{(i)} + B_{(j)} + AB_{(ij)} + C_{(k)} + AC_{(ik)} + BC_{(jk)} + ABC_{(ijk)} + \varepsilon_{ijk} \tag{14.9}$$

where $(i, j, k) = 1, 2, 3$ and the structure of the analysis of variance table for r replicates of the design is displayed below.

Source	d.f.
Reps	$r-1$
A	2
B	2
AB	4
C	2
AC	4
BC	4
ABC	8
Error	$26(r-1)$
Total	$27r-1$

14.4 Other Factorial Systems

We see from the last section that a 2^3 factorial experiment has eight treatment combinations. Similarly a 2^4 and 2^5 factorial experiments have, respectively, 16 and 32 treatment combinations and that a 3^n factorial experiment has n factors each at three levels.

We could also have mixed factorial of the form 3×4, 2×4 or $3 \times 4 \times 2$ systems. The latter being a three factor experiment each factor having three, four, and two levels, respectively, i.e., a total of 24 treatment combinations.

Example 14.3.1

The data in Table 14.12 refer to an experiment involving two factors A and B. A has four levels and B has three levels. The experiment was replicated twice ($r = 2$), and for illustrative purposes, we are assuming that the levels of the two factors are equally spaced.

Table 14.12 Coded data for this example

Replicate	Factor B	Factor A			
		a_1	a_2	a_3	a_4
1	b_1	7	8	9	7
	b_2	5	6	11	10
	b_3	4	6	10	12
2	b_1	7	9	9	8
	b_2	6	6	10	11
	b_3	6	7	10	12

Since the number of levels for factors A and B are four and three, respectively, it is possible to evaluate the linear, quadratic, and cubic effects of treatment A as well as the linear and quadratic effects of B. The joint effects (interaction) are measured by subdividing the interaction SS into $(A_L\ B_L), \cdots, (A_C\ B_Q)$. The treatment totals formed from the data in this example are presented in Table 14.13.

Table 14.13 Treatment sums formed from data in Table 14.12

	a_1	a_2	a_3	a_4	Total
b_1	14	17	18	15	64
b_2	11	12	21	21	65
b_3	10	13	20	24	67
Total	35	42	59	60	196

Replicates Total are for Rep1 and Rep2, respectively,

$$\text{Rep1} = 95; \qquad \text{Rep2} = 101$$

Analysis

Here, $r = 2, a = 4, b = 3$, therefore, we have a total of 24 observations. The relevant SSs are computed as follows:

$$\text{Total SS} = 7^2 + 5^2 + \cdots + 12^2 - \frac{196^2}{24} = 117.33$$

$$\text{Replicate SS} = \frac{95^2}{12} + \frac{101^2}{12} - \frac{196^2}{24} = 1.50$$

$$\text{SS(A)} = \frac{35^2 + 42^2 + \cdots + 60^2}{6} - \frac{196^2}{24} = 77.67$$

$$\text{SS(B)} = \frac{64^2 + 65^2 + 67^2}{8} - \frac{196^2}{24} = 0.583$$

$$\begin{aligned}\text{SS(AB)} &= \frac{14^2 + 17^2 + \cdots + 24^2}{2} - \text{CF-SS(A)-SS(B)} \\ &= 117.33 - \text{SS(A)-SS(B)} \\ &= 24.08.\end{aligned}$$

We present in Table 14.14 the initial analysis of variance table for the analysis of the data in Table 14.12.

Table 14.14 Initial analysis of variance

Source	d.f.	SS	MS	F
Replicates	1	1.5	1.5	
A	3	77.67	25.89	80.91*
B	2	0.58	0.29	0.91
AB	6	34.08	5.68	17.75*
Error	11	3.50	0.32	
Total	23	117.33		

We see that both main effect of A and the interaction terms AB are highly significant at the 5 % point. The above analysis is carried out in MINITAB as follows.

```
MTB > SET C1
DATA> (1:2)12
DATA> END
MTB > SET C2
DATA> 2(1:3)4
DATA> END
MTB > SET C3
DATA> 6(1:4)
DATA> END
MTB > SET C4
DATA> 7 8 9 7 5 6 11 10 4 6 10 12
DATA> 7 9 9 8 6 6 10 11 6 7 10 12
DATA> END
MTB > Print 'REP' 'B' 'A' 'Y'.

Data Display

 Row   REP     B     A      Y

   1     1     1     1      7
   2     1     1     2      8
   3     1     1     3      9
   4     1     1     4      7
   5     1     2     1      5
   6     1     2     2      6
   7     1     2     3     11
   8     1     2     4     10
   9     1     3     1      4
  10     1     3     2      6
  11     1     3     3     10
  12     1     3     4     12
  13     2     1     1      7
  14     2     1     2      9
  15     2     1     3      9
  16     2     1     4      8
  17     2     2     1      6
  18     2     2     2      6
  19     2     2     3     10
  20     2     2     4     11
  21     2     3     1      6
  22     2     3     2      7
  23     2     3     3     10
  24     2     3     4     12

MTB > GLM 'Y' = REP A B A*B;
SUBC>   Brief 2 .

General Linear Model: Y versus REP, A, B

Factor     Type Levels Values
REP       fixed      2 1 2
A         fixed      4 1 2 3 4
B         fixed      3 1 2 3

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
REP         1     1.5000     1.5000     1.5000    4.71  0.053
A           3    77.6667    77.6667    25.8889   81.37  0.000
B           2     0.5833     0.5833     0.2917    0.92  0.428
A*B         6    34.0833    34.0833     5.6806   17.85  0.000
Error      11     3.5000     3.5000     0.3182
Total      23   117.3333
```

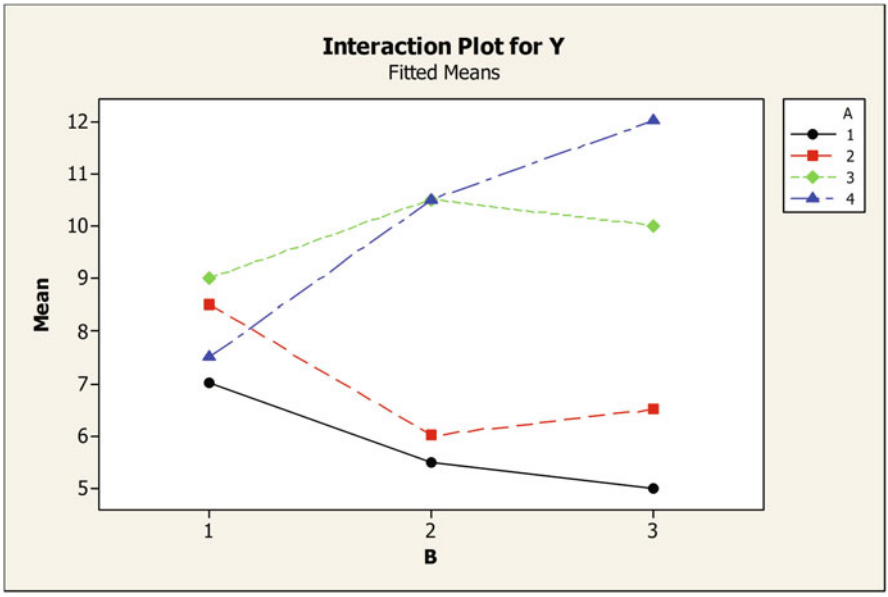


Fig. 14.9 Plot of the significant AB interaction term

We shall now partition the three sum of squares into their various components by making use of coefficients of orthogonal polynomials. For $k = 4$, there are linear, quadratic and cubic components. We give below their orthogonal coefficients (from Table 5 in the appendix)

Linear	−3	−1	1	3
Quadratic	1	−1	−1	1
Cubic	−1	3	−3	1

Similarly for $k = 3$, there are

Linear	−1	0	1
Quadratic	1	−2	1

For factor A: we have with the treatment totals,

−3	−1	1	3	L
1	−1	−1	1	Q
−1	3	−3	1	C
35	42	59	60	Totals

For factor A, we calculate below, the linear, quadratic, and cubic SS, which we have denoted here as A_L, A_Q, and A_C, respectively

$$A_L = \frac{[35(-3) + 42(-1) + 59(1) + 60(3)]^2}{6\{-3^2 + -1^2 + 1^2 + 3^2\}} = 70.53$$

$$\mathrm{A}_Q = \frac{[35(1) + 42(-1) + 59(1) + 60(1)]^2}{6\{4\}} = 1.50$$

$$\mathrm{A}_C = \frac{[35(-1) + 42(3) + 59(-3) + 60(1)]^2}{6\{20\}} = 5.63$$

We observe here that $\mathrm{A}_L SS + \mathrm{A}_Q SS + \mathrm{A}_Q SS = 70.53 + 1.50 + 5.63 = 77.66$. The three components SS are therefore pairwise orthogonal, each with 1 d.f.

For factor B: Similarly for factor B, both linear and quadratic components are again calculated from the totals for factor B levels and are again orthogonal.

−1	0	1	L
1	−2	1	Q
64	65	67	Totals

Hence,

$$\mathrm{B}_L = \frac{[64(-1) + 65(0) + 67(1)]^2}{2 \times 8} = 0.56$$

$$\mathrm{B}_Q = \frac{[64(1) + 65(-2) + 67(1)]^2}{6 \times 8} = 0.02.$$

To obtain the interaction contrasts; we first obtain the A contrasts of each level of B by using

$$L_A^1 = (-3, -1, 1, 3), \quad Q_A^1 = (1, -1, -1, 1), \quad C_A^1 = (-1, 3, -3, 1).$$

For the first level of B, we have:

$$b = 1 \ : -3(14) - 1(17) + 1(18) + 3(15) = 4$$

$$b = 2 \ : -3(11) - 1(12) + 1(21) + 3(21) = 39$$

$$b = 3 \ : -3(10) - 1(13) + 1(20) + 3(24) = 49.$$

For the second level of B, we have:

$$b = 1 \ : 1(14) - 1(17) - 1(18) + 1(15) = -6$$

$$b = 2 \ : 1(11) - 1(12) - 1(21) + 1(21) = -1$$

$$b = 3 \ : 1(10) - 1(13) - 1(20) + 1(24) = 1.$$

For the third level of B, we also have:

$$b = 1 \ : -1(14) + 3(17) - 3(18) + 1(15) = -2$$

$$b = 2 \ : -1(11) + 3(12) - 3(21) + 1(21) = -17$$

$$b = 3 \ : -1(10) + 3(13) - 3(20) + 1(24) = -7.$$

The A contrasts are given in Table 14.15.

Table 14.15 Factor A Contrasts

	Linear A	Quadratic A	Cubic A	Factor B Divisors
B_1	4	−6	−2	
B_2	39	−1	−17	
B_3	49	1	−7	
Linear B	45	7	−5	2
Quadratic B	−25	−3	25	6
A divisors	20	4	20	

Then, we multiply $L_B^1 = (-1, 0, 1)$, $Q_B^1 = (1, -2, 1)$, thus we have

$$L_A^1 \times L_B^1 = 4(-1) + 39(0) + 49(1) = 45.$$

Hence,

$$L_A^1 \times L_B^1 \quad \text{SS} = \text{A}_L\text{B}_L \text{ SS} = \frac{45^2}{2 \times 20 \times 2} = 25.31.$$

Similarly,

$$\text{A}_L\text{B}_Q = \frac{(-25)^2}{2 \times 20 \times 6} = 2.60$$

$$\text{A}_Q\text{B}_L = \frac{7^2}{2 \times 4 \times 2} = 3.60$$

$$\text{A}_Q\text{B}_Q = \frac{(-3)^2}{2 \times 4 \times 6} = 0.19$$

$$\text{A}_C\text{B}_L = \frac{(-5)^2}{2 \times 20 \times 2} = 0.31$$

$$\text{A}_C\text{B}_Q = \frac{(25)^2}{2 \times 20 \times 6} = 2.60$$

Table 14.16 gives the full analysis of variance for the data in Table 14.13.

From the F tables in the appendix, F(1,11) at $\alpha = 0.05 = 4.84$, hence, A_L, A_C, $A_L\, B_L, A_L\, B_Q, A_Q\, B_L$, and $A_C\, B_Q$ are therefore found to be significant at $\alpha = 0.05$. Obviously, the response of factor A can be least described by a third-degree polynomial. The linear × linear quadratic components too are also highly significant. Of course, we could have asked MINITAB to calculate these sum of squares in Table 14.16 (you may use the referencing Table command here) by coding the levels of A and B in MINITAB and run with the ensuing commands in MINITAB and partial output viz:

```
MTB > code (1) -1 (2) 0 (3) 1 c2 c5
MTB > code (1) 1 (2) -2 (3) 1 c2 c6
MTB > code (1) -3 (2) -1 (3) 1 (4) 3 c3 c7
MTB > code (1) 1 (2) -1 (3) -1 (4) 1 c3 c8
MTB > code (1) -1 (2) 3 (3) -3 (4) 1 c3 c9
MTB > let c10=c7*c5
MTB > let c11=c7*c6
MTB > let c12=c8*c5
MTB > let c13=c8*c6
MTB > let c14=c9*c5
MTB > let c15=c9*c6
MTB > print c1-c15

Data Display

Row  REP  B  A   Y  BL  BQ  AL  AQ  AC  ALBL  ALBQ  AQBL  AQBQ  ACBL  ACBQ
  1    1  1  1   7  -1   1  -3   1  -1     3    -3    -1     1     1    -1
  2    1  1  2   8  -1   1  -1  -1   3     1    -1     1    -1    -3     3
  3    1  1  3   9  -1   1   1  -1  -3    -1     1     1    -1     3    -3
  4    1  1  4   7  -1   1   3   1   1    -3     3    -1     1    -1     1
  5    1  2  1   5   0  -2  -3   1  -1     0     6     0    -2     0     2
  6    1  2  2   6   0  -2  -1  -1   3     0     2     0     2     0    -6
  7    1  2  3  11   0  -2   1  -1  -3     0    -2     0     2     0     6
  8    1  2  4  10   0  -2   3   1   1     0    -6     0    -2     0    -2
  9    1  3  1   4   1   1  -3   1  -1    -3    -3     1     1    -1    -1
 10    1  3  2   6   1   1  -1  -1   3    -1    -1    -1    -1     3     3
 11    1  3  3  10   1   1   1  -1  -3     1     1    -1    -1    -3    -3
 12    1  3  4  12   1   1   3   1   1     3     3     1     1     1     1
 13    2  1  1   7  -1   1  -3   1  -1     3    -3    -1     1     1    -1
 14    2  1  2   9  -1   1  -1  -1   3     1    -1     1    -1    -3     3
 15    2  1  3   9  -1   1   1  -1  -3    -1     1     1    -1     3    -3
 16    2  1  4   8  -1   1   3   1   1    -3     3    -1     1    -1     1
 17    2  2  1   6   0  -2  -3   1  -1     0     6     0    -2     0     2
 18    2  2  2   6   0  -2  -1  -1   3     0     2     0     2     0    -6
 19    2  2  3  10   0  -2   1  -1  -3     0    -2     0     2     0     6
 20    2  2  4  11   0  -2   3   1   1     0    -6     0    -2     0    -2
 21    2  3  1   6   1   1  -3   1  -1    -3    -3     1     1    -1    -1
 22    2  3  2   7   1   1  -1  -1   3    -1    -1    -1    -1     3     3
 23    2  3  3  10   1   1   1  -1  -3     1     1    -1    -1    -3    -3
 24    2  3  4  12   1   1   3   1   1     3     3     1     1     1     1

MTB > GLM 'Y' = REP AL AQ AC BL BQ ALBL ALBQ AQBL AQBQ ACBL ACBQ;
SUBC>   Covariates 'AL' 'AQ' 'AC' 'BL' 'BQ' 'ALBL' 'ALBQ' 'AQBL' 'AQBQ'  &
CONT>     'ACBL' 'ACBQ';
SUBC>   Brief 2 .

General Linear Model: Y versus REP

Factor  Type   Levels  Values
REP     fixed       2  1, 2

Analysis of Variance for Y, using Adjusted SS for Tests
```

Source	DF	Seq SS	Adj SS	Adj MS	F	P
REP	1	1.5000	1.5000	1.5000	4.71	0.053
AL	1	70.5333	70.5333	70.5333	221.68	0.000
AQ	1	1.5000	1.5000	1.5000	4.71	0.053
AC	1	5.6333	5.6333	5.6333	17.70	0.001
BL	1	0.5625	0.5625	0.5625	1.77	0.211
BQ	1	0.0208	0.0208	0.0208	0.07	0.803
ALBL	1	25.3125	25.3125	25.3125	79.55	0.000
ALBQ	1	2.6042	2.6042	2.6042	8.18	0.015
AQBL	1	3.0625	3.0625	3.0625	9.62	0.010
AQBQ	1	0.1875	0.1875	0.1875	0.59	0.459
ACBL	1	0.3125	0.3125	0.3125	0.98	0.343
ACBQ	1	2.6042	2.6042	2.6042	8.18	0.015
Error	11	3.5000	3.5000	0.3182		
Total	23	117.3333				

We obtain exactly the same results with less complications. (You may decide not to reproduce the entire print output-perhaps first 5 and last five lines!)

Example 14.3.2

An experiment was conducted on strawberries under cloches to investigate the response of four varieties to three

Table 14.16 Full analysis of variance table

Source	d.f.	SS	MS	F
Replicates	1	1.5	1.5	
Treatments				
A_L	1	70.53	70.53	240.4 *
A_Q	1	1.50	1.50	4.69
A_c	1	5.63	5.63	17.59*
B_L	1	0.56	0.56	1.75
B_Q	1	0.02	0.02	0.06
$A_L\ B_L$	1	25.31	25.31	79.09*
$A_L\ B_Q$	1	2.60	2.60	8.12
$A_Q\ B_L$	1	3.06	3.06	9.56*
$A_Q\ B_Q$	1	0.19	0.19	0.59
$A_C\ B_L$	1	0.31	0.31	0.97
$A_C\ B_Q$	1	2.60	2.60	8.12*
Error	11	3.50	0.32	
Total	23	117.33		

times of covering. A randomized block design was used, with four blocks and twelve treatment combinations. Table 14.17 gives the results from this experiment.

Table 14.17 Data for the 4 × 3 factorial experiment in this example

Time of covering	Variety	Blocks I	II	III	IV	Total
February	V	10.2	10.1	12.1	12.3	44.7
	R	11.1	9.8	8.6	9.4	38.9
	F	6.8	9.5	9.5	10.3	36.1
	G	5.3	7.5	4.6	7.3	24.7
March	V	8.0	9.7	12.0	7.8	37.5
	R	9.7	7.9	10.3	11.2	39.1
	F	8.6	9.6	9.5	10.0	37.7
	G	3.4	4.2	7.3	7.6	22.5
April	V	2.0	6.1	4.8	6.7	19.6
	R	10.9	8.4	6.5	9.2	35.0
	F	2.2	4.9	4.4	3.6	15.1
	G	2.1	0.9	3.4	2.3	8.7
Block totals		80.3	88.6	93.0	97.7	359.6

Here, there are two factors, variety and times of covering. Variety is at four levels (V, R, F, G) while time of covering has three levels (Feb, Mar, Apr). Thus, we have a total of $4 \times 3 = 12$ treatment combinations. In this experiment, therefore each block must have 12 plots and each treatment combination must be present in each block.

The initial analysis of variance (ignoring the factorial structure of treatments, that is, treating experiment as four blocks of 12 treatments each) gives,

Source	d.f.	SS	MS	F
Blocks	3	13.70	4.57	
Treatments	11	356.02	32.37	15.40 ***
Error	33	69.38	2.102	
Total	47	439.16		

The Treatments F value of 15.40 is highly significant at the 0.01 % point. We present in Table 14.18 the two-way interaction table for times and varieties.

Table 14.18 Two-way interaction table for times and varieties

Covering time	Variety V	R	F	G	Totals	Time
February	44.7	38.9	36.1	24.7		144.4
March	37.5	39.1	37.7	22.5		136.8
April	19.6	35.0	15.1	8.7		78.4
Variety totals	101.8	113.0	88.9	55.9		359.6

The relevant SS are calculated as follows:

$$\text{SS Main effect of varieties} = \frac{101.8^2}{12} + \frac{113^2}{12} + \frac{88.9^2}{12} + \frac{55.9^2}{12} - \text{CF}$$
$$= 152.69.$$

Since each of $101.8, \cdots, 55.9$ comes from 12 observations.

Table 14.19 Full analysis of variance table for the data in Table 14.17

Source	d.f.	SS	MS	F
Blocks	3	13.70	4.57	
Varieties	3	152.69	50.90	24.2 ***
Covering times	2	163.01	81.50	38.8 ***
Varieties × times	6	40.32	6.72	3.20*
Error	33	69.38	2.102	
Total	47	439.16		

Similarly,

$$\text{SS Main effect of Covering} = \frac{144.4^2}{16} + \frac{136.8^2}{16} + \frac{78.4^2}{16} - \text{CF}$$
$$= 163.01$$

$$\text{Interaction SS} = \frac{44.7^2}{4} + \frac{38.9^2}{4} + \cdots + \frac{8.7^2}{4} - \text{CF} - \text{SS(Varieties)}$$
$$- \text{SS (Times)}$$
$$= 40.32.$$

This could have been obtained as Treatment SS − Varieties SS − Times SS. The full analysis of variance is presented in Table 14.19.

As the interaction SS is significant, the results are presented in a two-way table of treatment means. The plot of these means is presented in Fig. 14.11. The s.e. of difference between two values in body of Table 14.20 is $= \sqrt{\frac{2S^2}{4}} = 1.03$. The S.E. of difference between two variety means $= \sqrt{\frac{2S^2}{16}} = 0.59$.

Summary of Results

For each variety, the difference between the means for the first two covering times was not significant; for all varieties except R, the third covering times gave a significantly lower yield than the other times. For the first two covering times, variety G gave a significantly lower yield than the other three varieties; for the third covering time, variety R gave a significantly higher yield than the other three varieties. (All significance statements refer to the 5 % significance levels).

Table 14.20 Table of treatment means

Covering time	Variety				Time mean
	V	R	F	G	
February	11.2	9.7	9.0	6.2	9.0
March	9.4	9.8	9.4	5.6	8.5
April	4.9	8.8	3.8	2.2	4.9
Variety mean	8.5	9.4	7.4	4.7	7.5

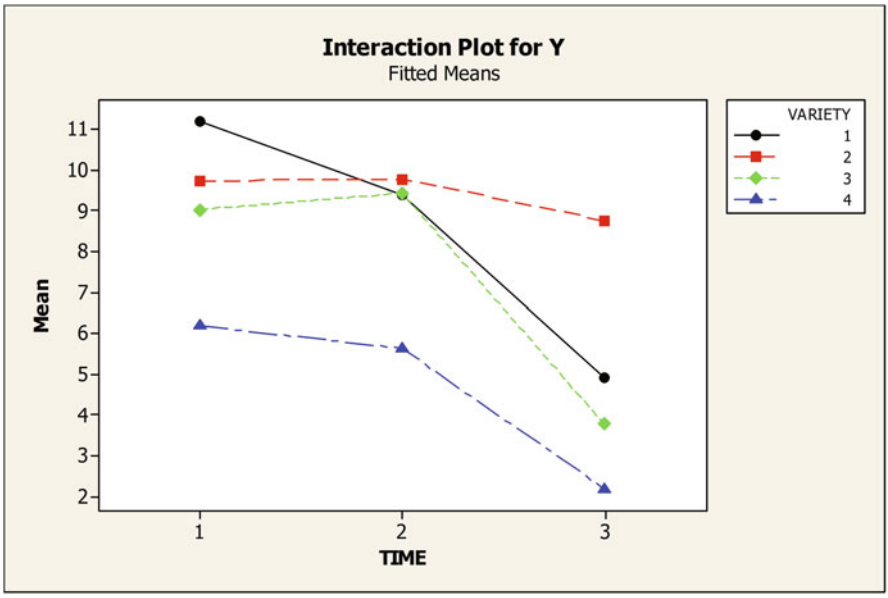


Fig. 14.10 Time and variety interaction plot

In the following MINITAB implementation, the time of covering (February, March, April) are coded (1, 2, 3), while varieties (V, R, F, G) are coded (1, 2, 3, 4), respectively.

```
MTB > SET C1
DATA> (1:3)16
DATA> END
MTB > SET C2
DATA> 3(1:4)4
DATA> END
MTB > SET C3
DATA> 12(1:4)
DATA> END
MTB > SET C4
DATA> 10.2 10.1 12.1 12.3 11.1 9.8 8.6 9.4
DATA> 6.8 9.5 9.5 10.3 5.3 7.5 4.6 7.3
DATA> 8.0 9.7 12.0 7.8 9.7 7.9 10.3 11.2
DATA> 8.6 9.6 9.5 10.0 3.4 4.2 7.3 7.6
DATA> 2.0 6.1 4.8 6.7 10.9 8.4 6.5 9.2
DATA> 2.2 4.9 4.4 3.6 2.1 0.9 3.4 2.3
DATA> END
MTB > PRINT C1-C4
```

```
Data Display

 Row   TIME  VARIETY  BLOCKS      Y

   1      1        1       1   10.2
   2      1        1       2   10.1
   3      1        1       3   12.1
   4      1        1       4   12.3
   5      1        2       1   11.1
   6      1        2       2    9.8
   7      1        2       3    8.6
   8      1        2       4    9.4
  ..............................
  ..............................
  41      3        3       1    2.2
  42      3        3       2    4.9
  43      3        3       3    4.4
  44      3        3       4    3.6
  45      3        4       1    2.1
  46      3        4       2    0.9
  47      3        4       3    3.4
  48      3        4       4    2.3

MTB > GLM 'Y' = BLOCKS TIME VARIETY TIME*VARIETY;
SUBC>   Brief 1 ;
SUBC>   Means TIME VARIETY TIME*VARIETY.

General Linear Model: Y versus BLOCKS, TIME, VARIETY


Factor     Type Levels Values
BLOCKS    fixed      4 1 2 3 4
TIME      fixed      3 1 2 3
VARIETY   fixed      4 1 2 3 4

Analysis of Variance for Y, using Adjusted SS for Tests

Source          DF     Seq SS     Adj SS     Adj MS       F      P
BLOCKS           3     13.692     13.692      4.564    2.17  0.110
TIME             2    163.007    163.007     81.503   38.73  0.000
VARIETY          3    152.685    152.685     50.895   24.18  0.000
TIME*VARIETY     6     40.320     40.320      6.720    3.19  0.014
Error           33     69.453     69.453      2.105
Total           47    439.157

Least Squares Means for Y

TIM              Mean   SE Mean
1               9.025    0.3627
2               8.550    0.3627
3               4.900    0.3627
VARIETY
1               8.483    0.4188
2               9.417    0.4188
3               7.408    0.4188
4               4.658    0.4188
TIME*VARIETY
1    1         11.175    0.7254
1    2          9.725    0.7254
```

```
1    3          9.025    0.7254
1    4          6.175    0.7254
2    1          9.375    0.7254
2    2          9.775    0.7254
2    3          9.425    0.7254
2    4          5.625    0.7254
3    1          4.900    0.7254
3    2          8.750    0.7254
3    3          3.775    0.7254
3    4          2.175    0.7254
```

Example 14.3.3

The data in Table 14.21 refer to an experiment with carrots to investigate the effect of sowing rate on yield for two stocks of seed. The experiment consisted of three randomized blocks of the eight treatment combinations. Calculate the analysis of variance, examining the effects of stock and sowing rate and the interaction between these two factors. Summarize the data in a table of means and report your conclusions.

Table 14.21 Yield in a two-factor 2×4 factorial experiment

Stocks	Sowing rate (lbs/acre)		Block I	Block II	Block III
T	1.5	(A)	4.20	4.94	4.45
	2	(B)	4.36	3.50	4.17
	2.5	(C)	5.40	4.55	5.75
	3	(D)	5.15	4.40	3.90
H		A	2.82	3.14	3.80
		B	3.74	4.43	2.92
		C	4.82	3.90	4.50
		D	4.57	5.32	4.35

Analysis

We commence the analysis of the data in Table 14.21 by first obtaining the totals for blocks, and the eight treatment combinations.

Block Totals		Treatment Total	
BK 1	35.06	T_A	13.59
BK 2	34.18	T_B	12.03
BK 3	33.8	T_C	15.70
Total	103.08	T_D	13.45
		H_A	9.76
		H_B	11.09
		H_C	13.22
		H_D	14.24
Total			103.08

The analysis of variance (Ignoring the factorial structure) is displayed as:

Source	d.f.	SS	MS	F
Blocks	2	0.0991	0.0496	
Treatments	7	8.1358	1.1623	3.43*
Error	14	4.7417	0.3386	
Total	23	12.9766		

Breaking the treatments SS into its three components, we have for SS computed as:

$$\text{SS Main effect of Stocks} = \frac{54.77^2 + 48.31^2}{12} - \text{CF} = 1.7388$$

$$\text{SS Main effect of rates} = \frac{23.35^2 + 23.12^2 + 28.92^2 + 27.69^2}{6} - \text{CF} = 4.4146$$

$$\text{Interaction SS} 8.1358 - 1.7388 - 4.4146 = 1.9824.$$

Consequently, the full analysis of variance Table is displayed in Table 14.22.

Table 14.22 The full ANOVA table for the data in Table 14.21

Source	d.f.	SS	MS	F
Blocks	2	0.0991	0.0496	
Stocks	1	1.7388	1.7388	5.14*
Rates	3	4.4146	1.4715	4.35*
Interaction	3	1.9824	0.6608	1.95
Error	14	4.7417	0.3386	
Total	23	12.9766		

The two way table of means for the data is presented in Table 14.23

Table 14.23 Two-way table of treatment means

	Rates				
Stocks	A	B	C	D	Mean
T	4.53	4.01	5.23	4.48	4.56
H	3.25	3.70	4.41	4.75	4.03
Mean	3.89	3.85	4.82	4.62	4.30

The S.E. of difference between two values in body of table $= \sqrt{\frac{2S^2}{3}} = 0.475.$

The S.E. of difference between two stock means $= \sqrt{\frac{2S^2}{12}} = 0.2380.$

The S.E. of difference between two rate means $= \sqrt{\frac{2S^2}{6}} = 0.3360.$

14.4.1 Summary of Results

The mean difference between stocks is 0.53 with a S.E. of 0.238. Significant variation is also found between the mean yields for different rates but the experiment provides insufficient evidence for interaction between the two factors. Examination of the table of treatment mean does, however, suggest that the two stocks react to changing the sowing rate in different ways—the yield for stock H increases steadily whereas the yields for stock T are irregular and the natural conclusion would be to conduct a more sensitive experiment over a wider range of rates. The analysis of variance for this example is again analyzed in MINITAB with the following. Here, the stocks are coded (T,H)=(1,2); the sowing rates are coded (1.5, 2, 2.5, 3) = (1,2,3,4), respectively. A partial output is presented.

```
MTB > SET C1
DATA> (1:2)12
DATA> END
MTB > SET C2
DATA> 2(1:4)3
DATA> END
MTB > SET C3
DATA> 8(1:3)
DATA> END
MTB > SET C4
DATA> 4.20 4.94 4.45 4.36 3.50 4.17
DATA> 5.40 4.55 5.75 5.15 4.40 3.90
DATA> 2.82 3.14 3.80 3.74 4.43 2.92
DATA> 4.82 3.90 4.50 4.57 5.32 4.35
DATA> END
MTB > PRINT C1-C4

Data Display

 Row   STOCK S   RATE S   BLOCK S       Y

   1         1        1         1    4.20
   2         1        1         2    4.94
   3         1        1         3    4.45
  ..................................
  22         2        4         1    4.57
  23         2        4         2    5.32
  24         2        4         3    4.35
```

```
MTB > GLM 'Y' = BLOCKS STOCKS RATES STOCKS*RATES;
SUBC>   Brief 1 .

General Linear Model: Y versus BLOCKS, STOCKS, RATES

Factor     Type Levels Values
BLOCKS    fixed      3 1 2 3
STOCKS    fixed      2 1 2
RATES     fixed      4 1 2 3 4

Analysis of Variance for Y, using Adjusted SS for Tests

Source           DF     Seq SS     Adj SS     Adj MS       F       P
BLOCKS            2     0.0991     0.0991     0.0496    0.15   0.865
STOCKS            1     1.7388     1.7388     1.7388    5.13   0.040
RATES             3     4.4146     4.4146     1.4715    4.34   0.023
STOCKS*RATES      3     1.9823     1.9823     0.6608    1.95   0.168
Error            14     4.7417     4.7417     0.3387
Total            23    12.9766
```

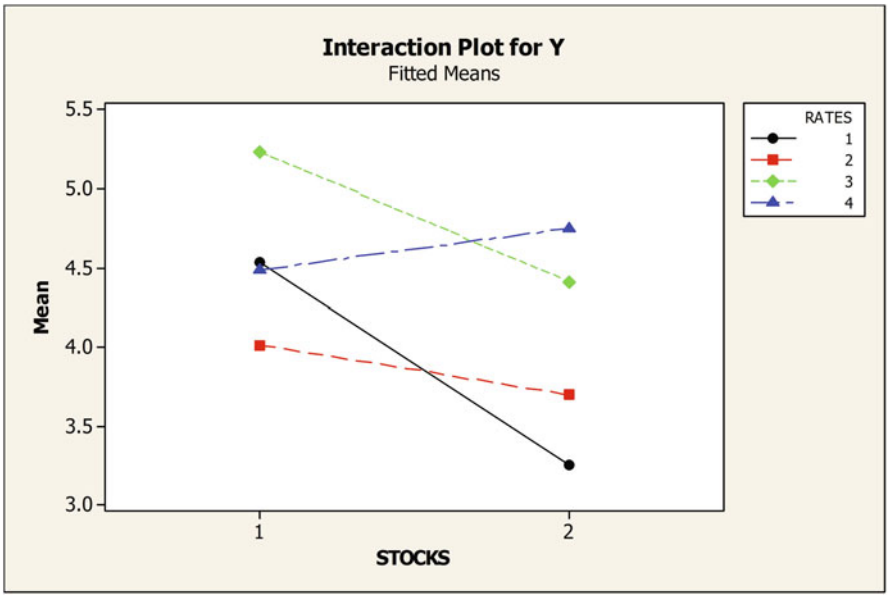


Fig. 14.11 Stock and rate interaction plot

We observe that the levels of rates were 1.5, 2, 2.5, and 3, that is, they are equally spaced and quantitative. We can therefore use this knowledge to partition the rates 3 df into three components, namely, linear, quadratic

and cubic each with 1 df by making use of the coefficients of orthogonal polynomials (Table 6 in Appendix). From the appendix we have:

Linear	−3	−1	1	3
Quadratic	1	−1	−1	1
Cubic	−1	3	−3	1
Yields	23.35	23.12	28.92	27.69

and the corresponding SS are calculated as follows:

$$\text{Linear SS} = \frac{[(-3)(23.35) + (-1)23.12 + (1)28.92 + (3)27.69]^2}{20 \times 6}$$

$$= \frac{(18.82)^2}{120} = 2.952$$

$$\text{Quadratic SS} = \frac{(23.35 - 23.12 - 28.92 + 27.69)^2}{4 \times 6}$$

$$= \frac{(-1)^2}{24} = 0.042$$

$$\text{Cubic} = \frac{[-23.35 + 3(23.12) - 3(28.92) + 1(27.69)]^2}{20 \times 6}$$

$$= \frac{(-13.06)^2}{120} = 1.421.$$

The above SS can be obtained for rates in MINITAB with the following commands and partial output.

```
MTB > %Fitline 'Y' 'RATES';
SUBC>   Poly 3.

Source      DF     Seq SS          F       P
Linear       1    2.95160    6.47734   0.018
Quadratic    1    0.04167    0.08765   0.770
Cubic        1    1.42136    3.32018   0.083
```

The F values for these components are 8.72, 0.12, and 4.19, respectively. We see that only the linear component is significant at the 5 % point (compare with F(1,14) = 4.60). A simple linear regression involving the levels of rates yield the following estimated equation with corresponding plot in Fig. 14.12.

$$\hat{Y}_i = 2.884 + 0.627\, X_i.$$

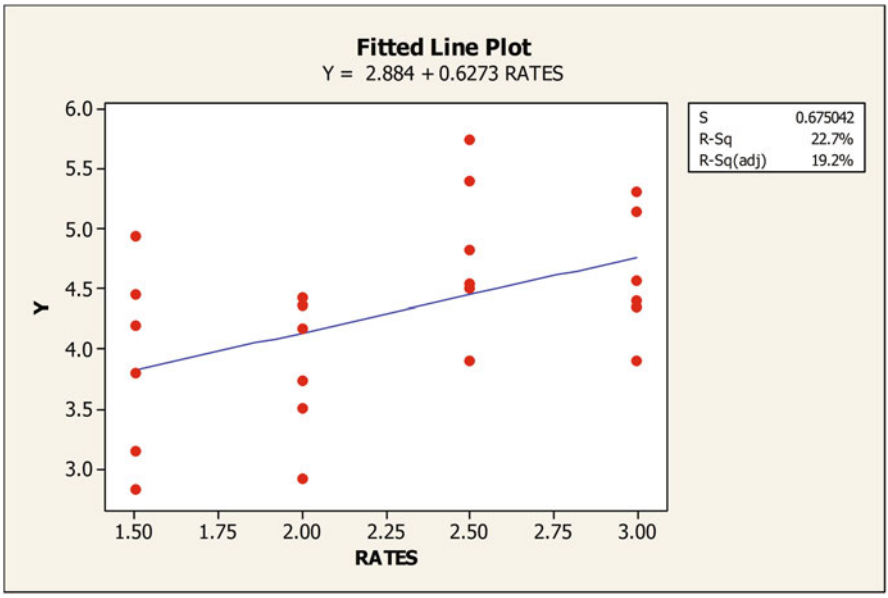


Fig. 14.12 Estimated simple regression response plot

14.5 Single Replicate Experiments

While the factorial structure and its advantages in terms of getting information on interaction effects tends to lend itself to the inclusion of more and more factors, there is the case when only a single replicate of the factorial structure is available or is desirable. The former situation usually arises when only one observation per cell is observed in a factorial experiment. The latter situation arises in large factorial experiments. Even for a 2^5 factorial, there are 32 treatment combinations, while a 2^6 has 64 treatment combinations. When resources are scarce or unavailable to sustain such a large factorial experiment, and we are unwilling to sacrifice information on some of the factors, we usually resort to single replicate factorial.

Single replicate experiments are most useful for screening experiments when several factors are under consideration. However, the single replicate is run at a cost. The error variance may be grossly estimated from the single replicate and moreover it has to be estimated from the combination of one or more interaction effects in the error line.

Consider for instance a 2^4 factorial indexed by factors A, B, C, D. Suppose only a single replicate of this design in blocks of 16 is employed. The model has the formulation:

$$y_{ijkl} = \mu + a_i + b_j + (ab)_{ij} + c_k + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk} + d_l + (ad)_{il} \tag{14.10}$$

$$+ (bd)_{jl} + (abd)_{ijl} + (cd)_{kl} + (acd)_{ikl} + (bcd)_{jkl} + (abcd)_{ijkl} + \epsilon_{ijkl}. \tag{14.11}$$

This leads to the following ANOVA structure.

Source	d.f	Source	d.f.
A	1	AD	1
B	1	BD	1
AB	1	ABD	1
C	1	CD	1
AC	1	ACD	1
BC	1	BCD	1
ABC	1	ABCD	1
D	1	Error	0
Total			15

We observe from the above that there is no degree of freedom for the error term and thus no estimate of σ^2 or the associated standard errors of treatment effects may be obtained. However, if we perceive that certain higher order interactions are insignificant, then we might be able to pull their sums of squares and corresponding sum of their degrees of freedom together to constitute the error SS and d.f., respectively. We give an example in the next example.

Example 14.4.1

In a 2^4 experiment on the yield of a chemical process, the treatment response from a single replicate of the experiment is given below.

		d_0		d_1	
		c_0	c_1	c_0	c_1
a_0	b_0	38	58	59	79
	b_1	27	30	53	53
a_1	b_0	40	55	62	75
	b_1	30	32	50	54

The treatment combinations can be extracted and arranged in the standard order as shown below.

Treatment Combinations	TC	Response
0000	(1)	38
1000	(a)	40
0100	(b)	27
1100	(ab)	30
0010	(c)	58
1010	(ac)	55
0110	(bc)	30
1110	(abc)	32
0001	(d)	59
1001	(ad)	62
0101	(bd)	53

Treatment Combinations	TC	Response
1101	(abd)	50
0011	(cd)	79
1011	(acd)	75
0111	(bcd)	53
1111	(abcd)	54

The initial MINITAB analysis is presented below.

```
MTB > SET C5
DATA> 38 40 27 30 58 55 30 32
DATA> 59 62 53 50 79 75 53 54
DATA> END

MTB > print c1-c5

Data Display
 Row     A     B     C     D      Y

   1     0     0     0     0     38
   2     1     0     0     0     40
   3     0     1     0     0     27
   4     1     1     0     0     30
   5     0     0     1     0     58
   6     1     0     1     0     55
   7     0     1     1     0     30
   8     1     1     1     0     32
   9     0     0     0     1     59
  10     1     0     0     1     62
  11     0     1     0     1     53
  12     1     1     0     1     50
  13     0     0     1     1     79
  14     1     0     1     1     75
  15     0     1     1     1     53
  16     1     1     1     1     54

MTB > GLM 'Y' = A B C D A*B A*C A*D B*C B*D C*D A*B*C A*B*D A*C*D B*C*D A &
CONT>     *B*C*D;
SUBC>   Brief 1 ;
SUBC>   Means A B C D A*C B*C C*D;
SUBC>   Residuals 'RESI1';
SUBC>   Coefficients 'COEF1';
SUBC>   Fits 'FITS1'.

General Linear Model: Y versus A, B, C, D


Factor     Type Levels Values
A         fixed      2 0 1
B         fixed      2 0 1
C         fixed      2 0 1
D         fixed      2 0 1

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F     Contributions
A           1       0.06       0.06       0.06      **         0.002
B           1    1173.06    1173.06    1173.06      **        31.624
C           1     370.56     370.56     370.56      **         9.990
D           1    1914.06    1914.06    1914.06      **        51.600
```

```
A*B          1      1.56      1.56      1.56      **         0.042
A*C          1      5.06      5.06      5.06      **         0.136
A*D          1      3.06      3.06      3.06      **         0.082
B*C          1    217.56    217.56    217.56      **         5.865
B*D          1      3.06      3.06      3.06      **         0.082
C*D          1      0.56      0.56      0.56      **         0.015
A*B*C        1     14.06     14.06     14.06      **         0.379
A*B*D        1      3.06      3.06      3.06      **         0.082
A*C*D        1      0.56      0.56      0.56      **         0.015
B*C*D        1      0.06      0.06      0.06      **         0.002
A*B*C*D      1      3.06      3.06      3.06      **         0.082
Error        0      0.00      0.00      0.00
Total       15   3709.44
```

Notice that the error df is zero and, hence, the F-values cannot be calculated. In order to know which effects are important, we can obtain the percentage of the total variation accounted for by each effect. For instance, for effect A, this equals

$$\frac{0.06}{3709.44} \times 100 = 0.002.$$

The contributions of each effect is presented in the last column. The effects B, C, D, and BC, accounted for 99.079 % of the total variation in Y. Hence, all other effects can be pooled together to form the error sum of squares on 11 df.

Usually, pooling the higher order interactions are sometimes also feasible. In this example, however, it is clear that the other effects are clearly not worth including in our final model. Thus, we refit the model in MINITAB to give the following ANOVA Table in the output.

```
MTB > Name c9 = 'RESI2' c10 = 'COEF2' c11 = 'FITS2'
MTB > GLM 'Y' = B C D B*C;
SUBC>   Brief 1 ;
SUBC>   Means B C D B*C;
SUBC>   Residuals 'RESI2';
SUBC>   Coefficients 'COEF2';
SUBC>   Fits 'FITS2'.

General Linear Model: Y versus B, C, D


Factor     Type Levels Values
B         fixed      2 0 1
C         fixed      2 0 1
D         fixed      2 0 1

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
B           1    1173.06    1173.06    1173.06  377.44  0.000
C           1     370.56     370.56     370.56  119.23  0.000
D           1    1914.06    1914.06    1914.06  615.86  0.000
B*C         1     217.56     217.56     217.56   70.00  0.000
Error      11      34.19      34.19       3.11
Total      15    3709.44
```

All the effects in the above ANOVA table are significant at the 5 % point. The fitted values and residuals from the above final model are displayed below.

Data Display

Row	A	B	C	D	Y	Fitted	Residuals
1	0	0	0	0	38	38.8125	-0.8125
2	1	0	0	0	40	38.8125	1.1875
3	0	1	0	0	27	29.0625	-2.0625
4	1	1	0	0	30	29.0625	0.9375
5	0	0	1	0	58	55.8125	2.1875
6	1	0	1	0	55	55.8125	-0.8125
7	0	1	1	0	30	31.3125	-1.3125
8	1	1	1	0	32	31.3125	0.6875
9	0	0	0	1	59	60.6875	-1.6875
10	1	0	0	1	62	60.6875	1.3125
11	0	1	0	1	53	50.9375	2.0625
12	1	1	0	1	50	50.9375	-0.9375
13	0	0	1	1	79	77.6875	1.3125
14	1	0	1	1	75	77.6875	-2.6875
15	0	1	1	1	53	53.1875	-0.1875
16	1	1	1	1	54	53.1875	0.8125

The parameter estimates of the model are also generated as:

$$49.6875, \quad 8.5625, \quad -4.8125, \quad -10.9375, \quad -3.6875$$

leading to the response model:

$$\hat{y} = 49.6875 + 8.5625x_2 - 4.8125x_3 - 10.9375x_4 - 3.6875x_2x_3 \qquad (14.12)$$

where x_2, x_3 and x_4 are variables associated with factors B, C, and D, respectively. Each of the x's takes values -1 or $+1$ depending on whether the factor with which it is associated is present or absent for the particular response being estimated in (14.12). Thus, if for the BD effect corresponding to the (ac) treatment combination, we have in this case $x_2 = +1, x_3 = -1, x_4 = +1$. Hence, on substitution $\hat{y} = 50.9375$. Note that

$$x_i = \begin{cases} +1 & \text{if letter is absent} \\ -1 & \text{if letter is present} \end{cases} \qquad (14.13)$$

14.6 Confounding in the Factorial System

A complete factorial system is the one in which all the treatment combinations are placed in one block. Of course, this arrangement can be repeated to give as many replicates as desired.

For instance, in a 2^3 experiment, all the eight treatment combinations must all be in the same block if the experiment were conducted as a randomized

complete block design. In such a case, we say that the experiment is in blocks of size 8 or simply block of 8.

Thus, for a 2^3 experiment in blocks of eight with r replications, we have the following sketch of the analysis of variance table.

Source	d.f.
Replications (Blocks)	$r-1$
Treatments	7
Error	$7(r-1)$
Total	$8r-1$

However, for large factorial, such as a 2^4 or a 2^5, to arrange these as complete factorial imply that we must have blocks of size 16 or 32, respectively. It may well be that for large factorials like these, it may not be possible to perform a complete replicate of the experiment in one block. That is, we may not have a block with as many as 16 or 32 plots to contain all these treatment combinations. The reason may be because, for example, the block might be one day, or one homogeneous batch of raw materials, etc. Confounding, therefore, is a design technique for arranging a complete factorial experiment in blocks, where the block size is smaller than the number of treatment combinations in one replicate. For example, for a 2^4 complete factorial experiment, we require for one single replicate, a block of 16. However, if for one reason or the other, a block of this size is not available, we may decide to do with two blocks of size 8, such that the first block contains a set of eight carefully chosen treatment combinations and the other block contains the remaining set of eight treatment combinations.

Example 14.6.1

In a 2^4 factorial experiment with factors A, B, C, and D, the 16 treatment combinations are,

(1), a, b, ab, c, ac, bc, abc,
d, ad, bd abd, cd, acd, bcd, abcd.

Suppose we do not have a block large enough to accommodate these treatment combinations, we could lay a single replicate of this experiment in two blocks of size 8 as presented in Table 14.24.

Table 14.24 A single replicate of a confounded design (without randomization)

(1)	ab	ac	ad	bc	bd	cd	abcd	Block 1
a	b	c	d	abc	abd	acd	bcd	Block 2

We notice from Table 14.24 that all the 16 treatment combinations are contained between the two blocks. However, half of these are in block 1 and the remaining half are in block 2. The choice of which treatment combinations go into one block or the other is a complicated process, and we will only give a brief discussion of the mechanics here. In this example, however, the above layout shows that the four order interaction ABCD has been confounded completely with blocks. That is, no information on this interaction can be obtained from this experiment in this replicate as its effect has been completely confounded with blocks.

To generate the appropriate treatment combinations that will go in each block, we need to do the following:

(a) Decide the confounding effects, in our case here, this is the ABCD interaction term.
(b) Generate treatment combinations that are even (0, 2, 4) or odd (1, 3) with the defining effects. I always like to work with the even. This will constitute the principal block.
(c) Generate the other treatment combinations in other blocks by multiplying the treatment combinations in the principal with the ones that are not already in the principal block to get all the treatment combinations, bearing in mind that any letter to the power of 2 becomes 1. For instance $a^2 = 1$ and ab^2cd becomes acd treatment combination.

In our example here, the treatment combinations that are even with the ABCD interaction are:

(1)	ab	ac	ad	bc	bd	cd	abcd	Block 1
a	b	c	d	abc	abd	acd	bcd	Block 2

We may note here that the treatment combination (1) is zero to the ABCD interaction, that is, it has no letter in common with ABCD. Block 1 above is the principal block. We notice that treatment combination (a) is not in that block. Hence, multiplying all the treatment combinations in block 1 with (a), generates the treatment combinations in the second block. Note that $a^2b = b$ etc. MINITAB can be used to generate the necessary treatment combinations for each block. We present the case for the ABCD interaction confounded with blocks below.

```
MTB > Name C1 "StdOrder" C2 "RunOrder" C3 "CenterPt" C4 "Blocks" C5 "A" C6 "B" C7 "C" &
CONT>  C8 "D"
MTB > FFDesign 4 16;
SUBC>   CTPT 'CenterPt';
SUBC>   Blocks 2;
SUBC>   SOrder 'StdOrder' 'RunOrder';
SUBC>   Alias 4;
SUBC>   XMatrix 'Blocks' 'A' 'B' 'C' 'D'.

Full Factorial Design

Factors:   4   Base Design:         4, 16   Resolution with blocks:  V
Runs:     16   Replicates:               1
Blocks:    2   Center pts (total):      0


Block Generators: ABCD


Alias Structure

I

Blk = ABCD

A
B
C
D
AB
AC
AD
BC
BD
CD
ABC
ABD
ACD
BCD

MTB > print c1-c8

Data Display

Row  StdOrder  RunOrder  CenterPt  Blocks   A   B   C   D
  1         1         1         1       1   1  -1  -1  -1
  2         2         2         1       1  -1   1  -1  -1
  3         3         3         1       1  -1  -1   1  -1
  4         4         4         1       1   1   1   1  -1
  5         5         5         1       1  -1  -1  -1   1
  6         6         6         1       1   1   1  -1   1
  7         7         7         1       1   1  -1   1   1
  8         8         8         1       1  -1   1   1   1
  9         9         9         1       2  -1  -1  -1  -1
 10        10        10         1       2   1   1  -1  -1
 11        11        11         1       2   1  -1   1  -1
 12        12        12         1       2  -1   1   1  -1
 13        13        13         1       2   1  -1  -1   1
 14        14        14         1       2  -1   1  -1   1
 15        15        15         1       2  -1  -1   1   1
 16        16        16         1       2   1   1   1   1
```

The output indicates that only 14 effects can be estimated. The main effects, say A main effect is obtained as:

$$\frac{(a-1)(b+1)(c+1)(d+1)}{2^4}$$
$$= \frac{abcd+abc+abd+ab+acd+ac+ad+a-bcd-bc-bd-b-cd-c-d-(1)}{16}$$

Notice that half of the positive treatment combinations are in block 1, and half of the negative treatment combinations are also in block 1. Similarly, half of the positive treatment combinations are in block 2 and half of the negatives are also in block 2. Thus, the effect of A can be estimated from both blocks. In comparison, the effect of the ABCD is such that all the positives are in block 1 and all the negatives are in block 2; thus, the effect of ABCD is no more than the difference between blocks 1 and 2 and thus the ABCD is said to be intrinsically confounded with blocks. The effects of the other effects can similarly be generated and noting that the interaction AB is simply the product of the entries in column "A" and "B" etc.

14.6.1 Replications in 2^n Confounding

If the above basic design in Table 14.24 is repeated, say three times, we could have the following layout in Table 14.25.

We notice that the treatment combinations have been randomized within blocks, otherwise the eight treatment combinations in Replicate I in Table 14.24 are still the same. The same for treatment combinations in Replicates II and III too. An analysis layout for the above design is given below:

Table 14.25 A 2^4 factorial in blocks of size 8 in three replicates

Rep. I		Rep. II		Rep. III	
Block 1	Block 2	Block 1	Block 2	Block 1	Block 2
(1)	a	b	(1)	(1)	acd
ab	b	bcd	ad	bd	a
ac	c	abc	bd	abcd	abd
ad	d	a	abcd	cd	d
bc	abc	acd	cd	bc	bcd
bd	abd	abd	bc	ad	abc
cd	acd	d	ac	ac	b
abcd	bcd	c	ab	ab	c

Source	d.f.		d.f.
Replications	2		
Blocks (ABCD)	1		
Replications & block interaction	2	Blocks within Reps	5
A	1		
B	1		
AB	1		
⋮	⋮		
D	1		
BCD	1		
Replications & all others (Error)	28		
Total	47		

Here, complete information is lost on the four-order interaction ABCD as this has been completely confounded with blocks.

In general, therefore, the technique of confounding causes information about certain treatment effects (usually high-order interactions) to be indistinguishable from or confounded with blocks; and we say that this type of design is an *incomplete block design* because each block does not contain all treatments or treatment combinations. We also note that a basic assumption of the analysis of variance is that the plots be homogeneous, that is, the variance of each of the plots is a constant and is equal to σ^2. However, having blocks of size 16, 32, or even 64 may make this assumption invalid. It is therefore imperative that for factorials of these type, we seldom allow a block to be more than of size 16. However, to conduct say a 2^5 experiment in blocks of 16 requires that we would have to employ the technique of confounding. Thus, for homogeneity purposes, confounding readily lends itself for consideration in the designs of experiments.

For 2^n factorials, the complete factorial would require a block of size 2^n. With confounding, however, blocks of sizes 2^{n-1}, 2^{n-2}, 2^{n-3}, that is, multiples of 2 are possible. However, for experiments in blocks of 2^{n-1}, we must sacrifice at least one effect or interaction. For experiments in blocks of 2^{n-2}, we must sacrifice two effects or interactions plus their generalized interaction—a total of three effects or interactions.

The choice of these effects is very technical and statisticians need to be consulted for this and similar designs. For example, for a 2^4 experiment, with factors A, B, C, and D, there are a total of 16 treatment combinations and a total of 15 d.f. for the effects and interactions. These effects and interactions are

A, B, AB, C, AC, BC, ABC, D
AD, BD, ABD, CD, ACD, BCD, ABCD.

Each of these has 1 df. Thus to conduct this experiment in blocks of eight, we need to confound any one of the above 15 effects or interactions, while to have the experiment in blocks of four will require the selection of two basic effects or interactions. Of course, their generalized interaction will also be confounded, e.g.,

(i) Suppose ABCD and BC are confounded, then their generalized interaction, ABCD × BC = $\mathrm{AB^2C^2D}$ = AD is also automatically confounded with blocks. This implies that complete information is lost on these three effects, including two 2-factor interaction terms. We surely do not want to lose complete information on two 2-factor effects.

(ii) Suppose we try as defining effects, ABCD and ABD. Again this implies that ABCD × ABD = $A^2B^2CD^2$ = C is also automatically confounded with blocks. Of course, we do not want to confound a main effect with block, thus the choice of our defining effects in this case is not wise.

(iii) Now suppose we choose to confound two 3-factor interactions, ABD and BCD. This also implies that their generalized interaction ABD × BCD = $AB^2\ CD^2$ = AC is also confounded with blocks. This would be a better choice, as it has fewer lower order interactions confounded.

The above indicates that care must, therefore, be taken so that main effects and possibly second order interactions are not confounded. We generally assume that higher order interactions may not be as important as lower order interactions. To generate a single replicate of the third choice, we generate the even treatment combinations with the defining effects and pull out the four common treatments combinations to both effects.

ABD	BCD
(1)	(1)
ab	bc
ad	bd
bd	cd
c	a
abc	abc
acd	abd
bcd	acd

Notice that treatment combination c and a are zero letters with ABD and BCD, respectively. Introducing them and multiplying all previously identified treatment combinations with either of them will generate the remaining combinations. From the above, we observe that only four treatment combinations are common to both. These are laid out in the principal block (Block 1) in Table 14.26.

Table 14.26 Single replicate of a 2^4 in blocks of 4

Block 1	Block 2	Block 3	Block 4
(1)	a	b	ab
bd	abd	d	ad
abc	bc	ac	c
acd	cd	abcd	bcd

The treatment combinations in Block 2 are generated by noting that (a) is not in block 1 and multiplying the treatment combinations in block 1 with a. Block 3 is generated by introducing the combination (b) and again multiplying through with (b). Block 4 is obtained by multiplying those in block 1 with the (ab) treatment combination. All the 16 treatment combinations in a 2^4 factorial are present in the four blocks, however, information on the ABD, BCD and AC effects are completely lost with the blocks, and therefore are not estimable. There is of course a need to randomize the treatment combinations within blocks in the design in Table 14.26.

Example 14.6.2

Three factors A (temperature), B (pressure) and C (catalyst concentration) are believed to influence the yield of a chemical reaction. These factors were all set at two levels and because of the limitations in the laboratory, the 2^3 factorial design was run in two blocks of size 4, with ABC as the confounded effect. Three replicates of the experiment were conducted were carried with treatment combinations randomized within blocks and blocks randomized within replicates. The data collected are presented in Table 14.27.

The analysis of the data was carried out in MINITAB as follows:

Table 14.27 An example of a 2^3 factorial in blocks of 4

Replicate I				Replicate II				Replicate III			
Block 1a		Block 1b		Block 2a		Block 2b		Block 3a		Block 3b	
(1)	4.6	(a)	10.1	(1)	4.4	(a)	7.8	(1)	4.8	(a)	9.2
(ab)	6.6	(b)	4.6	(ab)	6.3	(b)	5.9	(ab)	6.3	(b)	6.2
(ac)	6.2	(c)	7.6	(ac)	7.0	(c)	6.9	(ac)	7.5	(c)	8.0
(bc)	8.1	(abc)	7.7	(bc)	8.9	(abc)	8.2	(bc)	9.0	(abc)	8.2

```
Data Display

 Row     A     B     C       Y  BLOCKS   REP
   1     0     0     0     4.6       1     1
   2     1     1     0     6.6       1     1
   3     1     0     1     6.2       1     1
   4     0     1     1     8.1       1     1
   5     1     0     0    10.1       2     1
   6     0     1     0     4.6       2     1
   7     0     0     1     7.6       2     1
   8     1     1     1     7.7       2     1
   9     0     0     0     4.4       1     2
  10     1     1     0     6.3       1     2
  11     1     0     1     7.0       1     2
  12     0     1     1     8.9       1     2
  13     1     0     0     7.8       2     2
  14     0     1     0     5.9       2     2
  15     0     0     1     6.9       2     2
  16     1     1     1     8.2       2     2
  17     0     0     0     4.8       1     3
  18     1     1     0     6.3       1     3
  19     1     0     1     7.5       1     3
  20     0     1     1     9.0       1     3
  21     1     0     0     9.2       2     3
  22     0     1     0     6.2       2     3
  23     0     0     1     8.0       2     3
  24     1     1     1     8.2       2     3
```

```
MTB > GLM 'Y' = REP BLOCKS(REP) A B C A*B A*C B*C;
SUBC>   Brief 1 ;
SUBC>   Means A B C.

General Linear Model: Y versus REP, A, B, C, BLOCKS

Factor         Type Levels Values
REP           fixed      3 1 2 3
BLOCKS(REP) random      6 1 2 1 2 1 2
A             fixed      2 0 1
B             fixed      2 0 1
C             fixed      2 0 1

Analysis of Variance for Y, using Adjusted SS for Tests

Source         DF     Seq SS     Adj SS     Adj MS       F      P
REP             2     1.1725     1.1725     0.5863    0.34  0.735
BLOCKS(REP)     3     5.1362     5.1363     1.7121    4.22  0.030
A               1     6.1004     6.1004     6.1004   15.04  0.002
B               1     0.1504     0.1504     0.1504    0.37  0.554
C               1    11.3437    11.3438    11.3438   27.96  0.000
A*B             1     4.9504     4.9504     4.9504   12.20  0.004
A*C             1    15.8438    15.8438    15.8438   39.05  0.000
B*C             1     5.9004     5.9004     5.9004   14.54  0.002
Error          12     4.8683     4.8683     0.4057
Total          23    55.4663
```

The S.E.s

(i) The S.E. of a single yield $= \sqrt{0.4057} = 0.6369$.

(ii) The S.E. of difference of any two treatment mean is given by $\sqrt{\frac{2S^2}{r2^2}} = \sqrt{\frac{2\times 0.4057}{12}} = 0.2600$.

(iii) The S.E. for any interaction mean equals $\sqrt{\frac{2S^2}{r2^1}} = \sqrt{\frac{2\times 0.4057}{6}} = 0.3677$.

(iv) Estimated variance of an unconfounded total effects $= 24 \times 0.4057 = 9.7368$ and hence, has S.E. $= 3.1204$. We note that $r = 3$ in this example.

We observe that both the main effects, A and C are highly significant. However, the interaction terms AB, AC, and BC are also highly significant. Therefore, our focus should be directed to the interaction terms in order to explain the variations within the experimental factors.

In the completely confounded analysis above, there is total loss of information on the ABC interaction, since it has been confounded with blocks. We may, however, be able to recover some information on the confounded effect on an analysis that is based on blocks as a unit. This is known as *inter-block information* recovery as distinct from our analysis which is also often called *intra-block information* of variation of plots within blocks.

The inter-block analysis of the above data is carried out below with the accompanying output from MINITAB. Here, the BLOCKS within Replicate BLOCKS(REP) sum of squares of 5.1362 is broken down into the two components, namely, Blocks (ABC) and Blocks*Replicate interaction. The p-value for blocks is very significant at the 5 % point; hence, blocking has been very

effective in reducing our experimental variance. The interaction plots for this example are presented in Fig. 14.13. Note that information on the ABC interaction has been lost or confounded with blocks in the above example. We consider in the next section experiments in which information in confounded effects are not totally lost.

```
MTB > GLM 'Y' = REP BLOCKS BLOCKS*REP A B C A*B A*C B*C;
SUBC>   Brief 1 ;
SUBC>   Means A B C.


General Linear Model: Y versus REP, BLOCKS, A, B, C


Analysis of Variance for Y, using Adjusted SS for Tests

Source        DF     Seq SS     Adj SS     Adj MS       F       P
REP            2     1.1725     1.1725     0.5863    3.21   0.238
BLOCKS         1     4.7704     4.7704     4.7704   26.08   0.036
REP*BLOCKS     2     0.3658     0.3658     0.1829    0.45   0.647
A              1     6.1004     6.1004     6.1004   15.04   0.002
B              1     0.1504     0.1504     0.1504    0.37   0.554
C              1    11.3438    11.3438    11.3438   27.96   0.000
A*B            1     4.9504     4.9504     4.9504   12.20   0.004
A*C            1    15.8437    15.8437    15.8437   39.05   0.000
B*C            1     5.9004     5.9004     5.9004   14.54   0.002
Error         12     4.8683     4.8683     0.4057
Total         23    55.4663
```

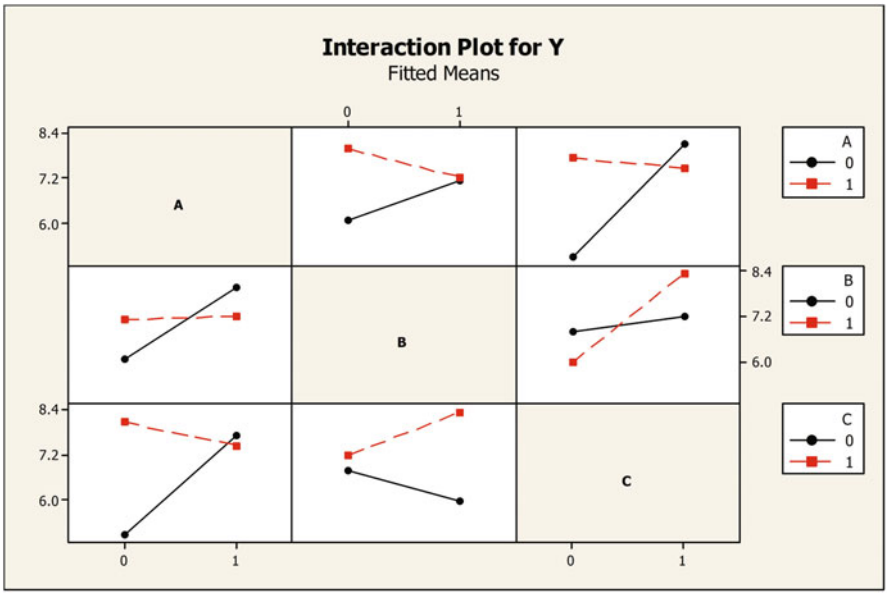


Fig. 14.13 Interaction plots for the significant two-way interactions

We present below the interaction means for AB, AC, and the BC interactions.

A	B		Mean
	0	1	
0	6.050	7.117	6.583
1	7.967	7.217	7.592
Mean	7.008	7.167	

A	C		Mean
	0	1	
0	5.083	8.083	6.583
1	7.717	7.467	7.592
Mean	6.400	7.775	

B	C		Mean
	0	1	
0	6.817	7.200	7.008
1	5.983	8.350	7.167
Mean	6.400	7.775	

Comparisons of the AB interaction means using Tukey's procedure yield the following results. Clearly, the significant difference is between the lower level yield for the two factors against at least one of the upper level combinations of the factors.

```
Grouping Information Using Tukey's Method and 95.0% Confidence

A  B  N   Mean  Grouping
1  0  6  7.967  A
1  1  6  7.217  A
0  1  6  7.117  A B
0  0  6  6.050    B

Means that do not share a letter are significantly different.

Grouping Information Using Tukey Method and 95.0% Confidence

A  C  N   Mean  Grouping
0  1  6  8.083  A
1  0  6  7.717  A
1  1  6  7.467  A
0  0  6  5.083    B

Means that do not share a letter are significantly different.

Grouping Information Using Tukey Method and 95.0% Confidence

B  C  N   Mean  Grouping
1  1  6  8.350  A
0  1  6  7.200    B
0  0  6  6.817    B  C
1  0  6  5.983       C

Means that do not share a letter are significantly different.
```

14.7 Partial Confounding

In complete confounding, information is totally lost on the confounded effects as they are mixed up with the block effects. However, to avoid the loss of total information on any of the confounded factorial effects, a different effect can be confounded in different replication group. Thus, a factorial effect is only confounded in one of the replications and its effect can thus be estimated from the other replicates. We illustrate this with the following example.

Example 14.7.1

An animal scientist conducted a study on the effects of heat stress and dietary intake of protein and saline water on laboratory mice. The three factors were each used at two levels in a 2^3 factorial structure. The levels of the factors were (A) protein (low, high); (B) water (normal, saline); and (C) heat stress (room temperature, heat stress). Blocks of size 4 were used with four mice from an individual litter used in each block. Each mouse was put in an individual cage and assigned one of the treatment combinations. One replication of the experiment consisted of two litters of mice. The weights gains (grams) for the mice are shown for each mouse next to the treatment combination.

Litter 1		Litter 2		Litter 3		Litter 4	
(1)	27.5	(ab)	24.3	(bc)	19.5	(abc)	19.7
(bc)	20.6	(c)	24.3	(a)	24.1	(b)	19.5
(abc)	22.0	(ac)	22.8	(ab)	22.4	(1)	22.5
(a)	28.6	(b)	24.6	(c)	22.0	(ac)	18.8

Replicate I Replicate II

Litter 5		Litter 6	
(1)	24.5	(a)	33.1
(c)	23.0	(b)	20.5
(ab)	23.4	(ac)	19.8
(abc)	21.7	(bc)	18.5

Replicate III

In replicate I, the BC effect has been confounded with litters 1 and 2, the litters serving as blocks in this case. Similarly, in replicate II, the AC effect has been confounded with litters 3 and 4. In replicate III, the AB effect has similarly been confounded with litters 5 and 6. Thus, the AB effect will be estimated from replicates I and II, the AC from replicates I and III, while the BC would be from replicates II and III. All other effects will be estimated from the three replicates.

Table 14.28 ANOVA for partially confounded 2^3 factorial

Source	Degrees of freedom (d.f.)
Replicates	2
Blocks within replicates[or AB(rep. III) + BC (rep. I) + AC (rep. II)	3
A	1
B	1
C	1
ABC	1
AB (from replicates I, II)	1
AC (from replicates I, III)	1
BC (from replicates II, III)	1
Error	11
Total	23

The analysis of the data in this example is again carried out in MINITAB with data read into columns C1, C2, C3, C4, C5, C6 as indicated below.

```
Data Display

 Row     A     B     C     Y      LITTER  REP
 ----------------------------------------------
   1     0     0     0   27.5       1     1
   2     0     1     1   20.6       1     1
   3     1     1     1   22.0       1     1
   4     1     0     0   28.6       1     1
   5     1     1     0   24.3       2     1
   6     0     0     1   24.3       2     1
   7     1     0     1   22.8       2     1
   8     0     1     0   24.6       2     1
   9     0     1     1   19.5       1     2
  10     1     0     0   24.1       1     2
  11     1     1     0   22.4       1     2
  12     0     0     1   22.0       1     2
  13     1     1     1   19.7       2     2
  14     0     1     0   19.5       2     2
  15     0     0     0   22.5       2     2
  16     1     0     1   18.8       2     2
  17     0     0     0   24.5       1     3
  18     0     0     1   23.0       1     3
  19     1     1     0   23.4       1     3
  20     1     1     1   21.7       1     3
  21     1     0     0   33.1       2     3
  22     0     1     0   20.5       2     3
  23     1     0     1   19.8       2     3
  24     0     1     1   18.5       2     3

MTB > GLM 'Y' = REP LITTER(REP) A B C A*B A*C B*C A*B*C;
SUBC>   Brief 2 .

General Linear Model: Y versus REP, A, B, C, LITTER
```

```
Factor        Type Levels Values
REP          fixed      3 1 2 3
LITTER(REP)  fixed      6 1 2 1 2 1 2
A            fixed      2 0 1
B            fixed      2 0 1
C            fixed      2 0 1

Analysis of Variance for Y, using Adjusted SS for Tests

Source        DF     Seq SS     Adj SS     Adj MS       F      P
REP            2     43.603     43.603     21.802    5.94  0.018
LITTER(REP)    3      8.004      2.234      0.745    0.20  0.892
A              1      7.820      7.820      7.820    2.13  0.172
B              1     49.020     49.020     49.020   13.35  0.004
C              1     74.554     74.554     74.554   20.30  0.001
A*B            1      2.402      2.402      2.402    0.65  0.436
A*C            1      9.610      9.610      9.610    2.62  0.134
B*C            1     12.602     12.602     12.602    3.43  0.091
A*B*C          1     14.260     14.260     14.260    3.88  0.074
Error         11     40.402     40.402      3.673
Total         23    262.280
```

$F(.05, 1, 11) = 4.84$. Only the main effects B and C are significant. None of the other effects or interactions have significant effect on weight gained by mice. There is a significant mean reduction in weight gain of 2.858 (s.e $= \sqrt{\frac{2S^2}{12}} = \sqrt{\frac{2(3.673)}{12}} = 0.7824$) if the mice are fed with saline water rather than normal water. Similarly, there is a significant mean reduction in weight gains of mice of 3.525 g for mice subjected to heat stress relative to those kept at room temperature with a corresponding S.E. of 0.7824.

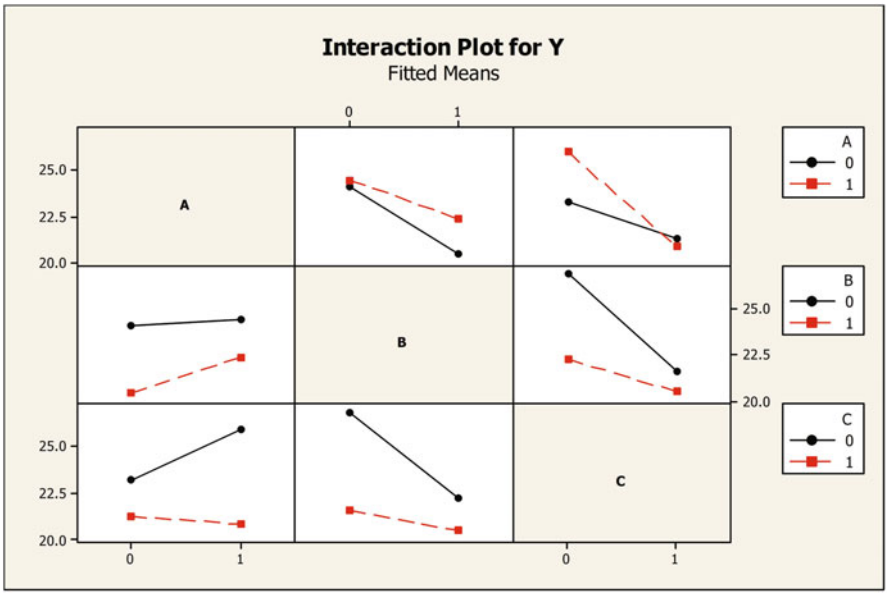


Fig. 14.14 Interaction plots for the partial confounding example

Example 14.7.2

Consider the following 2^3 factorial in blocks of two. That means that two effects and their generalized interaction will be confounded. Suppose we design such an experiment with each replicate having three confounded effects with a total of four replicates. That is 32 total observations. The design could be like this:

Replicate I: Confound BC and AC $\longrightarrow$ AB
Replicate II: Confound BC and ABC $\longrightarrow$ A
Replicate III: Confound AC and ABC $\longrightarrow$ B
Replicate IV: Confound AB and ABC $\longrightarrow$ C

We present the data for this design in the following Table 14.29.

Table 14.29 Synthetic data for the design

Replicate I: AB, AC, BC confounded			
Block 1	Block 2	Block 3	Block 4
75 (1)	89 (ab)	61 (a)	30 (b)
100 (abc)	73 (c)	45 (bc)	54 (ac)
Replicate II: A, BC, ABC confounded			
Block 1	Block 2	Block 3	Block 4
60 (1)	47 (a)	1 (b)	26 (ac)
34 (bc)	81 (abc)	35 (c)	52 (ab)
Replicate III: B, AC, ABC confounded			
Block 1	Block 2	Block 3	Block 4
58 (1)	48 (a)	18 (b)	68 (ab)
42 (ac)	52 (c)	82 (abc)	32 (bc)
Replicate IV: C, AB, ABC confounded			
Block 1	Block 2	Block 3	Block 4
47 (1)	34 (a)	50 (c)	37 (ac)
57 (ab)	4 (b)	80 (abc)	27 (bc)

We may notice here that the effects of main effects A, B, and C will be estimated from three replicates each, that is,

A, B, C, are estimable from three replicates each:

1. A $\longrightarrow$ Reps: I, III, IV
2. B $\longrightarrow$ Reps: I, II, IV
3. C $\longrightarrow$ Reps: I, II, III

AB, AC, BC are estimable from two replicates each:

1. AB $\longrightarrow$ Reps: II, III
2. AC $\longrightarrow$ Reps: II, IV
3. BC $\longrightarrow$ Reps: III, IV

ABC is estimable from only one replicate, namely, Rep I.

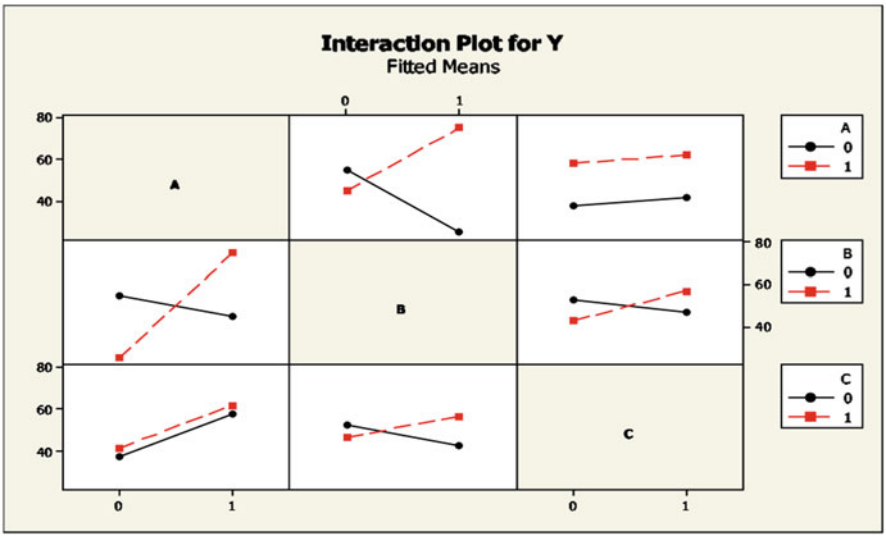
We present below the output from MINITAB analysis of the data in Table 14.29. The data were read in columns C1 to C6.

```
Data Display

Row  REP   BLK   A   B   C    Y
  1    1     1   0   0   0   75
  2    1     2   1   1   0   89
  3    1     3   1   0   0   61
  4    1     4   0   1   0   30
  5    1     1   1   1   1  100
  6    1     2   0   0   1   73
  7    1     3   0   1   1   45
  8    1     4   1   0   1   54
  9    2     1   0   0   0   60
 10    2     2   1   0   0   47
 11    2     3   0   1   0    1
 12    2     4   1   0   1   26
 13    2     1   0   1   1   34
 14    2     2   1   1   1   81
 15    2     3   0   0   1   35
 16    2     4   1   1   0   52
 17    3     1   0   0   0   58
 18    3     2   1   0   0   48
 19    3     3   0   1   0   18
 20    3     4   1   1   0   68
 21    3     1   1   0   1   42
 22    3     2   0   0   1   52
 23    3     3   1   1   1   82
 24    3     4   0   1   1   32
 25    4     1   0   0   0   47
 26    4     2   1   0   0   34
 27    4     3   0   0   1   50
 28    4     4   1   0   1   37
 29    4     1   1   1   0   57
 30    4     2   0   1   0    4
 31    4     3   1   1   1   80
 32    4     4   0   1   1   27

Interaction Plot for Y

MTB > Erase C4000.
MTB > GLM 'Y' = REP BLK( REP) A B A*B C A*C B*C A*B*C;
SUBC>   Brief 2 ;
SUBC>   Means A B 'C'A*B A*C B*C;
SUBC>   Pairwise A B 'C' A*B A*C B*C;
SUBC>     Tukey;
SUBC>     NoTest;
SUBC>     NoCI.

General Linear Model: Y versus REP, A, B, C, BLK

Factor    Type   Levels  Values
REP       fixed       4  1, 2, 3, 4
BLK(REP)  fixed      16  1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4
A         fixed       2  0, 1
B         fixed       2  0, 1
C         fixed       2  0, 1

Analysis of Variance for Y, using Adjusted SS for Tests

Source    DF   Seq SS   Adj SS   Adj MS          F      P
REP        3  3040.09  3040.09  1013.36   36481.12  0.000
BLK(REP)  12  7568.38   653.91    54.49    1961.72  0.000
A          1  2420.04  2420.04  2420.04   87121.50  0.000
B          1     0.04     0.04     0.04       1.50  0.252
A*B        1  3600.00  3600.00  3600.00  129600.00  0.000
C          1   100.04   100.04   100.04    3601.50  0.000
A*C        1     0.00     0.00     0.00       0.00  1.000
```

```
B*C        1    400.00   400.00   400.00   14400.00  0.000
A*B*C      1      0.12     0.12     0.12       4.50  0.063
Error      9      0.25     0.25     0.03
Total     31  17128.97

S = 0.166667   R-Sq = 100.00%   R-Sq(adj) =  99.99%

Unusual Observations for Y

Obs        Y     Fit  SE Fit  Residual  St Resid
  1   75.000  75.250   0.144    -0.250     -3.00 R
  5  100.000  99.750   0.144     0.250      3.00 R

R denotes an observation with a large standardized residual.

Least Squares Means for Y

A     Mean  SE Mean
0    39.93  0.04501
1    60.01  0.04501
B
0    49.93  0.04501
1    50.01  0.04501
C
0    47.93  0.04501
1    52.01  0.04501
A*B
0 0  54.89  0.07014
0 1  24.97  0.07014
1 0  44.97  0.07014
1 1  75.05  0.07014
A*C
0 0  37.89  0.07014
0 1  41.97  0.07014
1 0  57.97  0.07014
1 1  62.05  0.07014
B*C
0 0  52.89  0.07014
0 1  46.97  0.07014
1 0  42.97  0.07014
1 1  57.05  0.07014

Grouping Information Using Tukey Method and 95.0% Confidence

A   N   Mean  Grouping
1  16  60.01  A
0  16  39.93    B

Means that do not share a letter are significantly different.

Grouping Information Using Tukey Method and 95.0% Confidence

B   N   Mean  Grouping
1  16  50.01  A
0  16  49.93  A

Means that do not share a letter are significantly different.

Grouping Information Using Tukey Method and 95.0% Confidence

C   N   Mean  Grouping
1  16  52.01  A
0  16  47.93    B

Means that do not share a letter are significantly different.

Grouping Information Using Tukey Method and 95.0% Confidence
```

```
A  B  N   Mean  Grouping
1  1  8  75.05  A
0  0  8  54.89    B
1  0  8  44.97      C
0  1  8  24.97        D

Means that do not share a letter are significantly different.

Grouping Information Using Tukey Method and 95.0% Confidence

A  C  N   Mean  Grouping
1  1  8  62.05  A
1  0  8  57.97    B
0  1  8  41.97      C
0  0  8  37.89        D

Means that do not share a letter are significantly different.

Grouping Information Using Tukey Method and 95.0% Confidence

B  C  N   Mean  Grouping
1  1  8  57.05  A
0  0  8  52.89    B
0  1  8  46.97      C
1  0  8  42.97        D

Means that do not share a letter are significantly different.
```

We present in the figure below the interaction plots from the analysis.



14.7.1 Confounding in the 3^n Series

For the 3^n series, the block sizes are multiples of 3, e.g., for a 3^3 experiment, i.e., a total of 27 treatment combinations, it is possible to have confounded designs in blocks of nine or three.

A fairly simple design which uses this technique of confounding is the split-plot design discussed in the next chapter.

14.8 Fractional Replication

Complete or full factorial experiments cannot always be feasible because the number of treatment combinations may be very large, and adequate experimental material may not be available to have a balanced design. Fractional replication is a factorial experiment in which only an adequately chosen fraction of the treatment combinations is utilized in the experiment. In most practical situations, even though there may be many factors, but very few are usually important and these are often dominated by main effects and low-order interactions.

Let us consider the case of a 2^3 factorial experiment with factors A, B, and C each at two levels.

Effects representation and their contrasts in the 2^3 *design.*

Factor-level combinations	Constant (I)	A	B	AB	C	AC	BC	ABC	Response
(1)	+1	−1	−1	+1	−1	+1	+1	−1	y_0
(a)	+1	+1	−1	−1	−1	−1	+1	+1	y_1
(b)	+1	−1	+1	−1	−1	+1	−1	+1	y_2
(ab)	+1	+1	+1	+1	−1	−1	−1	−1	y_3
(c)	+1	−1	−1	+1	+1	−1	−1	+1	y_4
(ac)	+1	+1	−1	−1	+1	+1	−1	−1	y_5
(bc)	+1	−1	+1	−1	+1	−1	+1	−1	y_6
(abc)	+1	+1	+1	+1	+1	+1	+1	+1	y_7

14.8.1 Constructing a 2^{n-1} *Fractional Factorial Design*

As a simple example, suppose we wish to construct a 2^{3-1} fractional factorial. This means that we wish to use only $2^{3-1} = 2^2$ four treatment combinations (of the possible eight treatment combinations). The question arises, which four treatment combinations should we use for such a half replicate $\frac{2^3}{2^2} = \frac{1}{2}$. In general, fractional factorial of the form 2^{n-1} is described as half-replicate fractional factorial. To answer the earlier question, the principle of fractional factorial is based on prior information on higher level interaction term/s being assumed having negligible effects. In our case, let us assume that the ABC interaction has negligible effect. Thus, the ABC contrast in the above table can be confounded with blocks resulting in the choice of either the positive treatment combinations: *a, b, c, abc* or the negative treatment combinations: *(1), ab, ac, bc*. If we decide to go with the positive four treatment combinations, then they will be referred to as the *principal fraction*, while the negatives will be referred to as the *complimentary fraction* and in this case,

we would write I = +ABC as the defining relationship obtained by setting the confounded effect to I.

Table 14.30 Effects representation and their contrasts in the 2^{3-1} design, I = +ABC

Treatment combinations	I	A	B	AB	C	AC	BC	ABC	Response
(a)	+1	+1	−1	−1	−1	−1	+1	+1	y_1
(b)	+1	−1	+1	−1	−1	+1	−1	+1	y_2
(c)	+1	−1	−1	+1	+1	−1	−1	+1	y_4
(abc)	+1	+1	+1	+1	+1	+1	+1	+1	y_7

To obtain other confounding effects in a half fraction, we generate the contrasts for the other effects through multiplication as follows using the defining equation I = +ABC.

$$I = ABC$$
$$A = A^2BC = BC$$
$$B = AB^2\ C = AC$$
$$C = ABC^2 = AB$$

Thus, the contrasts that estimate A are the same as the contrasts that estimate BC. Similarly for B and AC and for C and AB. In other words, we would say that the main effects A, B and C are aliased with BC, AC, and AB, respectively. Similarly, if we have used $I = -ABC$ as the defining equation, then the corresponding confounding effects would in this case be:

$$I = -ABC$$
$$A = -A^2BC = -BC$$
$$B = -AB^2\ C = -AC$$
$$C = -ABC^2 = -AB$$

With corresponding effect and contrast representation as follows:

Table 14.31 Effects representation and their contrasts in the 2^{3-1} design. I = −ABC

Treatment combinations	I	A	B	AB	C	AC	BC	ABC	Response
(1)	+1	−1	−1	+1	−1	+1	+1	−1	y_0
(ab)	+1	+1	+1	+1	−1	−1	−1	−1	y_3
(ac)	+1	+1	−1	−1	+1	+1	−1	−1	y_5
(bc)	+1	−1	+1	−1	+1	−1	+1	−1	y_6

14.8.2 Calculating the Effects:

Using the four treatment combinations in Table 14.30, we can calculate the main effects as:

$$A = \frac{(a+abc)}{2} - \frac{(b+c)}{2} = \frac{1}{2}(a-b-c+abc) = \ell_A \tag{14.14}$$

$$B = \frac{(b+abc)}{2} - \frac{(a+c)}{2} = \frac{1}{2}(-a+b-c+abc) = \ell_B \tag{14.15}$$

$$C = \frac{(c+abc)}{2} - \frac{(a+b)}{2} = \frac{1}{2}(-a-b+c+abc) = \ell_C \tag{14.16}$$

Since we showed earlier that BC, AC and AB are aliases of A, B, and C, respectively, therefore, $\frac{1}{2}(a-b-c+abc)$ in effect is estimating A+BC (the main effect and the two-factor interaction term BC). We can, therefore, write the effects in a 2^{3-1} design as:

$$A + BC = \frac{1}{2}(a-b-c+abc) \tag{14.17}$$

$$B + AC = \frac{1}{2}(-a+b-c+abc) \tag{14.18}$$

$$C + AB = \frac{1}{2}(-a-b+c+abc) \tag{14.19}$$

In this design, we have four treatment combinations and hence, three degrees of freedom to estimate A+BC, B+AC and C+AB. This design is, therefore, useful if the two-way interactions are not important or of interest, since the two-way effects can only be estimated in combination with the main effects. This design is often referred to as *Resolution III Design*, because the generator ABC has three letters, but the properties of the design and all Resolution III designs are such that the main effects are confounded with the two-way interactions.

If the complimentary fraction with the defining equation, $I = -ABC$ has been used, then from Table 14.31, we see again that the estimating A, B, and C are equivalent to actually estimating A−BC, B−AC, and C−AB, respectively. That is,

$$A = \frac{(ab+ac)}{2} - \frac{(1+bc)}{2} = \frac{1}{2}(-1+ab+ac-bc) = \ell'_A$$

$$B = \frac{(ab+bc)}{2} - \frac{(1+ac)}{2} = \frac{1}{2}(-1+ab-ac+bc) = \ell'_B$$

$$C = \frac{(ac+bc)}{2} - \frac{(1+ab)}{2} = \frac{1}{2}(-1-ab+ac+bc) = \ell'_C$$

If we choose to run first the principal fraction and subsequently also run the complimentary fraction, the two can then be combined to form a full factorial. In such a situation, then,

$$\frac{1}{2}(\ell_A + \ell_A^{'}) = \frac{1}{2}(A + BC + A - BC) \rightarrow A$$
$$\frac{1}{2}(\ell_A - \ell_A^{'}) = \frac{1}{2}(A + BC - A + BC) \rightarrow BC$$

We note that in both Tables 14.30 and 14.31, the main effects are orthogonal to each other (Note: An experimental design is orthogonal if the effects of any factor balance out (sum to zero) across the effects of the other factors). In our case, the products of their corresponding elements sum to zero.

14.8.3 The One-Quarter Fraction of the 2^n Design: 2^{n-2}

Consider a 2^5 factorial indexed by factors A, B, C, D, and E each at two levels. We, thus, have a total of 32 treatment combinations. To construct a half replicate (a 2^{5-1}), we need only one defining contrast or generator, usually the highest order interaction. Let us say, the ABCDE interaction. Then, in this case,

$$I = ABCDE$$

In this case, as in the previous case, we would need just one generator, and the treatment combinations would be:

a b c d e abc abd abe acd ace ade bcd bce bde cde abcde

To construct a quarter replicate, denoted by 2^{n-2}, we would need two generators in the defining relationship and their generalized interaction. For instance, if we decide to use say $I = ABCDE = BCDE$, then their generalized interaction A is also confounded. Thus, the defining relationship in this case would be $I = ABCDE = BCDE = A$. This certainly will not be a good design since we are completely losing information on the main effect A. Suppose instead, we choose the defining relationship:

$$I = ABCD = ABE = CDE$$

The even {0, 2, 4} treatment combinations in the ABCD are:

(1), ab, ac, ad, bc, cd, abd, e, abe, ace, ade, bce, bde, cde, abcde

Similarly, the even treatment combinations having {0, 2, 4} letters in common to the generator ABE are:

(1), ab, ae, be, c, abc, ace, bce, d, abd, ade, bde, cd, abcd, acde, bcde

The eight common treatment combinations to these two generators (and of course also to the CDE) are:

(1), ab, cd, ace, bce, ade, bde, abcd

The corresponding aliases are therefore:

$$I = ABCD = ABE = CDE$$
$$A = BCD = BE = ACDE$$
$$B = ACD = AE = BCDE$$
$$C = ABD = ABCE = DE$$
$$D = ABC = ABDE = CE$$
$$E = ABCDE = AB = CD$$
$$AC = BD = BCE = ADE$$
$$AD = BC = BDE = ACE$$

We observe that main effects have two-factor aliases. Consequently, the only estimable two factor effects are AC = BD and AD = BC. The analysis of variance table would therefore look like (for one replicate)

Source	d.f.
Main effects	5
2-factor	2
Total	7

14.8.4 An Example:

In a $\frac{1}{4}$ fractional replicate experiment, an agronomist wishes to test the effects of five fertilizers each at two levels on the yield of maize. The data from this experiment is presented in Table 14.32.

Table 14.32 Yield in kg/acre from a $2^{(5-2)}$ experiment in two replicates

	Treatment combinations							
	(1)	ab	cd	ace	bce	ade	bde	abcd
Rep. I	127	135	158	142	138	129	146	132
Rep. II	131	136	161	140	140	134	150	133

We can implement the analysis of this data in MINITAB by coding the factor levels 0/1 and whereas, for instance, the treatment combination (1) is coded as {0 0 0 0 0}, while the combination *bde* is similarly coded as {0 1 0 1 1}.

The MINITAB codes as well as the analysis of variance table are presented below.

```
    A  B  C  D  E  Rep    Y
    0  0  0  0  0    1  127
    1  1  0  0  0    1  135
    0  0  1  1  0    1  158
    1  0  1  0  1    1  138
    0  1  1  0  1    1  138
    1  0  0  1  1    1  129
    0  1  0  1  1    1  146
    1  1  1  1  0    1  132
    0  0  0  0  0    2  131
    1  1  0  0  0    2  136
    0  0  1  1  0    2  161
    1  0  1  0  1    2  140
    0  1  1  0  1    2  140
    1  0  0  1  1    2  134
    0  1  0  1  1    2  150
    1  1  1  1  0    2  133

MTB > GLM 'Y' = Rep A B 'C' D E A*C A*D;
SUBC>   Brief 2 .

General Linear Model: Y versus Rep, A, B, C, D, E

Factor  Type   Levels  Values
Rep     fixed       2  1, 2
A       fixed       2  0, 1
B       fixed       2  0, 1
C       fixed       2  0, 1
D       fixed       2  0, 1
E       fixed       2  0, 1


Analysis of Variance for Y, using Adjusted SS for Tests

Source  DF   Seq SS  Adj SS  Adj MS       F      P
Rep      1    30.25   30.25   30.25   27.32  0.001
A        1   342.25  342.25  342.25  309.13  0.000
B        1     4.00    4.00    4.00    3.61  0.099
C        1   169.00  169.00  169.00  152.65  0.000
D        1   210.25  210.25  210.25  189.90  0.000
E        1     0.25    0.25    0.25    0.23  0.649
A*C      1    72.25   72.25   72.25   65.26  0.000
A*D      1   625.00  625.00  625.00  564.52  0.000
Error    7     7.75    7.75    1.11
Total   15  1461.00


S = 1.05221   R-Sq = 99.47%   R-Sq(adj) = 98.86%

MTB > GLM 'Y' = Rep A B 'C' D E A*C A*D;
SUBC>   SMeans C4000;
SUBC>   Brief 0;
SUBC>   Interact 'A' 'C'  'D'.
MTB > GFInt 'A' 'C'  'D';
SUBC>   Responses 'Y';
SUBC>   FMeans C4000.
```

We see that all the main effects except that of B and E at the 5 % level of significance are all significant. Further, the two interaction terms are highly significant and hence, we need to focus on these interaction terms rather than the main effects.

In general a $\frac{1}{2^p}$ fraction of a 2^n design is designated as a 2^{n-p} fractional factorial and such a design would require:

- p design generators with $2^p - p - 1$ generalized interactions.
- Each effect would have $2^p - 1$ aliases.

For instance, a 2^{6-2} fractional factorial would have $p = 2$ design generators and $2^2 - 2 - 1 = 1$ generalized interactions. There would, therefore, be $2^2 - 1 = 3$ aliases for each effect.

14.9 2^{n-p} Resolution III and IV Designs

Generally, fractional replications are designs grouped into classes based on their resolutions. Most common are Resolution III, IV, and V. However, we will concern ourselves in this chapter on the former two which are defined as follows:

Resolution III: This is a design in which no main effect is confounded with any other main effect, but main effects are confounded with two-factor interactions and two-factor interactions are confounded with other two-factor interactions.

Resolution IV: This is a design in which no main effect is confounded with any other main effect or two-factor interactions, but two-factor interactions are confounded with one another.

It seems that the resolution of a design is determined by the smallest number of letter appearing in the design generator. Thus, a 2^{3-1} generated from the ABC contrast is a Resolution III design and would be designated as 2^{3-1}_{III} design. Similarly, a 2^{4-1} design with the ABCD generator contrast is a design Resolution IV, that is, a 2^{4-1}_{IV} design.

For instance, a 2^{6-2} fractional design with the two generators ABDE and ABCF and, thus, their generalized Interaction $ABDE \times ABCF = A^2B^2CDEF = CDEF$ has the defining relationship:

$$I = ABDE = ABCF = CDEF$$

Here, the smallest number of letters for the generators is four, and thus, this design is Resolution IV design. That is, it is a 2^{6-2}_{IV} design. Similarly, suppose a 2^{7-4} fractional replication has the following four generators ABD, ACE, BCF, and ABCG and $2^4 - 4 - 1 = 11$ generalized interactions and each effect would have $2^4 - 1 = 15$ aliases. Since the smallest number of letters in the generators is three, we therefore have a 2^{7-4}_{III} design here. That is, a Resolution III design.

14.10 Logistic Regression for a Factorial Study

The data in Table 14.33 relate to data from Lombard and Doering (1947) from a survey of knowledge about cancer. The data has a 2^4 factorial treatment combinations, arrangement with n being the number of individuals surveyed in this category and r the number who gave a good score in the response to questions about cancer knowledge. The four factors are: (A) newspaper reading; (B) listening to radio; (C) solid reading; (D) attendance at lectures. (*Note: the data has previously been analyzed in Armitage and Berry (1987)*).

Table 14.33 A 2^4 factorial set of proportions

Factor combinations	Number of individuals n	Number with good score r	Factor combinations	Number of individuals n	Number with good score r
(1)	477	84	d	12	2
a	231	75	ad	13	7
b	63	13	bd	7	4
ab	94	35	abd	12	8
c	150	67	cd	11	3
ac	378	201	acd	45	27
bc	32	16	bcd	4	1
abc	169	102	abcd	31	23

If we let p_i be the proportion of individuals giving good score in treatment combination $i = 1, 2, \ldots, 16$, then the saturated model would have the form:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 D + \beta_{12} AB + \beta_{13} AC + \cdots + \beta_{1234} ABCD \tag{14.20}$$

The above model would not have any degree of freedom and would thus produce a perfect fit to the data. However, we can overcome this by fitting only main effects and the second order interactions. That is the model,

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 D + \beta_{12} AB + \beta_{13} AC + \cdots + \beta_{34} CD \tag{14.21}$$

The model in (14.21) is implemented in MINITAB with the following statements and the partial outputs including the data display are presented.

```
MTB > print c1-c6

Data Display

Row  A  B  C  D    n    r
  1  0  0  0  0  477   84
  2  1  0  0  0  231   75
  3  0  1  0  0   63   13
  4  1  1  0  0   94   35
  5  0  0  1  0  150   67
  6  1  0  1  0  378  201
  7  0  1  1  0   32   16
  8  1  1  1  0  169  102
  9  0  0  0  1   12    2
 10  1  0  0  1   13    7
 11  0  1  0  1    7    4
 12  1  1  0  1   12    8
 13  0  0  1  1   11    3
 14  1  0  1  1   45   27
 15  0  1  1  1    4    1
 16  1  1  1  1   31   23

MTB > Blogistic 'r' 'n' = A B 'C' D A*B A*C A*D B*C B*D   C*D  ;
SUBC>   ST;
SUBC>   Logit;
SUBC>   Brief 2.
Binary Logistic Regression: r, n versus A, B, C, D

Link Function: Logit


Response Information

Variable Value       Count
r        Event         668
         Non-event    1061
n        Total        1729


Logistic Regression Table

                                                    Odds     95% CI
Predictor        Coef   SE Coef        Z       P Ratio  Lower  Upper
Constant     -1.53668  0.116547  -13.19  0.000
A             0.779673  0.173051    4.51  0.000  2.18   1.55   3.06
B             0.247441  0.264111    0.94  0.349  1.28   0.76   2.15
C              1.29218  0.187651    6.89  0.000  3.64   2.52   5.26
D             0.196022  0.475627    0.41  0.680  1.22   0.48   3.09
A*B        -0.0035985  0.285735   -0.01  0.990  1.00   0.57   1.74
A*C         -0.387508  0.231841   -1.67  0.095  0.68   0.43   1.07
A*D          0.813517  0.469828    1.73  0.083  2.26   0.90   5.67
B*C         0.0069559  0.256503    0.03  0.978  1.01   0.61   1.66
B*D          0.484744  0.405235    1.20  0.232  1.62   0.73   3.59
C*D         -0.843755  0.430921   -1.96  0.050  0.43   0.18   1.00


Log-Likelihood = -1046.665
Test that all slopes are zero: G = 213.460, DF = 10, P-Value = 0.000


Goodness-of-Fit Tests

Method           Chi-Square  DF      P
Pearson             2.84098   5  0.724
Deviance            2.79968   5  0.731
Hosmer-Lemeshow     0.28266   4  0.991
```

We observe that the model fits the data well with a deviance of 2.79968 on 5 df and corresponding $p - value = 0.731$. The five degrees of freedom correspond collectively to the omitted effects, ABC, ABD, ACD, BCD, and

ABCD. However, all the second-order interactions are not significant at $\alpha = 0.05$ level of significance except the CD interaction (and barely too!). We, therefore, next include only the CD interaction in our next model, that is,

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 D + \beta_{34} CD \tag{14.22}$$

These results of implementing model (14.22) are presented in the MINITAB partial output below.

```
Logistic Regression Table

                                                        Odds      95% CI
Predictor       Coef    SE Coef       Z      P      Ratio  Lower  Upper
Constant    -1.48688  0.0979798  -15.18  0.000
A           0.646847   0.115566    5.60  0.000       1.91   1.52   2.39
B           0.301609   0.122295    2.47  0.014       1.35   1.06   1.72
C            1.03747   0.115302    9.00  0.000       2.82   2.25   3.54
D           0.894327   0.318555    2.81  0.005       2.45   1.31   4.57
C*D        -0.717191   0.391255   -1.83  0.067       0.49   0.23   1.05


Log-Likelihood = -1050.402
Test that all slopes are zero: G = 205.985, DF = 5, P-Value = 0.000


Goodness-of-Fit Tests

Method           Chi-Square  DF      P
Pearson              9.9194  10  0.448
Deviance            10.2742  10  0.417
Hosmer-Lemeshow      2.5278   4  0.640
```

Again, this model fits the data with a deviance of 10.2742 on 10 d.f. with a $p-value$ of 0.417. However, the model is not the most parsimonious because the CD interaction is no longer significant ($p-value = 0.067$). Hence, we are reduced to model (14.23) containing only the main effects of the study.

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 D \tag{14.23}$$

These results of implementing model (14.23) are presented in the MINITAB partial output below.

```
MTB > Name c7 "EPRO1"
MTB > Blogistic 'r' 'n' = A B 'C' D ;
SUBC>   ST;
SUBC>   Logit;
SUBC>   Eprobability 'EPRO1';
SUBC>   Brief 2.

Binary Logistic Regression: r, n versus A, B, C, D

Link Function: Logit

Logistic Regression Table

                                                        Odds      95% CI
Predictor      Coef     SE Coef        Z        P Ratio   Lower  Upper
Constant   -1.46043   0.0964047   -15.15    0.000
A          0.649798    0.115421     5.63    0.000  1.92    1.53   2.40
B          0.310105    0.122198     2.54    0.011  1.36    1.07   1.73
C          0.980614    0.110729     8.86    0.000  2.67    2.15   3.31
D          0.420353    0.190971     2.20    0.028  1.52    1.05   2.21


Log-Likelihood = -1052.060
Test that all slopes are zero: G = 202.669, DF = 4, P-Value = 0.000


Goodness-of-Fit Tests

Method           Chi-Square  DF      P
Pearson             13.6067  11  0.256
Deviance            13.5909  11  0.256
Hosmer-Lemeshow      2.8851   4  0.577
```

The model fits the data with a deviance of 13.5909 on 11 d.f. and a p value of 0.256. The estimated logistic model therefore is given by:

$$\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -1.4604 + 0.6498\,\mathbf{A} + 0.3101\,\mathbf{B} + 0.9806\,\mathbf{C} + 0.4204\,\mathbf{D} \tag{14.24}$$

While we recognize that the effects might not be orthogonal, however, our results here indicate that the odds of an individual surveyed giving a good score on the knowledge about cancer is 2.67 times for those with solid reading than those who have not read about cancer knowledge. It is 1.92 times higher for those who have read about it in newspapers as against those who had not. Similar interpretations can be given for factors B and D, with B being the least effective score in the response to questions about cancer knowledge.

14.11 Exercises

1. In a 2^4 experiment on the yield of a chemical process, the treatment response from a single replicate of the experiment is given below.

		d_0		d_1	
		c_0	c_1	c_0	c_1
a_0	b_0	8	31	79	77
	b_1	53	12	73	49
a_1	b_0	4	9	68	38
	b_1	43	36	8	23

   Fit an appropriate model to this single replicate data.
2. Construct a half replicate of a 2^5 factorial system such that there would be two blocks in each of the two replications. Give a sketch of the analysis of variance.
3. The following exercise is adapted from Hoshmand (1994). The data in the table below relate to a 2^5 experiment on corn yield conducted in a half-replicate with two blocks each containing two replicates.

	Block 1			Block 2	
Treatment	Rep. I	Rep. II	Treatment	Rep. I	Rep. II
(1)	132	125	ac	112	114
ac	151	145	ad	122	132
ab	136	138	bc	145	148
be	144	142	bd	161	158
acde	171	162	de	145	138
cd	154	159	ce	162	168
bcde	143	138	abde	144	140
abcd	132	136	abce	155	159

Grain yield from the experiment in (bu/acre)

   (a) Perform the analysis of variance.
   (b) Which of the main and interaction effects are highly significant?
   (c) Plot any two way significant interactions and draw your conclusions.

4. A 2^4 factorial experiment was conducted in four blocks of four experimental units each and BCD, and ABD were used as defining contrasts.

   (a) What other effects were confounded with blocks?
   (b) Could there have been a better choice of defining contrasts for this design? Explain.
   (c) Write down the structure of ANOVA table for this design for a single replicate? Two replicates?

5. For the following list of designs:

$$\begin{array}{lll} \text{(i) } 2^{5-1} & \text{(ii) } 2^{7-3} & \text{(iii) } 2^{6-1} \\ \text{(iv) } 2^{6-2} & \text{(v) } 2^{6-3} & \text{(vi) } 2^{7-4} \end{array}$$

For each of the design above:

(a) What is the fraction of the full design?
(b) Number of generators required?
(c) Number of generalized interactions.
(d) Number of aliases for each effect.
(e) The number of experimental units required to run the design.

6. A 2^{5-2} fractional factorial design is proposed with two competing generating relations:

$$\begin{aligned} &\text{(i)} \quad I = ABCD \quad = BCE \\ &\text{(ii)} \quad I = ABCDE = ABCD \end{aligned}$$

(i) What fraction of a 2^5 design will it be?
(ii) Generate the aliases and other generalized interactions for the two designs.
(iii) Generate the treatment combinations for a single replicate of each design.
(iv) Which defining relationship is preferred for the design? Explain.

7. Analyze the following 2^3 factorial involving three factors, A, B, and C. Identify which effect is confounded with blocks in each replicate.

Replicate I				Replicate II				Replicate III			
Block 1		Block 2		Block 3		Block 4		Block 5		Block 6	
ab	(101)	b	(88)	(1)	(125)	ab	(115)	bc	(75)	a	(53)
abc	(111)	a	(90)	abc	(95)	c	(95)	ac	(100)	abc	(76)
(1)	(75)	bc	(115)	ac	(80)	bc	(90)	(1)	(55)	b	(65)
c	(55)	ac	(75)	b	(100)	a	(80)	ab	(92)	c	(82)

8. The data below is adapted from and give the yields (in cwt. per acre) of a 2^4 experiment on soybeans. The treatments are all combinations of:

Dung	:10 tons per acre (d) or nil
Nitrochalk	:$\frac{1}{2}$ cwt. per acre (n)or nil
Superphosphate	:$\frac{1}{2}$ cwt. per acre (p) or nil
Muriate of potash	:1 cwt. per acre (k) or nil

The design plan is:

Block 1 A				Block 1B			
nk	64.4	(1)	40.1	p	33.6	npk	51.9
np	32.4	dn	63.4	dnp	56.9	n	27.3
dp	53.7	pk	44.7	d	58.6	k	43.2
dk	47.4	dnpk	65.2	dnk	69.8	dpk	59.7

Block 2 A				Block 2B			
(1)	62.6	dp	72.8	n	64.1	k	59.7
nk	67.3	dk	77.2	dnk	96.4	dpk	60.2
np	49.6	dnpk	78.2	p	52.8	dnp	68.5
dn	74.7	pk	74.1	d	70.9	npk	52.4

(a) Which effect(s) are confounded?
(b) Analyze the data and draw your conclusions.
(c) Obtain the S.E.s of the effect means.

9. Construct a single replicate of a 2^5 factorial in blocks of four using as your defining contrasts effects ADE and BCE.
10. A study is conducted to determine the effect of water level and type of plant on the overall stem length of pea plants. Three water levels and two plant types are used. Eighteen leafless plants are available for study. These plants are randomly divided into three subgroups, and then water levels are randomly assigned to the groups. A similar procedure is followed with 18 conventional plants. These data resulted (stem length is given in centimeters) in the following table:

Factor B (plant type)	Factor A (water level)			Total
	Low	Medium	High	
Leafless	69.0	96.1	121.0	1788
	71.3	102.3	122.9	
	73.2	107.5	123.1	
	75.1	103.6	125.7	
	74.4	100.7	125.2	
	75.0	101.8	120.1	
Sub-Total	438	612	738	1788
Conventional	71.1	81.0	101.1	1578
	69.2	85.8	103.2	
	70.4	86.0	106.1	
	73.2	87.5	109.7	
	71.2	88.1	109.0	
	70.9	87.6	106.9	
Sub-Total	426	516	636	1578
Total	864	1128	1374	3366

Use MINITAB to conduct the analysis and answer the following questions:

a How many treatment combinations are in this experiment?
b How many replications are there for each treatment combination?
c Why is replication necessary?
d What type of design was employed for this experiment?

11. The partially completed ANOVA table for a balanced ANOVA is given below:

Source	DF	SS	MS	F
Factor A	3	605.272	-	-
Factor B	-	-	1145.679	-
$A * B$	6	-	-	7.90
Error	24	-	10.943	
Total	-	-		

(a) What type of design was employed in this experiment?
(b) Determine the number of levels of each factor?
(c) Complete the ANOVA Table.
(d) Test to determine significant effects. Use $\alpha = .10$.
(e) How many replications are used in this experiment?.

12. The partially completed ANOVA table for a balanced ANOVA is given below:

Source	DF	SS	MS	F
Blocks	2	9.00	–	
A	2	12.00	–	–
B	–	–	34.50	–
$A * B$	4	–	–	0.53
Error	16	–	1.50	
Total	–	–		

(a) What type of design was employed in this experiment?
(b) Determine the number of levels of each factor?
(c) Complete the ANOVA Table.
(d) Test to determine significant effects. Use $\alpha = .10$.
(e) How many replications are used in this experiment?

13. An experiment was conducted on survival of *Salmonella typhimurium* to investigate the effects of three levels of sorbic acid and six levels of water activity (a_w). The data displayed are log (density/ml) measured 7 days after the imposition of the treatments.

Factor A (Sorbic acid level)	Factor B (water activity level)					
	0.98	0.94	0.90	0.86	0.82	0.78
0	8.19	6.65	5.87	5.06	4.85	4.31
	8.37	6.70	5.98	5.35	4.31	4.34
	8.33	6.25	6.14	5.01	4.52	4.20
100 ppm	7.64	6.52	5.01	4.85	4.29	4.13
	7.79	6.19	5.28	4.95	4.43	4.39
	7.59	6.51	5.78	4.29	4.18	4.18
200 ppm	7.14	6.33	5.20	4.41	4.26	4.93
	6.92	6.18	5.10	4.40	4.27	4.12
	7.19	6.43	5.43	4.79	4.37	4.15

Answer the following questions:

a How many treatment combinations are in this experiment?
b How many replications are there for each treatment combination?
c Why is replication necessary?
d What type of design was employed for this experiment?
c Study the SAS program and output attached for this problem and interpret the printout. You will need to test each effect and draw your conclusions.

Chapter 15
The Split-Plot Design

15.1 Introduction

For the split-plot design, we are concerned with two or more factors, but we wish for more precise information on some of them than on others. If we are interested in more accurate information, for instance, on factor B than on A, then the usual scheme is to assign the various levels of factor A at random to whole plots (main plots) in each replicate as in a randomized complete block design. Following this, the levels of B are assigned at random to the split plots (subplots) within each whole plot. Such a scheme of randomization may arise not only from the desire for more precise information on one factor that on another but also because of the nature of the factors and the way in which they must be applied to the experimental units.

In agriculture, whole plots are usually large areas of land and the subplots are small areas of land. For example, several varieties of a crop could be planted in different fields (whole plots), one variety to a field. Then each field could be divided into five subplots, for example, and each subplot could be treated with a different type of fertilizer. Here the crop varieties are the main treatments and the different fertilizers are the subtreatments. The split-plot design is also very useful in many scientific or industrial experiments where some factors require large experimental units and other factors require small ones. We notice that the whole plot treatments in a split-plot design are confounded with the whole plots, and the subplot treatments are not confounded. Therefore, it is better to assign the factor of interest to the subplots if possible.

The linear statistical model for a split plot design with one observation per experimental unit is

$$Y_{ijk} = \mu + b_i + \alpha_j + \delta_{ij} + \beta_k + (\alpha\beta)_{jk} + e_{ijk}. \tag{15.1}$$

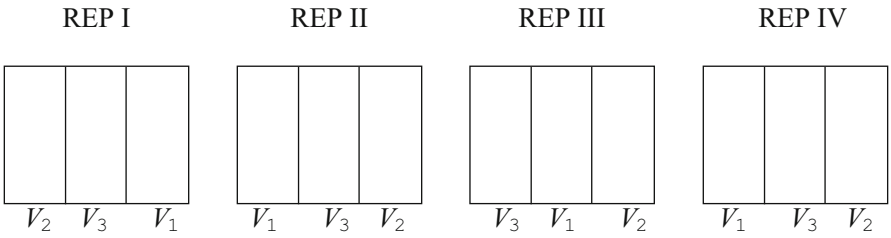
$i = 1, 2, \cdots, r$; $j = 1, 2, \cdots, t$; and $k = 1, 2, \cdots, s$ where b_i and α_j are replicate (block) and whole plot treatment effects respectively. δ_{ij} is the whole plot error, while β_k, $(\alpha\beta)_{jk}$, and e_{ijk}, respectively represent the subplot treatment, interaction effects, and random errors.

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_15,

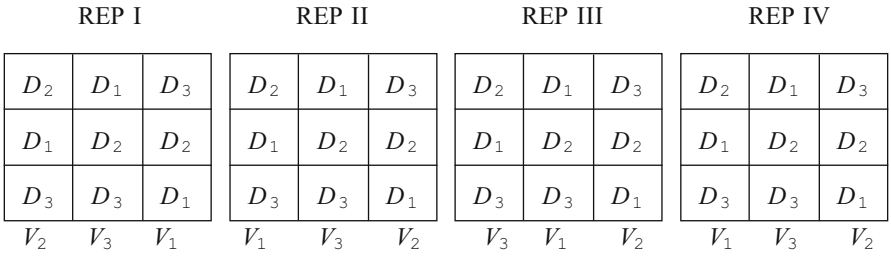
Basically, the split may be viewed as consisting of two designs: (i) a *main-plot design* and (ii) a *subplot design*. The main-plot design is used to allocate treatments to the main plots. For a randomized complete block split-plot design, for example, the main-plot design is a randomized complete block (RCB). That is, the main-plot treatments are assigned to blocks. In terms of replicates, each replicate contains one main-plot treatment as, say for instance in a completely randomized design.

Example 15.1.1

Suppose we are investigating the yield of $a = 3$ varieties of millet (A_1, A_2, A_3) at $b = 3$ three densities D_1, D_2, and D_3 and we wish to replicate the experiment $r = 4$ times. If the varieties are to go as main-plot treatments, the layout below divides the experimental area into four replicates since $r = 4$. These are then further divided each into three blocks, each block containing the randomized main-plot treatments V_1, V_2, and V_3.



Since we have three subplot treatments, now divide each block into three units and randomly assign the three plant densities as shown in the layout below. Note that columns are blocks in this layout



The structure of the analysis of variance table becomes:

Source	d.f.	Example d.f.
Blocks (replicates)	$r-1$	3
A	$a-1$	2
Main-plot error	$(a-1)(r-1)$	6
B	$(b-1)$	2
AB	$(a-1)(b-1)$	4
Subplot error	$a(r-1)(b-1)$	18
Total	$abr-1$	35

The main effects A will be tested from

$$F = \frac{\text{MSA}}{\text{Main Plot MS}}$$

where MSA is the mean square for A and Main Plot MS is the mean square obtained from the main-plot error line. The effects of B and AB are tested as usual with the subplot error mean square. The main-plot error sum of squares is the interaction between blocks and A sum of squares.

Example 15.1.2: Analysis of Split-Plot Experiment

The response of six varieties of lettuce, grown in frames, to various uncovering dates was investigated in a split-plot experiment with four blocks. The main-plot treatments were three uncovering dates and each main plot was split into six split plots for the six varieties. The data in Table 15.1 is from this experiment.

Table 15.1 Data for this experiment

Uncovering date	Variety	Blocks I	II	III	IV	Treatments totals
X	A	11.8	7.5	9.7	6.4	35.4
	B	8.3	8.4	11.8	8.5	37.0
	C	9.2	10.6	11.4	7.2	38.4
	D	15.6	10.8	10.3	14.7	51.4
	E	16.2	11.2	14.0	11.5	52.9
	F	9.9	10.8	4.8	9.8	35.3
Main-plot total		71.0	59.3	62.0	58.1	
Y	A	9.7	8.8	12.5	9.4	40.4
	B	5.4	12.9	11.2	7.8	37.3
	C	12.1	15.7	7.6	9.4	44.8
	D	13.2	11.3	11.0	10.7	46.2
	E	16.5	11.1	10.8	8.5	46.9
	F	12.5	14.3	15.9	7.5	50.2
Main-plot total		69.4	74.1	69.0	53.3	
Z	A	7.0	9.1	7.1	6.3	29.5
	B	5.7	8.4	6.1	8.8	29.0
	C	3.3	6.9	1.0	2.6	13.8
	D	12.6	15.4	14.2	11.3	53.5
	E	12.6	12.3	14.4	14.1	53.4
	F	10.2	11.6	10.4	12.2	44.4
Main-plot total		51.4	63.7	53.2	55.3	
Block totals		191.8	197.1	184.2	166.7	739.8

Analysis

(i) **Analysis of Main-Plot Section**
Since there are three main plots and four replicates, the analysis of the main-plot section therefore involves analysis based on observations in $3 \times 4 = 12$ plot observations and their corresponding subtotals. These are presented in Table 15.2.

Table 15.2 Subtotals from the 4×3 main-plot observations

	Blocks				
	I	II	III	IV	Total
X	71.0	59.3	62.0	58.1	250.4
Y	69.4	74.1	69.0	53.3	265.8
Z	51.4	63.7	53.2	55.3	223.6
Total	191.8	187.1	184.2	166.7	739.8

There are $4 \times 3 \times 6$ plots altogether in the experiment. From Table 15.2, therefore, we have, $\text{CF} = (739.8)^2/72 = 7601.44$. Hence,

$$\text{Block SS} = \frac{191.8^2}{18} + \frac{187.1^2}{18} + \cdots + \frac{166.7^2}{18} - \text{CF} = 29.35$$

$$\text{Uncovering date SS} = \frac{250.4^2}{24} + \frac{265.8^2}{24} + \frac{223.6^2}{24} - \text{CF} = 38.01$$

$$\text{Main plot SS} = \frac{71.0^2}{6} + \frac{59.3^2}{6} + \cdots + \frac{55.3^2}{6} - \text{CF} = 110.92$$

The main-plot analysis of variance table is, therefore, presented in Table 15.3. Here, main plot error is obtained by subtraction

(ii) **Analysis of Split-Plot Section**

The split-plot analysis is based on the $6 \times 3 = 18$ split-plot observations and subtotals in Table 15.4.

Again, the SS are obtained from the following calculations:

Table 15.3 Main-plot ANOVA table

Source	d.f.	SS	MS	F
Blocks	3	29.35		
Uncovering date	2	38.01	19.00	2.62
Error (a)	6	43.56	7.26	
Main-plot total SS	11	110.92		

Table 15.4 Split-plot observations and subtotals

Split-plot	Main-plot treatments			
treatments	X	Y	Z	Total
A	35.4	40.4	29.5	105.3
B	37.0	37.3	29.0	103.3
C	38.4	44.8	13.8	97.0
D	51.4	46.2	53.5	151.1
E	52.9	46.9	53.4	153.2
F	35.3	50.2	44.4	129.9
Total	250.4	265.8	223.6	739.8

$$\text{Varieties SS} = \frac{105.3^2}{12} + \frac{103.3^2}{12} + \cdots + \frac{129.9^2}{12} - \text{CF}$$
$$= 260.51$$

$$\text{Dates} \times \text{Varieties} = \frac{35.4^2}{4} + \frac{40.4^2}{4} + \cdots + \frac{44.4^2}{4} - \text{CF}$$
$$- \text{Varieties SS} - \text{Main effects SS} = 163.70$$

$$\text{Error (b)} = \text{By subtraction} = 227.27$$

The full analysis of variance table is presented in Table 15.5.

Table 15.5 Full analysis of variance table

Source	d.f.	SS	MS	F
Blocks	3	29.35		
Uncovering date	2	38.01	19.00	2.62
Error (a) (B × D) (whole plot error)	6	43.56	7.26	
Varieties	5	260.51	52.10	10.32***
Dates × varieties	10	163.70	16.37	3.24**
Error (b) (subplot error)	45	227.27	5.05	
Total	71	762.40		

** Significant at 1 %; *** Significant at 0.01 %

Note that from Table 15.5 the subplot error is less than the whole plot error. This is the usual case in split-plot designs, since the subplots are generally more homogeneous than the whole plots. The table of means from this analysis is provided in Table 15.6.

Table 15.6 Table of means

	Uncovering date			
Variety	X	Y	Z	Mean
A	8.8	10.1	7.4	8.8
B	9.2	9.3	7.1	8.6
C	9.6	11.2	3.4	8.1
D	12.8	11.6	13.4	12.6
E	13.2	11.7	13.4	12.8
F	8.8	12.6	11.1	10.8
Mean	10.4	11.1	9.3	10.3

S.E. for comparing two date means $= \sqrt{\frac{2\,\text{E}_a}{24}} = 0.78$ (6 d.f.).

S.E. for comparing two variety means $= \sqrt{\frac{2\,\text{E}_b}{12}} = 0.9175$ (45 d.f.).

S.E. for comparing two varieties at a single date $= \sqrt{\frac{2\,\text{E}_b}{4}} = 1.59$ (45 d.f.).

S.E. for comparing differences between two varieties for two dates $= \sqrt{\frac{4\,\text{E}_b}{4}} = 2.25$ (45 d.f.).

S.E. for comparing two dates, either for the same variety or for different varieties

$$= \sqrt{\frac{2(5\,\text{E}_b + \text{E}_a)}{4 \times 6}} = 1.65.$$

Note that in general, the formula for the last S.E. is

$$\sqrt{\frac{2[(b-1)\,\text{E}_b + \text{E}_a]}{r \times b}}$$

where r is the number of blocks (replicates), and b is the number of split-plot treatments.

Because both error mean squares are involved in this S.E., no exact t test is possible.

The analysis of the data in this example (Table 15.1) is implemented in MINITAB as follows:

The interaction plot of Variety and Date is presented in Fig. 15.1, while the main effects' plots for variety and dates is also presented in Fig. 15.2

```
MTB > SET C1
DATA> (1:3)24
DATA> END
MTB > SET C2
DATA> 3(1:6)4
DATA> END
DATA> END
MTB > SET C3
DATA> 18(1:4)
DATA> END
MTB > SET C4
DATA> 11.8 7.5 9.7 6.4 8.3 8.4 11.8 8.5
DATA> 9.2 10.6 11.4 7.2 15.6 10.8 10.3 14.7
DATA> 16.2 11.2 14.0 11.5 9.9 10.8 4.8 9.8
DATA> 9.7 8.8 12.5 9.4 5.4 12.9 11.2 7.8
DATA> 12.1 15.7 7.6 9.4 13.2 11.3 11.0 10.7
DATA> 16.5 11.1 10.8 8.5 12.5 14.3 15.9 7.5
DATA> 7 9.1 7.1 6.3 5.7 8.4 6.1 8.8 3.3 6.9
DATA> 1.0 2.6 12.6 15.4 14.2 11.3 12.6 12.3
DATA> 14.4 14.1 10.2 11.6 10.4 12.2
DATA> END

MTB > GLM 'YIELD' = BLOCKS DATE BLOCKS*DATE VARIETY DATE*VARIETY;
SUBC>   Random 'BLOCKS';
SUBC>   SSquares 1;
SUBC>   Brief 1 .

Factor     Type Levels Values
BLOCKS   random      4 1 2 3 4
DATE      fixed      3 1 2 3
VARIETY   fixed      6 1 2 3 4 5 6

General Linear Model: YIELD versus DATE, BLOCKS, VARIETY

Analysis of Variance for YIELD, using Sequential SS for Tests

Source         DF     Seq SS     Adj SS     Seq MS       F      P
-------------------------------------------------------------------
BLOCKS          3     29.343     29.343      9.781    1.35  0.345
DATE            2     38.003     38.003     19.002    2.62  0.152
BLOCKS*DATE     6     43.566     43.566      7.261    1.44  0.222
VARIETY         5    260.508    260.508     52.102   10.32  0.000
DATE*VARIETY   10    163.698    163.698     16.370    3.24  0.003
Error          45    227.277    227.277      5.051
Total          71    762.395
```

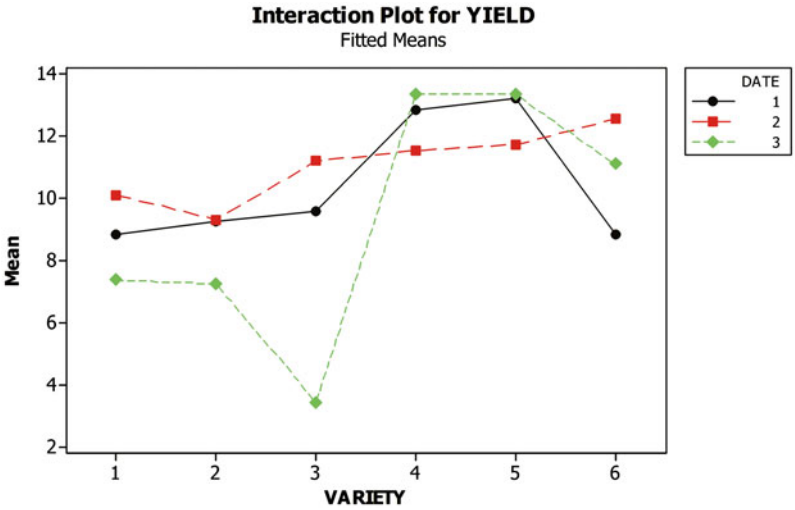


Fig. 15.1 Interaction plots

15.1.1 Summary of Results

Differences between varieties varied with uncovering dates. For uncovering date X, varieties D and E out-yielded all other varieties significantly; for Y differences between varieties were small; for Z varieties D, E, and F gave significantly higher yields than varieties A, B, and C. Although the mean yield declined between Y and Z for four of the six varieties, the decline was significant for variety C only. (All significance statements refer to the 5 % significance level.)

Example 15.1.3

Four strains of perennial rye grass were grown as swards at each of two fertilizer levels. The four strains were S23, New Zealand, Kent, and X (a "hypothetical" strain introduced to illustrate some points of statistical interest). The fertilizer levels were denoted by H, heavy and A, average. The experiment was laid out as four blocks of four whole plots for the varieties, each split in two for the application of fertilizer. The midsummer dry matter yields, in units of 10 lb/acre, are displayed in Data for example 15.1.3.

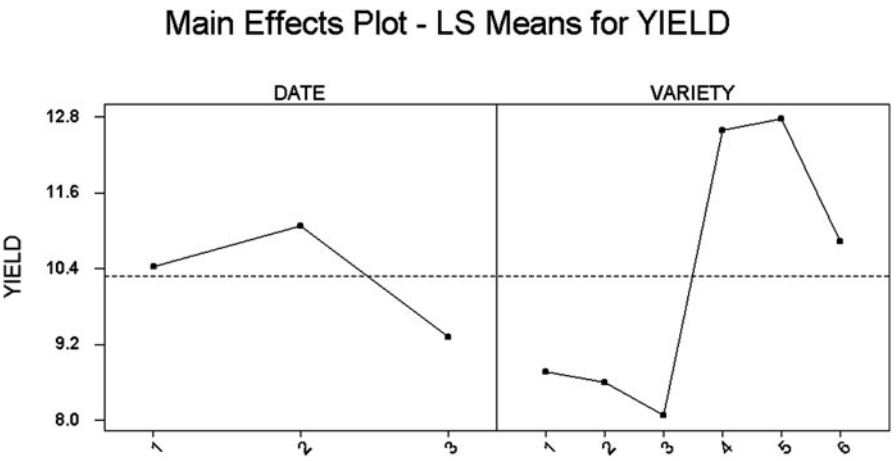


Fig. 15.2 Main effects plots

Table 15.7 Data for Example 14.1.3

	Manuring	Blocks				Manuring	Strain
		1	2	3	4	total	total
S23	H	299	318	284	279	1180	1983
	A	247	202	171	183	803	
New Zealand	H	315	247	289	307	1158	1952
	A	257	175	188	174	794	
X	H	403	439	355	324	1521	2281
	A	222	170	192	176	760	
Kent	H	382	353	383	310	1428	2220
	A	233	216	200	143	792	
Total		2358	2120	2062	1896		8436

Analysis

(i) **Whole-Plot Analysis**

The whole-plot analysis will be based on $4 \times 4 = 16$ observations and subtotals which are displayed in Table 15.8.

Table 15.8 The 16 plot observations for the main-plot analysis

	Blocks				
Variety	1	2	3	4	Total
S23	546	520	455	462	1983
New Zealand	572	422	477	481	1952
X	625	609	547	500	2281
Kent	615	569	583	453	2220
Total	2358	2120	2062	1896	8436

There are $4 \times 4 \times 2 = 32$ plots altogether in the experiment with CF $= 2,223,940.5$. Hence,

$$\text{Block SS} = \frac{2358^2}{8} + \frac{2120^2}{8} + \cdots + \frac{1896^2}{8} - \text{CF} = 13,712.5$$

$$\text{Strains SS} = \frac{1983^2}{8} + \frac{1952^2}{8} + \cdots + \frac{2220^2}{8} - \text{CF} = 10,303.75$$

$$\text{Whole Plot Totals} = \frac{546^2}{2} + \frac{520^2}{2} + \cdots + \frac{453^2}{2} - \text{CF} = 31,530.5$$

The main-plot analysis of variance table is, therefore, presented in Table 15.9.

(ii) **Analysis of Split-Plot Section**

Table 15.10 gives the $2 \times 4 = 8$ observations and subtotals needed for this analysis.

Table 15.9 Main-plot ANOVA table

Source	d.f.	SS	MS	F
Blocks	3	13,712.5		
Strains	3	10,303.75	3434.583	5.16
Error (a)	9	7514.25	834.917	
Main-plot totals	15	31,530.4		

Table 15.10 Subplot observations and subtotals

Split-plot treatments	Main-plot treatments S23	NZ	X	Kent	Total
H	1180	1158	1521	1428	5287
A	803	794	760	792	3149
Total	1983	1952	2281	2220	8436

$$\begin{aligned}\text{Fertilizer SS} &= \frac{5287^2}{16} + \frac{3149^2}{16} - \text{CF} \\ &= 142,845.125\end{aligned}$$

$$\begin{aligned}\text{Stain} \times \text{fertilizer} &= \frac{1180^2}{4} + \frac{1158^2}{4} + \cdots + \frac{792^2}{4} \\ &\quad - \text{CF} - \text{Fertilizer SS} - \text{Main effects SS} \\ &= 14,435.125\end{aligned}$$

$$\text{Error (b)} = \text{By subtraction} = 7982.75$$

The full analysis of variance for the data in Table 15.7 is presented in Table 15.11.

Table 15.11 Full analysis of variance table

Source	d.f.	SS	MS	F
Blocks	3	13,713		
Strains	3	10,304	3435	4.11
Error (a)	9	7514	834.917	
Whole plot totals	15	31,531		
Fertilizer	1	142,845	142,845	214.80***
Strains × fertilizer	3	14,435	4812	7.24***
Error (b)	12	7983	665	
Total	31	196,794		

** Significant at 1 %; *** Significant at 0.01 %

Note that from Table 15.11, again the subplot error is less than the whole-plot error in this example. We present in Table 15.12, the table of means of the interactions between Strains and Fertilizers.

S.E. for comparing two strain means $= \sqrt{\frac{2\,\mathrm{E}_a}{8}} = 14.447$ (9 d.f.).

S.E. for the response of a single strain $= \sqrt{\frac{\mathrm{E}_b}{2}} = 18.23$ (12 d.f.).

S.E. for comparing two fertilizer means $= \sqrt{\frac{2\,\mathrm{E}_b}{2}} = 9.119$ (12 d.f.).

S.E. for comparing two fertilizers at a single strain level $= \sqrt{\frac{2\,\mathrm{E}_b}{4}} = 18.238$ (12 d.f.).

Table 15.12 Table of means

	Strains				Mean
Fertilizer	S23	NZ	X	Kent	(units of 10 lb/acre)
H	295.00	289.5	380.25	357.0	330.438
A	200.75	198.5	190.0	198.00	196.813
Mean	247.875	244.00	285.125	277.500	263.625

S.E. for comparing two strains responses $= \sqrt{\frac{4\,\mathrm{E}_b}{4}} = 25.792$ (12 d.f.)

S.E. for the average response $= \sqrt{\frac{2\,\mathrm{E}_b}{16}} = 9.119$ (12 d.f.)

S.E. for comparing two strains, at a single fertilizer level =

$$= \sqrt{\frac{2(\mathrm{E}_b + \mathrm{E}_a)}{4 \times 2}} = 19.366$$

Note that in general, the formula for the last S.E. is

$$\sqrt{\frac{2[(b-1)\,\mathrm{E}_b + \mathrm{E}_a]}{r \times b}}$$

where r is the number of blocks (replicates), and b is the number of split-plot treatments. Here, $b = 1$ in this example.

Because both error mean squares are involved in this S.E., no exact t test is possible.

(Not all of these comparisons are particularly meaningful in this example but the standard errors are included for reference purposes.) The above analysis is again analyzed with MINITAB with the following statements and corresponding output.

```
MTB > set c1
DATA> 8(1:4)
DATA> end
MTB > set c1
DATA> (1:4)8
DATA> end
MTB > set c2
DATA> 4(1:2)4
DATA> end
MTB > set c3
DATA> 8(1:4)
DATA> end
MTB > name c1 'strains' c2 'variety' c3 'blocks'
MTB > set c4
DATA> 299  318 284 279 247 202 171 183
DATA> 315 247 289 307 257 175 188 174
DATA> 403 439 355 324 222 170 192 176
DATA> 382 353 383 310 233 216 200 143
DATA> end
MTB > name c4 'yield'
 Row  strains  variety  blocks  yield

   1        1        1       1    299
   2        1        1       2    318
   3        1        1       3    284
   4        1        1       4    279
   5        1        2       1    247
   6        1        2       2    202
   7        1        2       3    171
   8        1        2       4    183
   9        2        1       1    315
  10        2        1       2    247
  11        2        1       3    289
  12        2        1       4    307
  13        2        2       1    257
  14        2        2       2    175
  15        2        2       3    188
  16        2        2       4    174
  17        3        1       1    403
  18        3        1       2    439
  19        3        1       3    355
  20        3        1       4    324
  21        3        2       1    222
  22        3        2       2    170
  23        3        2       3    192
  24        3        2       4    176
  25        4        1       1    382
  26        4        1       2    353
  27        4        1       3    383
  28        4        1       4    310
  29        4        2       1    233
  30        4        2       2    216
  31        4        2       3    200
  32        4        2       4    143
```

```
MTB > GLM 'yield' = strains blocks strains*blocks variety strains*variety;
SUBC>   Random 'blocks';
SUBC>   Brief 1.

General Linear Model: yield versus strains, blocks, variety


Factor     Type Levels Values
strains   fixed      4 1 2 3 4
blocks   random      4 1 2 3 4
variety   fixed      2 1 2

Analysis of Variance for yield, using Adjusted SS for Tests

Source           DF     Seq SS     Adj SS     Adj MS       F      P
strains           3    10303.8    10303.8     3434.6    4.11  0.043
blocks            3    13712.5    13712.5     4570.8    5.47  0.020
strains*blocks    9     7514.3     7514.3      834.9    1.26  0.349
variety           1   142845.1   142845.1   142845.1  214.73  0.000
strains*variety   3    14435.1    14435.1     4811.7    7.23  0.005
Error            12     7982.8     7982.8      665.2
Total            31   196793.5
```

Note that the *Random* statement in the MINITAB program ensures that the appropriate error mean square is used for testing the effect of the main-plot treatment. Because the interaction of strains and variety is significant (p value $= 0.005$), we present this interaction plot from this analysis is presented in Fig. 15.3. The plot is generated from the interaction Table of means in Table 15.12

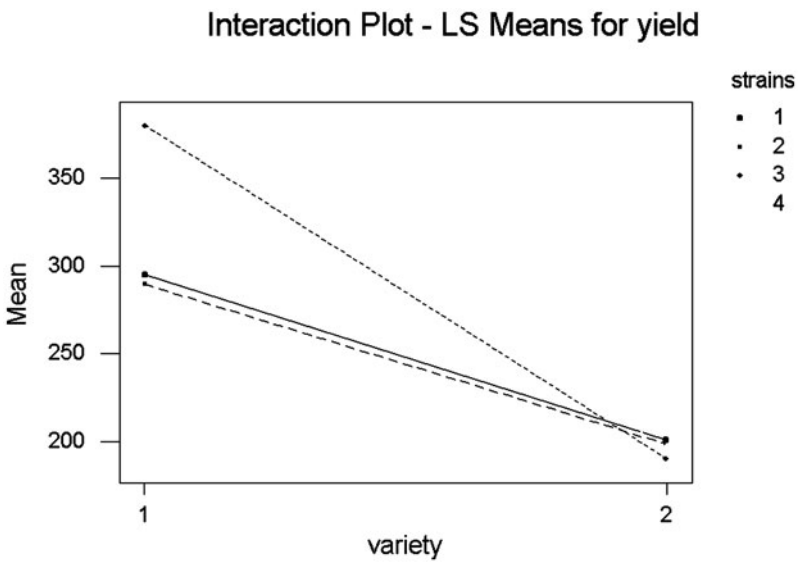


Fig. 15.3 Interaction plots

15.1.2 Summary of the Results of the Experiment

At the high fertilizer level, X and Kent significantly out-yielded S23 and New Zealand. At the average fertilizer level, there were no significant differences in yield between the four strains. All four strains showed a statistically significant response to fertilizer; the average response of X and Kent being significantly greater than that of S23 and New Zealand by 810 lb/acre with a standard error $\sqrt{\frac{4\,\mathrm{E}_b}{8}} = 10 \times 18.23 = 128$ lb/acre.

15.1.3 Missing Data in Split-Plot Design

If for some reasons (death of animal, crop destruction, natural disasters, or any other reason) data on an observation are missing, we can use the formula below to compute the missing value:

$$\hat{Y} = \frac{rM_0 + bT_0 - P_e}{(r-1)(b-1)} \tag{15.2}$$

where:

- $\hat{Y}$ is the estimated missing value.
- M_0 is the total observed values of the specific main plot that contains the missing data.
- T_0 is the total observed value of the treatment combination that contains the missing data.
- P_e is the total observed values of the main-plot treatment that contains the missing plot.
- r is the number of replications.
- b is the level of the subplot factors.

The table below gives the standard errors for the split-plot design with a missing plot value.

Comparison	Measured as	Standard error
Two main plot A means	$a_i - a_j$	$\sqrt{\frac{2(E_a + fE_b)}{rb}}$
Two subplot B means	$b_i - b_j$	$\sqrt{\frac{2E_b(1+\frac{fb}{a})}{ra}}$
Two B means at the same main plot means	$a_ib_j - a_ib_k$	$\sqrt{\frac{2E_b(1+\frac{fb}{a})}{r}}$
Two A means at		
(a) the same level of B	$a_ib_j - a_kb_j$	$\sqrt{\frac{2E_a+2E_b[(b-1)+fb^2]}{rb}}$
(b) different levels of B	$a_ib_j - a_kb_l$	ditto

where for a single missing plot,

$$f = \frac{1}{2(r-1)(b-1)}$$

15.2 Latin Square Split-Plot Example

This example is adapted from SAS. It relates to sugar beet yield in six varieties, labeled (1)–(6) in Table 15.13. The main plot represents the varieties while the subplots are the two harvesting times. The experiment is laid out as a Latin square design. The structure of the analysis of variance for this analysis is presented in Table 15.14.

To implement the analysis of the data in Table 15.13, the MINITAB GLM procedure could not handle this type of analysis correctly. So we have done the analysis in SAS which has capabilities for handling these type of data. The results from the SAS implementation are presented in the following:

```
    Latin Square Split-Plot Design
          The GLM Procedure

       Class Level Information

Class         Levels    Values
```

Table 15.13 The 6×6 LS data for the example

		Columns					
Harvest	Rows	I	II	III	IV	V	VI
1	1	(3) 19.1	(6) 18.3	(5) 19.6	(1) 18.6	(2) 18.2	(4) 18.5
1	2	(6) 18.1	(2) 19.5	(4) 17.6	(3) 18.7	(1) 18.7	(5) 19.9
1	3	(1) 18.1	(5) 20.2	(6) 18.5	(4) 20.1	(3) 18.6	(2) 19.2
1	4	(2) 19.1	(3) 18.8	(1) 18.7	(5) 20.2	(4) 18.6	(6) 18.5
1	5	(4) 17.5	(1) 18.1	(2) 18.7	(6) 18.2	(5) 20.4	(3) 18.5
1	6	(5) 17.7	(4) 17.8	(3) 17.4	(2) 17.0	(6) 17.6	(1) 17.6
2	1	(3) 16.2	(6) 17.0	(5) 18.1	(1) 16.6	(2) 17.7	(4) 16.3
2	2	(6) 16.0	(2) 15.3	(4) 16.0	(3) 17.1	(1) 16.5	(5) 17.6
2	3	(1) 16.5	(5) 18.1	(6) 16.7	(4) 16.2	(3) 16.7	(2) 17.3
2	4	(2) 17.5	(3) 16.0	(1) 16.4	(5) 18.0	(4) 16.6	(6) 16.1
2	5	(4) 15.7	(1) 16.1	(2) 16.7	(6) 16.3	(5) 17.8	(3) 16.2
2	6	(5) 18.3	(4) 16.6	(3) 16.4	(2) 17.6	(6) 17.1	(1) 16.5

Table 15.14 The degrees of freedom under the split-plot model arranged as an LS

Source	d.f.	This example
Between whole plot	(t^2-1)	35
Rows	$t-1$	5
Columns	$t-1$	5
A	$t-1$	5
Main-plot error (a)	$(t-1)(t-2)$	20
Within subplots	$t^2(b-1)$	36
B	$(b-1)$	1
AB	$(t-1)(b-1)$	5
Subplot error (b)	$t(t-1)(b-1)$	30
Total	(bt^2-1)	71

```
Column            6    1 2 3 4 5 6
Rep               6    1 2 3 4 5 6
Variety           6    1 2 3 4 5 6
Harvest           2    1 2

    Number of observations    72
         The GLM Procedure

Dependent Variable: Y
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	41	91.1973611	2.2243259	4.40	<.0001
Error	30	15.1658333	0.5055278		
Corrected Total	71	106.3631944			

R-Square	Coeff Var	Root MSE	Y Mean
0.857415	4.019184	0.711005	17.69028

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Variety	5	20.61902778	4.12380556	8.16	<.0001
Rep	5	4.32069444	0.86413889	1.71	0.1629
Column	5	1.57402778	0.31480556	0.62	0.6836
Column*Rep*Variety	20	3.25444444	0.16272222	0.32	0.9948
Harvest	1	60.68347222	60.68347222	120.04	<.0001
Variety*Harvest	5	0.74569444	0.14913889	0.30	0.9119

```
         The GLM Procedure

Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: Y
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Variety	5	20.619028	4.123806	25.34	<.0001
Rep	5	4.320694	0.864139	5.31	0.0029
Column	5	1.574028	0.314806	1.93	0.1333
Error	20	3.254444	0.162722		

```
Error: MS(Column*Rep*Variety)
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Column*Rep*Variety	20	3.254444	0.162722	0.32	0.9948
Harvest	1	60.683472	60.683472	120.04	<.0001
Variety*Harvest	5	0.745694	0.149139	0.30	0.9119
Error: MS(Error)	30	15.165833	0.505528		

```
MTB > GLM 'Y' = Varty  Col   Rep( Col)  harvest harvest* Varty;
SUBC>   Random 'Rep';
SUBC>   Brief 2 .

General Linear Model: Y versus Varty, Col, harvest, Rep

Factor    Type    Levels  Values
Varty     fixed        6  1, 2, 3, 4, 5, 6
Col       fixed        6  1, 2, 3, 4, 5, 6
Rep(Col)  random      36  1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6,
                          1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6
harvest   fixed        2  1, 2
```

```
Analysis of Variance for Y, using Adjusted SS for Tests

               Model  Reduced
Source            DF       DF     Seq SS
Varty              5        5    20.6190
Col                5        5     1.5740
Rep(Col)          30       25+    7.5751
harvest            1        1    60.6835
Varty*harvest      5        5     0.7457
Error             25       30    15.1658
Total             71       71   106.3632

+ Rank deficiency due to empty cells, unbalanced nesting, collinearity, or an
    undeclared covariate. No storage of results or further analysis will be
    done.


S = 0.711005   R-Sq = 85.74%   R-Sq(adj) = 66.25%
```

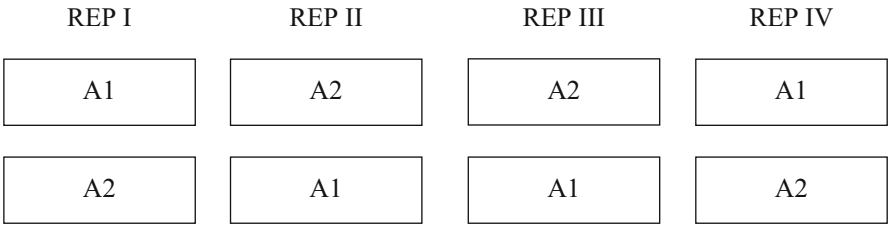
15.3 The Split-Split-Plot Design

Several variations of the split-plot design are available. One of these is the split-split-plot design which is used when we have more than two factors in our study. Here split-plot is further divided to accommodate the third factor. The design, therefore, allows for three different precision levels for each factor. The order of precision is

Main-effect treatment $\Longrightarrow$ Split-plot treatment $\Longrightarrow$ Split-split-plot treatment

with main effect having the lowest and the split-split-plot treatment having the highest precision. In this design, each level of factor A is assigned at random to r whole plots. A total of ra whole plots are therefore required. The b levels of factor B are then assigned at random to the subplots (the subplots are obtained by dividing each of the whole plots into b subplots) within each whole plot and each subplot is again divided into c subsubplots. The c levels of factor C are then assigned randomly to each of the subsubplots. In this design, factor A serves as whole plot, factor B serves as subplot, and factor C serves as the subsubplot. We present a typical layout of this design for a $2 \times 2 \times 2$ treatment combination where $a = 2, b = 2, c = 2$, and the design is replicated $r = 4$ times.

(i) First we divide the experimental area into four replications with each replication into two main plots to take factor level A A_1 and A_2 and duly randomized.



(ii) We now subdivide each main plot into two subplots (since factor B has two levels) and again assign randomly the two levels of factor B (B_1 and B_2) within the main plots. This is presented in the figure below.

REP I		REP II		REP III		REP IV	
A_1B_1	A_1B_2	A_2B_2	A_2B_1	A_2B_1	A_2B_2	A_1B_2	A_1B_1
A_2B_1	A_2B_2	A_1B_2	A_1B_1	A_1B_1	A_1B_2	A_2B_2	A_2B_1

(iii) We now divide each of the subplots into two subsubplots (the number of levels of factor C) and assign again randomly the two levels C_1 and C_2 of factor C (again here factor C has two levels) to the subsubplots. A possible arrangement is again presented in the figure below.

REP I		REP II		REP III		REP IV	
$A_1B_1C_2$	$A_1B_2C_1$	$A_2B_2C_1$	$A_2B_1C_2$	$A_2B_1C_1$	$A_2B_2C_2$	$A_1B_2C_2$	$A_1B_1C_1$
$A_1B_1C_1$	$A_1B_2C_2$	$A_2B_2C_2$	$A_2B_1C_1$	$A_2B_1C_2$	$A_2B_2C_1$	$A_1B_2C_2$	$A_1B_1C_1$
$A_2B_1C_2$	$A_2B_2C_1$	$A_1B_2C_1$	$A_1B_1C_2$	$A_1B_1C_1$	$A_1B_2C_2$	$A_2B_2C_2$	$A_2B_1C_1$
$A_2B_1C_1$	$A_2B_2C_2$	$A_1B_2C_2$	$A_1B_1C_1$	$A_1B_1C_2$	$A_1B_2C_1$	$A_2B_2C_1$	$A_2B_1C_2$

The structure of the analysis of variance table for this example is presented in the table below.

Source of variation	d.f.
Reps	3
A	1
Main-plot Error	3
B	1
AB	1
Subplot Error	6
C	1
AC	1
BC	1
ABC	1
Subsubplot error	12
Total	31

For the general split-split-plot design, the linear model formulation is presented in (15.3).

$$Y_{ijkm} = \mu + \alpha_i + e_{m(i)} + \beta_j + \alpha\beta_{ij} + e'_{m(ij)} + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \epsilon_{ijkm} \quad (15.3)$$

The structure of the analysis of variance table is presented in Table 15.15. The main effects of A in expression (15.3) will be tested with the main-plot error mean square ($e_{m(i)}$). The B and AB interaction effects will be tested with the subplot error mean square ($e'_{m(ij)}$) in (15.3), while the C, AC, BC, and ABC effects are tested with the subsubplot error mean square, namely (ϵ_{ijkm}) also in expression (15.3).

Table 15.15 The degrees of freedom under the split-split-plot model

Source	d.f.
Between whole plots	$ra - 1$
A	$a - 1$
Main-plot error	$a(r - 1)$
Within subplots	$ra(b - 1)$
B	$(b - 1)$
AB	$(a - 1)(b - 1)$
Subplot error	$a(r - 1)(b - 1)$
Within subsubplots	$rab(c - 1)$
C	$c - 1$
AC	$(a - 1)(c - 1)$
BC	$(b - 1)(c - 1)$
ABC	$(a - 1)(b - 1)(c - 1)$
Subsubplot error	$ab(r - 1)(c - 1)$
Total	$abcr - 1$

Example 15.3.1

A study conducted at Samaru to determine the influence of plant density and hybrids on corn (*Zea mays* L.) yield. The experiment was a $2 \times 2 \times 3$ factorial replicated four times in a randomized complete block design arranged in a split-split-plot layout. In this experiment, factor A is the two corn hybrids (P3730 and B70 $\times$ LH55) assigned to the main plots, factor B is the two row spacings (12 and 25 in) assigned to the subplots, and factor C is the three target plant densities (12,000, 16,000, and 20,000 plants/acre) assigned to the subsubplots. The data from the experiment are displayed in Table 15.16.

Table 15.16 Yield of two corn hybrids with two row spacings and three plant densities

Hybrid	Row spacing (in)	Plant density (plants/acre)	Grain yields (bushels/acre) Replications I	II	III	IV
P3730	12	12,000	140	138	130	142
		16,000	145	146	150	147
		20,000	150	149	146	150
			435	433	426	439
	25	12,000	136	132	134	138
		16,000	140	134	136	140
		20,000	145	138	138	142
			421	404	408	420
B70 × LH55	12	12,000	142	132	128	140
		16,000	146	136	140	141
		20,000	148	140	142	140
			436	408	410	421
	25	12,000	132	130	136	134
		16,000	138	132	130	132
		20,000	140	134	130	136
			410	396	396	402

We present below the analysis of the above data in MINITAB. The data are read into columns C1–C5.

```
MTB > set c4
DATA> 12(1:4)
DATA> end
MTB > set c3
DATA> 4(1:3)4
DATA> end
MTB > set c2
DATA> 2(1:2)12
DATA> end
MTB > set c1
DATA> (1:2)24
DATA> end
MTB > set c5
DATA> 140 138 130 142 145 146 150 147
DATA> 150 149 146 150 136 132 134 138
DATA> 140 134 136 140 145 138 138 142
DATA> 142 132 128 140 146 136 140 141
DATA> 148 140 142 140 132 130 136 134
DATA> 138 132 130 132 140 134 130 136
```

```
DATA> end
MTB > print c1-c5

Data Display

 Row     H     S     D   REP      Y

   1     1     1     1     1    140
   2     1     1     1     2    138
   3     1     1     1     3    130
   4     1     1     1     4    142
   5     1     1     2     1    145
   6     1     1     2     2    146
   7     1     1     2     3    150
   8     1     1     2     4    147
   9     1     1     3     1    150
  10     1     1     3     2    149
  11     1     1     3     3    146
  12     1     1     3     4    150
  13     1     2     1     1    136
  14     1     2     1     2    132
  15     1     2     1     3    134
  16     1     2     1     4    138
  17     1     2     2     1    140
  18     1     2     2     2    134
  19     1     2     2     3    136
  20     1     2     2     4    140
  21     1     2     3     1    145
  22     1     2     3     2    138
  23     1     2     3     3    138
  24     1     2     3     4    142
  25     2     1     1     1    142
  26     2     1     1     2    132
  27     2     1     1     3    128
  28     2     1     1     4    140
  29     2     1     2     1    146
  30     2     1     2     2    136
  31     2     1     2     3    140
  32     2     1     2     4    141
  33     2     1     3     1    148
  34     2     1     3     2    140
  35     2     1     3     3    142
  36     2     1     3     4    140
  37     2     2     1     1    132
  38     2     2     1     2    130
  39     2     2     1     3    136
  40     2     2     1     4    134
  41     2     2     2     1    138
  42     2     2     2     2    132
  43     2     2     2     3    130
  44     2     2     2     4    132
  45     2     2     3     1    140
  46     2     2     3     2    134
  47     2     2     3     3    130
  48     2     2     3     4    136
```

```
MTB > GLM 'Y' = H REP(H) S H*S S*REP(H) D H*D S*D h*s*d;
SUBC>   Random 'REP';
SUBC>   SSquares 1;
SUBC>   Brief 2 .

General Linear Model: Y versus H, S, D, REP


Factor     Type Levels Values
H         fixed      2 1 2
REP(H)   random      8 1 2 3 4 1 2 3 4
S         fixed      2 1 2
D         fixed      3 1 2 3

Analysis of Variance for Y, using Sequential SS for Tests

 Source     DF     Seq SS     Adj SS     Seq MS       F      P
 H           1    238.521    238.521    238.521    5.27  0.061
 REP(H)      6    271.625    271.625     45.271    6.83  0.017
 S           1    475.021    475.021    475.021   71.63  0.000
 H*S         1      1.687      1.687      1.687    0.25  0.632
 S*REP(H)    6     39.792     39.792      6.632    0.75  0.612
 D           2    350.042    350.042    175.021   19.92  0.000
 H*D         2     37.042     37.042     18.521    2.11  0.143
 S*D         2     87.792     87.792     43.896    5.00  0.015
 H*S*D       2      1.625      1.625      0.812    0.09  0.912
 Error      24    210.833    210.833      8.785
 Total      47   1713.979
```

The analysis obtained is in conformity with the structure of ANOVA table presented in Table 15.15. We see that the S*D interaction is significant, as well as the S effect. Often times the 6 d.f. for the REP(H) SS can be partitioned into two components representing REP SS and REP*H SS. The main effect of A can then be tested with the REP*H mean square. This is presented in the output below with H now seemingly important or significant.

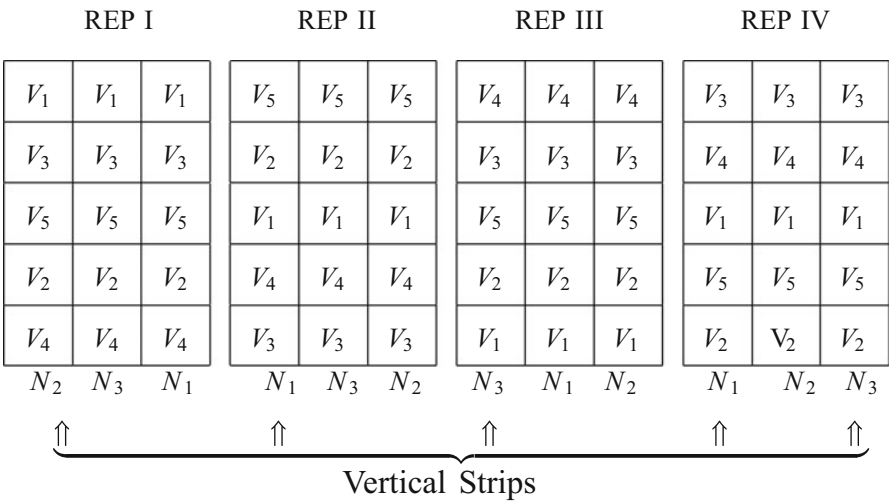
```
MTB > GLM 'Y' = REP H REP*H S H*S REP*S(H) D H*D S*D H*S*D;
SUBC>   Random 'REP' ;
SUBC>   Brief 2 .

General Linear Model: Y versus REP, H, S, D


Factor     Type Levels Values
REP      random      4 1 2 3 4
H         fixed      2 1 2
S         fixed      2 1 2
D         fixed      3 1 2 3

Source     DF     Seq SS     Adj SS     Adj MS       F      P
REP         3    237.729    237.729     79.243   22.51  0.514 x
H           1    238.521    238.521    238.521   21.11  0.019
REP*H       3     33.896     33.896     11.299    1.24  0.316
```

In addition to the standard errors given for treatment comparisons for the split-plot design discussed earlier (we need to divide by additional factor $\sqrt{c}$, though), the following additional standard errors for comparisons involving the split-split unit with c levels are split-split presented below (Cochran and Cox 1957, p. 305).

Treatment comparison	Standard error
$c_i - c_j$	$\sqrt{\frac{2E_c}{rab}}$
$a_i c_j - a_i c_k$	$\sqrt{\frac{2E_c}{rb}}$
$b_i c_j - b_i c_k$	$\sqrt{\frac{2E_c}{ra}}$
$a_i b_j c_k - a_i b_j c_l$	$\sqrt{\frac{2E_c}{r}}$
$b_i c_k - b_j c_k$ or $b_i c_k - b_j c_l$	$\sqrt{\frac{2[(c-1)E_c+E_b]}{rac}}$
$a_i b_j c_l - a_i b_k c_l$	$\sqrt{\frac{2[(c-1)E_c+E_b]}{rc}}$
$a_i c_k - a_j c_k$ or $a_i c_k - a_j c_l$	$\sqrt{\frac{2[(c-1)E_c+E_a]}{rbc}}$
$a_i b_k c_l - a_j b_k c_l$	$\sqrt{\frac{2[b(c-1)E_c+(b-1)E_b+E_a]}{rbc}}$

15.4 The Strip-Plot Design

These designs are also called split-block designs and are suitable for two-factor experiments, where factor A is applied to whole plots like the usual split-plot designs but factor B is also applied to strips which are actually a new set of whole plots orthogonal to the original plots used for factor A (that is perpendicular to each other). The scheme below gives the layout of this design. Here we have five varieties V_1, V_2, V_3, V_4, and V_5 ($a = 5$) as horizontal strips in four replicates ($r = 4$). Notice that the varieties are randomized within each replicate horizontally. Next, we divide each replicate into three vertical strips (the number of levels of the second factor). Here, we assume that we have three nitrogen rates N_1, N_2, and N_3 ($b = 3$). These treatments are again randomly assigned to give the scheme displayed below. Therefore, we have for this design, the *horizontal strip plot*, the *vertical strip plot*, and the *intersection plot*. The latter accommodates the interaction between the two factors.



The linear statistical model for this two factor design is:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \gamma_k + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + \varepsilon_{ijk} \begin{cases} i = 1, 2, \ldots, r \\ j = 1, 2, \ldots, a \\ k = 1, 2, \ldots, b \end{cases}$$

where, $(\tau\beta)_{ij}$, $(\tau\gamma)_{ik}$, and ε_{ijk} are the errors used to test factor A, factor B and interaction AB, respectively. Table 15.17 shows the structure of the analysis of variance assuming A and B to be fixed and blocks or replicates to be random.

Table 15.17 The degrees of freedom under the strip-plot model

Source	d.f.
Replications	$r - 1 = 3$
Horizontal factor (A)	$a - 1 = 4$
Error (a)	$(r - 1)(a - 1) = 12$
Vertical factor (B)	$b - 1 = 2$
B	$(b - 1)$
Error (B)	$(r - 1)(b - 1) = 6$
A × B	$(a - 1)(b - 1) = 8$
Error (c)	$(r - 1)(a - 1)(b - 1) = 24$
Total	$rab - 1 = 59$

Example

This example is from Hosmond (1993) and is data on four soft red winter wheat cultivars (Arthur 71, Auburn, Caldwell, and Compton) and grown with four nitrogen rates in a strip-plot design with four replicates. The data are presented in Table 15.18.

Table 15.18 Yield for the strip-plot experiment

Cultivar	Nitrogen rate (lb/acre)	Grain yield (bushels/acre)			
		Rep I	Rep II	Rep III	Rep IV
Arthur 71	40	72	74	76	70
	80	76	75	74	78
	120	72	74	73	75
	160	74	76	82	86
Auburn	40	60	62	64	65
	80	61	63	69	68
	120	70	72	69	70
	160	72	70	82	86
Caldwell	40	75	73	72	80
	80	77	78	77	82
	120	80	82	86	88
	160	84	82	84	89
Compton	40	65	68	63	72
	80	68	72	74	76
	120	69	68	70	72
	160	73	75	74	76

The analysis of the data in Table 15.18 using MINITAB is displayed below. The computed p values indicate that there are significantly different mean yields among the cultivars (p value $= 0.000$). Similarly, there are significant differences in the mean yields of nitrogen rates (p value $= 0.000$). Our analysis also indicates that the interaction term between the cultivars and nitrogen rates is highly significant (p value $= 0.0006$). All the p values are $<< 0.05$.

```
MTB > print c1-c4

Data Display

Row  A  B  REP   Y
  1  1  1    1  72
  2  1  1    2  74
  3  1  1    3  76
  4  1  1    4  70
  5  1  2    1  76
  6  1  2    2  75
  7  1  2    3  74
  8  1  2    4  78
  9  1  3    1  72
 10  1  3    2  74
.................
.................
 55  4  2    3  74
 56  4  2    4  76
 57  4  3    1  69
 58  4  3    2  68
 59  4  3    3  70
 60  4  3    4  72
 61  4  4    1  73
 62  4  4    2  75
 63  4  4    3  74
 64  4  4    4  76
```

```
MTB > GLM 'Y' = REP A REP*A B REP*B  A*B;
SUBC>   Random 'REP';
SUBC>   Brief 2 .

General Linear Model: Y versus REP, A, B

Factor  Type   Levels  Values
REP     random      4  1, 2, 3, 4
A       fixed       4  1, 2, 3, 4
B       fixed       4  1, 2, 3, 4

Analysis of Variance for Y, using Adjusted SS for Tests

Source  DF    Seq SS    Adj SS   Adj MS      F      P
REP      3   257.562   257.562   85.854  15.45  0.037 x
A        3  1282.187  1282.187  427.396  76.64  0.000
REP*A    9    50.188    50.188    5.576   0.72  0.682
B        3   761.313   761.313  253.771  33.07  0.000
REP*B    9    69.062    69.062    7.674   1.00  0.465
A*B      9   237.937   237.937   26.437   3.44  0.006
Error   27   207.687   207.687    7.692
Total   63  2865.938

x Not an exact F-test.


S = 2.77347   R-Sq = 92.75%   R-Sq(adj) = 83.09%
1    75.44
2    68.94
3    80.56
4    70.94
B
1    69.44
2    73.00
3    74.38
4    79.06
A*B    1     2     3     4
1    73.00 75.75 73.50 79.50
2    62.75 65.25 70.25 77.50
3    75.00 78.50 84.00 84.75
4    67.00 72.50 69.75 74.50
```

Now let us compute the coefficients of variation corresponding to each of the three error mean squares. These are (here, G is the grand mean):

$$cv(a) = \frac{\sqrt{E_a}}{\mathrm{G}} \times 100 = \frac{\sqrt{5.576}}{73.96} = 3.19\,\%$$

$$cv(b) = \frac{\sqrt{E_b}}{\mathrm{G}} \times 100 = \frac{\sqrt{7.674}}{73.96} = 3.75\,\%$$

$$cv(c) = \frac{\sqrt{E_c}}{\mathrm{G}} \times 100 = \frac{\sqrt{7.692}}{73.96} = 3.75\,\%$$

The $cv(a)$ value for instance indicates the degree of precision associated with the horizontal factor, $cv(b)$ with the vertical factor, and $cv(c)$ with the interaction between the two factors. $cv(c)$ is expected to be the smallest, and hence, highest precision. There is no rule as to which of the other two cv's should be higher or lower.

The standard errors for treatment comparisons in a strip-plot design are displayed in Table 15.19.

Table 15.19 Standard errors for the subplot strips

Treatment comparison	Standard error
(i) $a_i - a_j$	$\sqrt{\frac{2E_a}{rb}}$
(ii) $b_i - b_j$	$\sqrt{\frac{2E_b}{ra}}$
(iii) $a_i b_k - a_j b_k$	$\sqrt{\frac{2[(b-1)E_c+E_a]}{rb}}$
(iv) $a_i b_j - a_i b_k$	$\sqrt{\frac{2[(a-1)E_c+E_b]}{ra}}$

The last two standard errors in Table 15.19 are respectively:

(a) The standard error of difference between two main-plot (A) level means at the same level of B means.
(b) The standard error of difference between two strip-plot (B) level means at the same level of A means.

To obtain confidence intervals or conduct tests of significance based on the above standard errors, the Student's t critical value (say at 5 %) is obtained by multiplying the S.E. for (i) and (ii) by the respective $t_{.}05$ obtained for the appropriate d.f. However, for (iii) and (iv), the S.E. of mean differences involve two error terms, and we would use the following expressions to respectively compute the weighted Student's t values:
For case (iii), we have:

$$t = \frac{(b-1)E_c t_c + E_a t_a}{(b-1)E_c + E_a}$$

Similarly, for case (iv), we have:

$$t = \frac{(a-1)E_c t_c + E_b t_b}{(a-1)E_c + E_b}$$

where t_a, t_b, and t_c are t values at error d.f. for E_a, E_b, and E_c respectively.

15.5 Exercises

1. This example is taken from Steel and Torrie (1960), and is hereby duly acknowledged. An experiment compared the yields of four lots of oats for three chemical seed treatments and untreated check. The seed lots are Vicland(1), Vicland(2), with Vicland(1) infected with *H. victoriae* and Vicland(2) uninfected. The other two seed lots are Clinton and Branch oats which are resistant to *H. victoriae*. The seed lots are the main factor and assigned randomly to blocks, while the seed protectants, factor B, were assigned at random to the subplots within each whole plots. The experiment was laid out a CRBD of four blocks. The yield in bushels per acre are presented below.

Factor A		Factor B			
Seed lot	Blocks	Check	Ceresan M	Panogen	Agrox
Vicland(1)	1	42.9	53.8	49.5	44.4
	2	41.6	58.5	53.8	41.8
	3	28.9	43.9	40.7	28.3
	4	30.8	46.3	39.4	34.7
Vicland(2)	1	53.3	57.6	59.8	64.1
	2	69.6	69.6	65.8	57.4
	3	45.4	42.4	41.4	44.1
	4	35.1	51.9	45.4	51.6
Clinton	1	62.3	63.4	64.5	63.6
	2	58.5	50.4	46.1	56.1
	3	44.6	45.0	62.6	52.7
	4	50.3	46.7	50.3	51.8
Branch	1	75.4	70.3	68.8	71.6
	2	65.6	67.3	65.3	69.4
	3	54.0	57.6	45.6	56.6
	4	52.7	58.5	51.0	47.4

Yields of oats in bushels per acre

   (a) Carry out a split-plot analysis for the above data and draw your conclusions.
   (b) Calculate the appropriate standard errors for the difference between (i) two main-plot means and (ii) to subplot means.

2. In an experiment, three formulations of a diet (factor A) were compared on the basis of a certain chemical being absorbed in the diet by the kidneys of experimental rats. The researcher is also interested in comparing three techniques (factor B) for measuring the absorbed amounts. Four litters, each containing three rats were used in the study. Within each litter, the animals were randomly assigned to the three diets. After two weeks on the diets, the animals were sacrificed and three sample specimens were selected from each animal's kidney. The three methods were randomly assigned to

the three specimens and the absorbed amount of the chemical measured. The data for the experiment is presented below (Source: Rao 1998).

		Litter			
Diet	Method	1	2	3	4
	1	26.97	26.12	27.83	27.47
1	2	22.60	22.91	19.83	21.63
	3	30.71	29.53	27.51	28.62
	1	17.47	18.13	18.01	17.97
2	2	16.90	16.31	16.52	15.93
	3	23.95	22.84	23.84	23.45
	1	20.72	20.41	21.01	21.34
3	2	24.32	25.06	25.92	25.33
	3	28.31	29.02	29.13	29.36

Absorption data for diet experiment

Analyze the data as split-plot design and draw your conclusions.

3. An environmental horticulturist is interested in finding out whether (1) stress-adapted landscapes save water (2) whether irrigation equal to 15 % or less reference evapotranspiration (ET_0) can be applied to established shrubs and ground cover without any drought-related injury. The study is a three-factor experiment with three irrigation regimes (no irrigation, 12.0 in, and 24.0 in of water), two different irrigation methods (drip and furrow) on the growth of shrubs and ground covers such as *Xylosma*, oleander, *Cotoneaster*, juniper, ice plant, and *Hedera*. The experiment is a strip-split-plot design replicated three times. The data collected over a 2-year period are presented below.

	Irrigation	Water	Growth (in)		
Plantings	method	applied (in)	Rep I	Rep II	Rep III
Xylosma	Drip	0.0	8.0	8.4	9.5
		12.0	19.5	20.1	20.2
		24.0	30.6	31.0	31.4
	Furrow	0.0	6.0	5.4	5.8
		12.0	12.8	16.9	17.4
		24.0	28.2	27.6	29.4
Oleander	Drip	0.0	18.0	19.4	19.5
		12.0	39.5	40.1	40.3
		24.0	60.6	59.0	61.4
	Furrow	0.0	16.0	15.4	15.7
		12.0	22.8	36.9	37.4
		24.0	48.2	47.6	49.4
Coton easter	Drip	0.0	6.0	6.4	6.5
		12.0	35.5	31.1	30.6
		24.0	40.6	41.0	41.3
	Furrow	0.0	4.0	4.4	4.8
		12.0	19.8	16.9	18.4
		24.0	25.2	27.6	29.5

Plantings	Irrigation method	Water applied (in)	Growth (in) Rep I	Rep II	Rep III
Juniper	Drip	0.0	12.0	12.4	12.7
		12.0	10.5	10.1	10.2
		24.0	20.6	18.0	19.3
	Furrow	0.0	10.0	11.1	10.8
		12.0	9.8	6.9	7.4
		24.0	13.2	14.8	15.4
Ice plant	Drip	0.0	22.0	18.8	19.5
		12.0	26.5	28.1	27.2
		24.0	33.6	33.0	32.4
	Furrow	0.0	16.0	15.4	15.8
		12.0	22.5	26.9	25.4
		24	28.8	29.6	29.9
Hedera	Drip	0.0	16.0	17.4	18.5
		12.0	20.5	22.1	20.7
		24.0	40.6	41.0	41.4
	Furrow	0.0	12.0	13.2	14.6
		12.0	15.8	14.9	14.6
		24.0	28.2	27.6	29.4

Growth of shrubs and ground cover as a function of irrigation water received from April to August (Source: Hosmond 2004).

(a) Analyze the data as a strip-split plot experiment.
(b) What conclusions can you draw from the analysis?
(c) Compute the coefficient of variation for the factors.

4. A split-plot experiment in a randomized complete block design evaluated the effects of nitrogen, water, and phosphorous rates on the water use efficiency in a drip irrigation culture of sweet corn. Two rates of phosphorus ($P_1 = 0$ and $P_2 = 245$ lb P_2O_5/acre) were randomized to whole plots in a randomized complete block design. The 3×3 factorial treatments of nitrogen (0, 130, and 260 lb N/acre) and water (16, 22, and 28 in) were randomized to subplots within each of the main plots. The data presented below give the water use efficiency for each subplot (Source: Dr. T. Doerge, Department of Soil and Water Science, University of Arizona).

Water	Nitrogen	Block I P_1	Block I P_2	Block II P_1	Block II P_2
16	0	8.1	9.7	8.6	15.5
	130	36.0	34.2	34.5	33.1
	260	34.6	34.0	40.7	39.3
22	0	10.0	6.2	5.1	10.9
	130	21.5	19.7	19.9	21.9
	260	30.7	28.9	26.4	25.7
28	0	10.6	6.3	4.5	10.4
	130	19.4	19.7	21.7	19.9
	260	23.2	23.0	19.4	23.2

(a) Conduct the analysis of variance for the data using MINITAB.
(b) Construct a summary table of means and marginal means for each factor and interactions.
(c) Compute the estimated standard errors for comparing two different levels of:

- Phosphorous rates
- Water rates
- Nitrogen rates

(d) Test the hypotheses for all interaction and main effect terms and interpret your results.
(e) Partition the sum of squares for water and nitrogen into linear and quadratic terms including their interaction. Interpret your results and plot a graph of the observed means to assist.

5. A split-plot experiment was conducted in a completely randomized design with whole-plot treatments as a 2×2 factorial (factors A and B) and the subplot treatments as three levels of factor C. There were four replications of the experimental units.

(a) Outline the analysis of variance table showing the sources of variation and degrees of freedom.
(b) Suppose the above experiment was carried out as a Latin square design. Repeat part (a) above for this case.

Chapter 16
Incomplete Block Design

16.1 Introduction

The designs considered in the previous chapters, namely, randomized complete block and Latin square design assume that each block always contain enough experimental units to allow for each treatment (or treatment combination in case of a factorial design) to be contained at least once in each block or in the case of Latin square design in each row or column. In particular, when the number of treatments equals the number of units in a block, the design is very very simple and the analysis becomes straightforward. However, when the number of units in a block is less (in some cases could be more) than the number of treatments, the design is no longer simple and so does the analysis.

This sometimes becomes necessary under certain circumstances when, for instance, we do not have enough homogeneous units to form a block with the required number of treatments. This situation can be resolved by the use of incomplete block designs, which are not too difficult to construct and can readily be analyzed with MINITAB or SAS. We saw in Chap. 14 the construction of a design for the 2^4 factorial in blocks of 8. Since there are 16 treatment combinations in this experiment, ideally, we should have blocks of size 16 to have this as a randomized complete block or a 16×16 Latin square design. For the randomized complete block design (RCBD), this makes it difficult to have homogeneous units within each block (this would be very difficult in field experiment for instance where sloping, water logging in some parts of the field may make this impossible). A possible layout of a simple replicate for instance can be in blocks of eight or even in blocks of four.

a	b	c	abc	d	abd	acd	bcd

ab	(1)	bc	ac	bd	ad	abcd	cd

In the above design, we have used blocks of eight units. This is therefore a very simple example of an incomplete block design. It is incomplete because we are not able to apply all the 16 treatments in every block. In order words,

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_16,

an incomplete block design is simply one in which there are more treatments that can be put in a single block. That is, it is not possible to include all factor combinations in every block. In the above, we recognize that the fourth order interaction ABCD has been confounded with blocks.

Example 16.1.1

As an example, consider again, the 2^4 factorial which is to be laid out in blocks of size 4. This calls for four blocks in a single replicate with complete confounding of 3 of the 15 effects. Suppose we number the 16 treatment combinations, 1–16 in the order presented in Chap. 14, then we have:

1	4	13	16

Block 1

2	3	14	15

Block 2

5	8	9	12

Block 3

6	7	10	11

Block 4

The above is a single unrandomized replicate of an incomplete block design with $t = 16$ and $k = 4$. Several replicates of this basic design can be repeated depending on whether we want complete or partial confounding. In each case, however, apart from randomization of treatments within blocks and blocks within replicates, same group of treatments occur together in every block. We consider in the next section, balanced incomplete block design where pairs of treatments occur together the same number of times in the experiment. Indeed, we shall focus only on balanced designs in this chapter.

In general, suppose we have b blocks of k plots each, and t treatments each replicated r times, then

$$N = bk = tr. \tag{16.1}$$

The design is incomplete because

1. $k < t$. That is, the number of plots in each block is less than the number of treatments.
2. No treatment occurs more than once in any block.

For any two distinct treatments i and j, the *concurrence* λ_{ij} of i and j is the number of blocks which contain both i and j. As an example, consider the case of a design with $t = 6,\ r = 2,\ b = 4$, and $k = 3$. Then, the four blocks are:

A	B	C

Block 1

B	D	F

Block 2

A	D	E

Block 3

C	E	F

Block 4

Here, $\lambda_{12} = \lambda_{13} = 1$ and $\lambda_{16} = 0$. Thus, in general, for unbalanced incomplete block design, λ_{ij} are not all equal.

16.2 Balanced Incomplete Block Design

A balanced incomplete block (BIB) design is an incomplete block design in which every pair of treatments occurs the same number of times in the experiment. This guarantees equal precision on all pairwise comparisons among treatments. The BIB design has the following properties:

1. Each block contains the same number ($= k$) of treatments. k is the block size.
2. Each treatment occurs the same number of times ($= r$) in the entire experiment. Here, r is the number of replications.
3. Each pair of treatments occurs the same number of times λ in each block and appears as many times as any other pair of treatments in the design. That is, $\lambda_{ij} = \lambda$ for all i and j.

Notation

- t = the number of treatments
- b = the number of blocks in the experiment
- k = the number of units/plots per block, that is, the block size
- r = the number of replicates of a given treatment throughout the experiment

A BIB design is therefore specified by its parameters (t, b, r, k, λ). For a BIB design with parameters t and k, then parameters b, r, and λ are obtained from the following expressions.

$$b = \frac{t!}{k!(t-k)!} \tag{16.2a}$$

$$r = \frac{(t-1)!}{(k-1)!(t-k)!} \tag{16.2b}$$

$$\lambda = \frac{(t-2)!}{(k-2)!(t-k)!} \tag{16.2c}$$

where $a!$ is read "a" factorial.
If there is a common factor between b, r, and λ, they will be divided by this common factor to obtain a *reduced design*. As an example, consider the case when we have $t = 6$ treatments in blocks of $k = 3$ units. Then, we would have

$$b = \frac{6!}{3!3!} = 20$$

$$r = \frac{5!}{2!3!} = 10$$

$$\lambda = \frac{4!}{1!3!} = 4$$

b, r, and λ have 2 as a common factor; hence, a reduced model can be obtained with parameters $b = 10$, $r = 5$, and $\lambda = 2$. That is, we have BIB design with parameters $(6, 10, 5, 3, 2)$. In general, for a BIB design with parameters (t, b, r, k, λ), the following relationship holds (from 16.2b, c).

$$tr = bk = N, \text{the number of observations in the experiment}$$
$$\lambda = \frac{r(k-1)}{t-1}.$$

Most often, a BIB design (usually designated as BIBD) is commonly written as simply (t, k, λ), where b and r are given in terms of t, k, and λ by (also from 16.2b, c)

$$b = \frac{t(t-1)}{k(k-1)}$$
$$r = \frac{\lambda(t-1)}{k-1}.$$

16.3 Statistical Model for a Balanced Incomplete Block (BIB) Design

$$y_{ij} = \mu + t_i + b_j + e_{ij} \tag{16.3}$$

where

$y_{ij} \sim$ is the i-th observation in the j-th block
$\mu \sim$ grand mean
$t_i \sim$ effect of the i-th treatment
$b_j \sim$ is the effect of the j-th block
$e_{ij} \sim$ is the NID$(0, \sigma^2)$ random error term

Analysis

The analysis of a BIB design is based on obtaining the following SS:

$$\text{SS}_{Total} = \sum\sum y_{ij}^2 - \frac{y_{++}^2}{N} \quad \text{that is,}$$
$$\text{SS}_T = \text{SS}_{Trt(adjusted)} + \text{SS}_{blocks} + \text{SSE}$$

Because each treatment is represented in a different set of r blocks, the adjustment is necessary to extract the treatment effect from blocks.

$$\text{Blocks SS} = \frac{1}{k}\sum y_{+j}^2 - \frac{y_{++}^2}{N}. \tag{16.4}$$

The adjusted treatment SS

$$\text{SS}_{Trt(adj.)} = \frac{k}{\lambda t}\sum_{i=1}^{t} Q_i^2 \tag{16.5}$$

where Q_i is the adjusted total for the ith treatment, where

$$Q_i = y_{i+} - \frac{1}{k}\sum_{j=1}^{k} n_{ij} y_{+j} \tag{16.6}$$

with

$$n_{ij} = \begin{cases} 1 & \text{if treatment } i \text{ appears in block } j \\ 0 & \text{if treatment } i \text{ does not appear in block } j \end{cases}$$

Note that $\sum Q_i = 0$.

Example 16.3.1

Consider an experiment with $t = 4$ treatments, but only 3 units/block are available. Such an experiment could be an animal experiment involving four feeds but, while we have four, three, five, and four litters from each of four mothers. It would make sense to use only three litters from each mother (for orthogonality point of few). Hence, we are guaranteed of homogeneity between litters from the same mother and below is the result of weights of the animals from the experiment. The data for this example are presented in Table 16.1.

Table 16.1 Data for this example

	Blocks				
Trt	1	2	3	4	Total
1	73	74	–	71	218
2	–	75	67	72	214
3	73	75	68	–	216
4	75	–	72	75	222
Total	221	224	207	218	870

In this experiment, $t = 4$, $b = 4$, $k = 3$, $r = 3$, and $\lambda = 2 = 3(2)/3$ and $N = bk = tr = 12$. That is, the experiment used a total of 12 experimental units, hence

$$\text{Total SS} = 73^2 + 74^2 + \cdots + 75^2 - \frac{870^2}{12} = 81.000$$

$$\text{Blocks SS} = \frac{221^2}{3} + \frac{224^2}{3} + \frac{207^2}{3} + \frac{218^2}{3} - \frac{870^2}{12} = 55.000.$$

To obtain the adjusted treatments SS, first we note that:

$$y_{1+} = 218, \quad y_{2+} = 214, \quad y_{3+} = 216, \quad y_{4+} = 222$$
$$y_{+1} = 221, \quad y_{+2} = 224, \quad y_{+3} = 207, \quad y_{+4} = 218$$

Hence,

$$\begin{aligned} Q_1 &= 218 - \frac{1}{3}(n_{11}y_{+1} + n_{12}y_{+2} + n_{14}y_{+4}) \\ &= 218 - \frac{1}{3}(221 + 224 + 218) \\ &= -3 \end{aligned}$$

Since treatment (trt 1) appears in blocks 1, 2, and 4 only, hence, $n_{1j} = 1$ for $j = 1, 2, 4$ and 0 for $j = 3$. Similarly,

$$Q_2 = 214 - \frac{1}{3}(224 + 207 + 218) = -2.3333$$
$$Q_3 = 216 - \frac{1}{3}(221 + 224 + 207) = -1.3333$$
$$Q_4 = 222 - \frac{1}{3}(221 + 207 + 218) = 6.6666$$

$$\begin{aligned} \text{Trt SS}_{adj.} &= \frac{k}{\lambda t}\sum_{i=1}^{t} Q_i^2 \\ &= \frac{3}{2 \times 4}\left[(-3)^2 + (-2.3333)^2 + (-1.3333)^2 + (6.6666)^2\right] \\ &= (0.375)(60.6655) \\ &= 22.7496 \end{aligned}$$

The analysis of variance table is therefore given by:

Source	d.f.	SS	MS	F
Blocks	3	55.0000	18.3333	
$\text{Trt}_{adj.}$	3	22.7496	7.5832	11.665
Error	5	3.2504	0.6501	
Total	11	81.0000		

Parameter estimates of the treatment effects are computed from:

$$\hat{t}_i = \frac{kQ_i}{\lambda t} \tag{16.7}$$

Thus,

$$\hat{t}_1 = \frac{-3 \times 3}{2 \times 4} = -1.125$$
$$\hat{t}_2 = \frac{-2.3333 \times 3}{2 \times 4} = -0.875$$
$$\hat{t}_3 = \frac{-1.3333 \times 3}{2 \times 4} = -0.500$$
$$\hat{t}_4 = \frac{6.6666 \times 3}{2 \times 4} = 2.500$$

We present below the MINITAB implementation of the above analysis.

```
MTB > SET C1
DATA> (1:4)3
DATA> END
MTB > SET C2
DATA> 1 2 4 2 3 4
DATA> 1 2 3 1 3 4
DATA> END
MTB > SET C3
DATA> 73 74 71 75 67 72
DATA> 73 75 68 75 72 75
DATA> END
MTB > NAME C1 'TRT' C2 'BLOCKS' C3 'Y'

MTB > Print 'TRT' 'BLOCKS' 'Y'.

Data Display

 Row   TRT  BLOCKS      Y

   1     1       1     73
   2     1       2     74
   3     1       4     71
   4     2       2     75
   5     2       3     67
   6     2       4     72
   7     3       1     73
   8     3       2     75
   9     3       3     68
  10     4       1     75
  11     4       3     72
  12     4       4     75

MTB > GLM 'Y' = BLOCKS TRT;
SUBC>   Brief 1 ;
SUBC>   Means TRT.

General Linear Model: Y versus BLOCKS, TRT

Factor     Type Levels Values
BLOCKS    fixed      4 1 2 3 4
TRT       fixed      4 1 2 3 4
```

```
Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS     Adj MS       F      P
BLOCKS      3     55.000     66.083     22.028   33.89  0.001
TRT         3     22.750     22.750      7.583   11.67  0.011
Error       5      3.250      3.250      0.650
Total      11     81.000

Least Squares Means for Y

TRT      Mean   SE Mean
1       71.38    0.4868
2       71.63    0.4868
3       72.00    0.4868
4       75.00    0.4868

Tukey Simultaneous Tests
Response Variable Y
All Pairwise Comparisons among Levels of TRT

TRT = 1 subtracted from:

Level  Difference       SE of             Adjusted
TRT      of Means  Difference   T-Value    P-Value
2          0.2500      0.6982    0.3581     0.9825
3          0.6250      0.6982    0.8951     0.8085
4          3.6250      0.6982    5.1918     0.0130

TRT = 2 subtracted from:

Level  Difference       SE of             Adjusted
TRT      of Means  Difference   T-Value    P-Value
3          0.3750      0.6982    0.5371     0.9462
4          3.3750      0.6982    4.8338     0.0175

TRT = 3 subtracted from:

Level  Difference       SE of             Adjusted
TRT      of Means  Difference   T-Value    P-Value
4             3.000      0.6982     4.297     0.0281
```

From the p value of 0.011 given in the MINITAB output, we can conclude that there are significant differences in the adjusted means of the four treatments. The standard error of difference between two treatment means is

$$\sqrt{\frac{2kS^2}{\lambda t}} = \sqrt{2 \times 3 \times 0.6501/2 \times 4} = 0.6983$$

and the standard error of a treatment means is

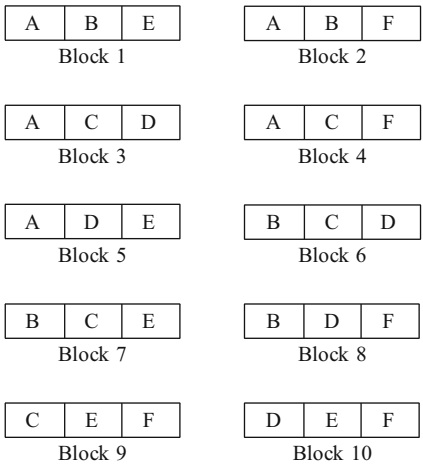
$$\sqrt{\frac{kS^2}{\lambda t}} = \sqrt{3 \times 0.6501/2 \times 4} = 0.4938.$$

Based on Tukey's test, we would say that treatments 1, 2, and 3 are not significantly different but each of them is significantly different from treatment 4, which has the highest least squares treatment mean of 75.0. This result is displayed below.

4 3 2 1

16.4 Constructing a Balanced Incomplete Block (BIB) Design

For our example involving $t = 6,\ k = 3,\ r = 5,\ b = 10$, and $\lambda = 2$ we can construct a BIB design for this reduced design with the following layout.



It is noteworthy to observe here that a BIB design may not exist for some combinations of the five parameters mentioned above. Further, construction of a BIB design can be very cumbersome and may not easily be achieved. Fortunately, tables of balanced incomplete block designs have been presented in Cochran and Cox (1957), Davies (1957), and Fisher and Yates (1953). For a given set of parameters, the chosen design must be followed by randomly assigning the treatments to blocks and using a separate randomization for each block.

In the event that we decide to use the unreduced BIB design for $t = 6$ treatments in blocks of three units, then we would require $b = 20,\ r = 10$, and $\lambda = 4$. Thus, a BIB design with $(t, k, \lambda) = (6, 3, 4)$ is displayed below.

A	B	C	D	E	F	A	B	D	C	E	F
Block 1			Block 2			Block 3			Block 4		
A	B	E	C	D	F	A	B	F	C	D	E
Block 5			Block 6			Block 7			Block 8		
A	C	D	B	E	F	A	C	E	B	D	F
Block 9			Block 10			Block 11			Block 12		
A	C	F	B	D	E	A	D	E	B	C	F
Block 13			Block 14			Block 15			Block 16		
A	D	F	B	C	E	A	E	F	B	C	D
Block 17			Block 18			Block 19			Block 20		

Again, randomization of treatments within blocks and of blocks must be carried out. In the above layout, blocks 1 and 2 can be considered as first replicate, blocks $(3, 4)$ as second replicate,$\cdots$,$(19, 20)$ blocks as the tenth replicate. Notice that each replicate contains all the treatments of interest.

16.5 Efficiency of Incomplete Block Designs

The efficiency of one design relative to another design is measured by the ratio of the variances for comparing two treatment means in both designs. Thus, the variance for two treatment means in a randomized complete block design (RCBD) is $2\sigma^2_{rcb}/r$. Similarly, the corresponding variance in a BIB design is $2k\sigma^2_{bib}/\lambda t$. Thus BIB and RCBD have the same number of treatments and replications, then, the efficiency of the BIB design relative to the RCBD is given by

$$\text{Relative efficiency} = \frac{(2\sigma^2_{rcb}/r)}{2k\sigma^2_{bib}/\lambda t} = \frac{\sigma^2_{rcb}}{\sigma^2_{bib}} \times \frac{\lambda t}{rk}. \tag{16.8}$$

The quantity $E = \frac{\lambda t}{rk}$ in expression (16.8) is often referred to as the *efficiency factor*. Thus, a BIB design is more precise for comparing two treatment means than the RCBD if

$$\frac{\sigma^2_{rcb}}{\sigma^2_{bib}} < \frac{\lambda t}{rk}. \tag{16.9}$$

Thus, for a BIB design with $t = 6$, $r = 5$, $k = 3$, $b = 10$, and $\lambda = 2$, the efficiency factor would be

$$\frac{2(6)}{5(3)} = 0.8.$$

That is, σ^2_{bib} has to be about 20 % smaller than σ^2_{rcb} for the BIB design to be as precise as the RCBD with the same number of treatments and replications. Thus, the aim of a BIB design would be to reduce σ^2_{bib}, and thus increase the precision for comparing two treatment means. A good BIB design, therefore, is one such that σ^2_{bib} would be reduced or as small as possible relative to the σ^2_{rcb} such that the inequality in (16.9) would be achieved.

16.6 Lattice Design

Sometimes, the number of treatments t can be very large in an experiment. Most often, experiments in plant breeding usually call for a large number of treatments. In such cases, a completely randomized design of t units would not be suitable because of heterogeneous within-blocks variations, which might result in inconsistent parameter estimates. As observed in the previous section, we could have the treatments arranged in balanced incomplete blocks.

A design which has the characteristics that the number of treatments must be a perfect square, that is, $t = k^2$, and that the block sizes must be k, the square root of t. Further, the number of replications is $(k + 1)$, that is, $r = (k+1)$ is called a *balanced lattice design*. Thus, the number of treatments must be 9, 16, 25, 36, 49, 81, etc. We shall discuss the randomization and layout of this design in the next section.

16.6.1 Construction of Lattice Design

For $t = k^2$ treatments, there would be $(k + 1)$ replicates, with each replicate consisting of k blocks and each block containing k units. The construction of lattice designs is based on the theory of orthogonal Latin squares. Suppose we have $t = 16$ treatments. This calls for $k = \sqrt{16} = 4$ Latin squares, and there would be $(k - 1) = 3$ orthogonal Latin squares, except for $k = 6$, 10, and 12 since these are not perfect squares. A typical layout for $t = 16$ is presented below:

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

1	2	3	4
3	4	1	2
4	3	2	1
2	1	4	3

α	β	γ	δ
δ	γ	β	α
β	α	δ	γ
γ	δ	α	β

For $k = 4$, we would need to generate $k(k+1) = 20$ blocks of size 4, thus, each replication will be divided into k incomplete blocks of size k. In our example, there would be four incomplete blocks each containing four experimental units. To construct these 20 blocks, first we start with the Latin letters. Let

the treatments be designated $1, 2, \cdots, 16$. Thus, filling in the 4×4 cells with these treatments, we have:

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

From the four rows of the above layout, we generate blocks 1–4 as follows:

Block 1	1	2	3	4
Block 2	5	6	7	8
Block 3	9	10	11	12
Block 4	13	14	15	16

From the columns, we can also generate blocks 5–8, viz.:

Block 5	1	5	9	13
Block 6	2	6	10	14
Block 7	3	7	11	15
Block 8	4	8	12	16

To generate blocks 9–12, we need to superimpose the treatments numbered on the first orthogonal square (Latin letters) and we would, thus, have

A 1	B 2	C 3	D 4
B 5	A 6	D 7	C 8
C 9	D 10	A 11	B 12
D 13	C 14	B 15	A 16

Now generate blocks 9–12 with the treatments, viz.:

Block 9	1	6	11	16	for treatment appearing with letter A
Block 10	2	5	12	15	for treatment appearing with letter B
Block 11	3	8	9	14	for treatment appearing with letter C
Block 12	4	7	10	13	for treatment appearing with letter D

To generate blocks 13–16, next we superimpose the treatments on the numerals in the second orthogonal 4×4 square. This leads to

1_1	2_2	3_3	4_4
3_5	4_6	1_7	2_8
4_9	3_{10}	2_{11}	1_{12}
2_{13}	1_{14}	4_{15}	3_{16}

Again, extracting the treatments to form blocks 13–16, we have

Block 13	1	7	12	14	for treatment appearing with numeral 1
Block 14	2	8	11	13	for treatment appearing with numeral 2
Block 15	3	5	10	16	for treatment appearing with numeral 3
Block 16	4	6	9	15	for treatment appearing with numeral 4

Finally, to generate blocks 17–20, again from the third orthogonal square (Greek letters), we can superimpose the treatments, again leading to:

α	1	β	2	γ	3	δ	4
δ	5	γ	6	β	7	α	8
β	9	α	10	δ	11	γ	12
γ	13	δ	14	α	15	β	16

Extracting the treatments to form blocks 17–20, we have

Block 17	1	8	10	15	for treatment appearing with α
Block 18	2	7	9	16	for treatment appearing with β
Block 19	3	6	12	13	for treatment appearing with γ
Block 20	4	5	11	14	for treatment appearing with δ

If we were to randomize the above layout, then the replicates, blocks, and treatments may look like the arrangement in Table 16.2. Notice that replicate I, for instance, represents the arrangement based on the first orthogonal square, and further, the treatments have been randomized within each block, and the blocks have also been similarly randomized within replicate. That is, this is a balanced lattice design with parameters $t = 16$, $k = 4$, $r = 5$, $b = 20$, and $\lambda = 1$.

Table 16.2 Balanced lattice design with $t = 16$, $k = 4$, $r = 5$, $b = 20$, and $\lambda = 1$

Block	Rep. I				Block	Rep. II			
(1)	11	1	16	6	(5)	3	13	6	12
(2)	7	4	10	13	(6)	11	4	14	5
(3)	2	15	12	5	(7)	10	1	15	8
(4)	9	8	14	3	(8)	9	16	7	2

Block	Rep. III				Block	Rep. IV			
(9)	6	8	7	5	(13)	6	14	10	2
(10)	12	10	9	11	(14)	16	8	4	12
(11)	3	1	4	2	(15)	9	5	13	1
(12)	16	14	15	13	(16)	11	7	15	3

Block	Rep. V			
(17)	5	16	10	3
(18)	14	12	1	7
(19)	9	4	15	6
(20)	11	2	13	8

We present in Table 16.3 the structure of the analysis of variance table for the balanced lattice design, viz.:

Table 16.3 Structure of ANOVA in balanced lattice design

Source	d.f.
Replication	k
Treatments (unadj.)	$k^2 - 1$
Block (adj.)	$k^2 - 1$
Intrablock error	$(k-1)(k^2-1)$
Treatments (adj.)	$k^2 - 2$
Effective error	$(k-1)(k^2-1)$
Total	$(k+1)k^2 - 1$

We may observe here that the balanced lattice design is balanced incomplete block design (BIBD) with, $t = k^2$, $b = k(k+1)$, $r = k+1$, and $\lambda = 1$.

Example 16.6.1

Plant breeders are interested in determining the spikelet initiation differences among nine winter wheat cultivars. The number of spikelets per plant from a field experiment which followed a 3×3 balanced lattice design with four replications is given below. Each cultivar is given a treatment number and they are: Turkey (1), Pawnee (2), Scout (3), Larned (4), Newton (5), Hawk (6), Vona (7), HW (8), and Bounty 100 (9). The collected data from each replicate are presented below:

Block	Rep I			Block	Rep II		
1	18.1 (8)	18.4 (6)	17.6 (1)	4	18.2 (8)	20.2 (7)	16.5 (9)
2	16.5 (3)	18.7 (5)	17.9 (7)	5	15.2 (3)	19.9 (2)	17.8 (1)
3	16.0 (4)	18.0 (2)	16.0 (9)	6	17.8 (6)	18.1 (5)	16.4 (4)

Block	Rep III			Block	Rep IV		
7	17.1 (8)	18.4 (5)	18.6 (2)	10	16.2 (3)	15.9 (4)	18.5 (8)
8	16.2 (4)	17.7 (7)	16.9 (1)	11	17.2 (6)	18.9 (2)	17.6 (7)
9	16.5 (3)	18.9 (6)	16.2 (9)	12	15.4 (9)	18.9 (5)	17.4 (1)

Analysis

1. Calculate block totals (B), replicate totals (R), treatment totals (T), and the grand total (G)

Reps. Totals			
1	2	3	4
157.2	160.1	156.5	156.0

Blocks Totals (B_j)											
1	2	3	4	5	6	7	8	9	10	11	12
54.1	53.1	50.0	54.9	52.9	52.3	54.1	50.8	51.6	50.6	53.7	51.7

and,

Treatment Totals (T_j)									Grand
1	2	3	4	5	6	7	8	9	Total (G)
69.7	75.4	64.4	64.5	74.1	72.3	73.4	71.9	64.1	629.8

2. Next, we compute the block totals over all blocks $B_j^{'}$ in which a particular treatment appears. For example,

$B_1^{'} = 54.1 + 52.9 + 50.8 + 51.7 = 209.5$ (these are from blocks 1, 5, 8, and 12)

$\vdots = \quad \vdots \quad \vdots \quad \vdots$

$B_9^{'} = 50.0 + 54.9 + 51.6 + 51.7 = 208.2$ (these are from blocks 3, 4, 9, and 12)

1. Now calculate

$$Q_j = kT_j - (k+1)B_j^{'} + G, \quad \text{that is,}$$
$$Q_j = 3T_j - 4B_j^{'} + G.$$

For instance,

$$Q_1 = (3 \times 69.7) - (4 \times 209.5) + 629.8 = 0.9.$$

Similar calculations lead to results presented in the next table.

trt (j)	T_j	$B_j^{'}$	Q_j
1	69.7	209.5	0.9
2	75.4	210.7	13.2
3	64.4	208.2	−9.8
4	64.5	203.7	8.5
5	74.1	211.2	7.3
6	72.3	211.7	−0.10
7	73.4	212.5	0.0
8	71.9	213.7	−9.3
9	64.1	208.2	−10.7
Total	629.8	1889.40	0

2. The correction factor (CF) is computed as:

$$\text{CF} = \frac{(629.8)^2}{k^2(k+1)} = \frac{(629.8)^2}{36} = 11,018.0011$$

$$\text{Total SS} = 18.1^2 + \cdots + 17.4^2 - \text{CF} = 52.7189$$

$$\text{Replicate SS} = \frac{157.2^2}{9} + \cdots + \frac{156^2}{9} - \text{CF} = 1.1211$$

$$\text{Treatments (unadj.) SS} = \frac{69.7^2}{4} + \cdots + \frac{64.1^2}{4} - \text{CF} = 40.7339$$

and,

$$\text{Blocks (adj.) SS} = \frac{\sum Q_j^2}{k^3(k+1)} - \text{C.F.} = \frac{0.9^2}{108} + \cdots + \frac{10.7^2}{108} - \text{CF}$$
$$= 5.5335$$

Hence,

$$\text{Intrablocks Error SS} = \text{Total SS} - \text{Reps SS} - \text{Trt SS} - \text{Blocks (adj.) SS}$$
$$= 5.3304.$$

The above leads to the analysis of variance table based on intrablock analysis.

Source	d.f.	SS	MS	F
Reps	3	1.1211		
Treatments (unadj.)	8	40.7339	5.0917	7.499
Blocks (adj.)	8	5.5335	0.6917	
Intrablock	16	5.3304	0.3332	
Total	35	52.7189		

3. Since the intrablock mean square is 0.3332, which is less than the adjusted block mean square of 0.6917, it is, therefore, imperative to compute adjusted treatment totals, and hence, adjusted treatment means. To do this, we compute the following:

$$\mathrm{T}_j^{'} = T_j + \mu Q_j \tag{16.10}$$

where

$$\mu = \frac{\text{Block}(adj.)\text{ MS} - \text{Intrablock Error MS}}{k^2[\text{Block(adj.) MS}]} = \frac{E_b - E_e}{k^2\ E_b} \tag{16.11}$$

where E_b and E_e are the adjusted block and intrablock error mean squares, respectively. Thus, for our data,

$$\mu = \frac{0.6917 - 0.3332}{9(0.6917)} = 0.0576$$

and the adjusted treatment totals, say for treatment 1, is computed as

$$T_1^{'} = 69.7 + 0.0576(0.9) = 69.7518, \quad \text{hence adj. } \bar{T}'_{adj} = 17.4380.$$

Continuing with the calculations, we have the following adjusted treatment totals and means:

Treatment (j)	Totals (adj.)	Means (adj.)
1	69.7518	17.4380
2	76.1604	19.0401
3	63.8356	15.9589
4	64.9896	16.2474
5	74.5204	18.6301
6	72.2944	18.0736
7	73.4000	18.3500
8	71.3644	17.8411
9	63.4836	15.8709

We can now calculate the adjusted Treatment SS as

$$\text{Treatment}_{adj.}\ \text{SS} = \frac{1}{(k+1)} \sum T_j^{'2} - \text{C}$$

$$\text{Treatments (adj.) SS} = \frac{69.7518^2}{4} + \cdots + \frac{63.4836^2}{4} - \frac{G^2}{36} = 45.6808$$

$$\text{Treatment (adj.) MS} = \frac{45.6808}{8} = 5.5335$$

and

$$\text{The effective error MS} = (\text{intrablock error MS})\ (1 + k\mu) = 0.3908$$

Hence,

$$F = \frac{\text{Treatment (adj.) MS}}{\text{Effective error MS}} = \frac{5.5335}{0.3908} = 14.611.$$

The computed F is highly significant. If the intrablock error MS had been greater than the block (adj.) MS, the value of μ computed earlier would have been negative and would, therefore, be assumed to be zero. In this case, the adjustment analysis above would not be necessary and would have based our F test on the first ANOVA table, that was based on unadjusted treatment SS. We present a MINITAB analysis for the data. The results obtained agree with those obtained from our calculations.

```
MTB > GLM 'y' = Rep block( Rep) trt;
SUBC>   Brief 2 ;
SUBC>   Means trt;
SUBC>   Pairwise trt;
SUBC>     Tukey;
SUBC>     NoTest;
SUBC>     NoCI.

General Linear Model: y versus Rep, trt, block

Factor      Type   Levels  Values

Rep        fixed        4  1, 2, 3, 4
block(Rep) fixed       12  1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3
trt        fixed        9  1, 2, 3, 4, 5, 6, 7, 8, 9

Analysis of Variance for y, using Adjusted SS for Tests

Source      DF   Seq SS  Adj SS   Adj MS      F      P
Rep          3   1.1211  1.1211   0.3737   1.12  0.370
block(Rep)   8   7.9044  5.5335   0.6917   2.08  0.102
trt          8  38.3630 38.3630   4.7954  14.39  0.000
Error       16   5.3304  5.3304   0.3331
Total       35  52.7189


S = 0.577190   R-Sq = 89.89%   R-Sq(adj) = 77.88%

Least Squares Means for y

trt   Mean  SE Mean
1    17.45   0.3286
2    19.22   0.3286
3    15.83   0.3286
4    16.36   0.3286
5    18.73   0.3286
6    18.07   0.3286
7    18.35   0.3286
8    17.72   0.3286
9    15.73   0.3286

Grouping Information Using Tukey Method and 95.0% Confidence
```

```
trt  N   Mean  Grouping
2    4  19.22  A
5    4  18.73  A B
7    4  18.35  A B
6    4  18.07  A B
8    4  17.72  A B C
1    4  17.45    B C D
4    4  16.36      C D E
3    4  15.83        D E
9    4  15.73          E

Means that do not share a letter are significantly different.
```

16.7 Relative Efficiency for Lattice Design

The relative efficiency of a balanced lattice design over that of the randomized complete block design (RCBD) is given by the following expression Cochran and Cox (1957),

$$\text{R.E.} = \frac{\text{Pooled Mean Square}}{\text{Effective error Mean Square}} \times 100 \tag{16.12}$$

where, the effective error mean square, designated as, $E_e^{'}$ is computed as

$$E_e^{'} = E_e(1 + k\mu) = 0.3332(1 + 3 \times 0.0576) = 0.3907$$

and the pooled mean square is obtained by adding the adjusted blocks SS with the intrablock SS and dividing by the appropriate degree of freedom of $k(k^2 - 1)$. Thus, in our case, the pooled MS is computed as

$$E_{\text{pooled}} = \frac{5.5335 + 5.3304}{24} = \frac{10.8639}{24} = 0.4527$$

Hence, the relative efficiency (R.E.) in this example equals $\dfrac{0.4527}{0.3907} \times 100 = 115.9\,\%$.
We see that the experimental precision has been increased by 15.9 % over that of the randomized complete block design.

Cochran and Cox (1957) listed lattice designs up to $t = 144$ treatments. We may also note here that a lattice design may also be used when the number of treatments is not a perfect square. For instance, if we have 14 treatments in an experiment, then two of the treatments might be included twice to make $t = 16$. Some of the textbooks earlier mentioned in this chapter have extensive discussions on both balanced and partially balanced lattice designs and similar incomplete block designs.

16.8 Exercises

1. Construct a balanced lattice design with $t = 9$.
2. An incomplete block design consists of the following arrangement of blocks (1, 2, 3, 4, 5) and treatments (A, B, C, D, E).

Block		
	1	(B, C, D, E)
	2	(A, B, D, E)
	3	(A, C, D, E)
	4	(A, B, C, D)
	5	(A, B, C, E)

   (a) What are the design parameters t, r, k, and b?
   (b) Verify that the design is balanced.

3. A horticulturalist studied the germination of tomato seed with four different temperatures (25 °C, 30 °C, 35 °C, and 40 °C) in a balanced incomplete block design because there were only two growth chambers available for the study. Each run of the experiment was an incomplete block consisting of the two growth chambers as the experimental units ($k = 2$). Two experimental temperatures were randomly assigned to the chambers for each run. The data for the experiment are as presented below and give the germination rates of the tomato seed.

Run	25 °C	30 °C	35 °C	40 °C
1	24.65	–	–	1.34
2	–	24.38	–	2.24
3	29.17	21.25	–	–
4	–	–	5.90	1.83
5	28.90	–	18.27	–
6	25.53	8.42	–	

Source: Robert Kuehl (1998)

   (a) How many times did each treatment pair occur together in the same block?
   (b) What is the efficiency factor for this design?
   (c) Analyze the data and obtain the standard errors between two levels of temperatures.
   (d) Partition the treatment SS into its components.
   (e) What degree of polynomial would you recommend?

4. An experiment to examine the preferences of cabbage root flies for six different substances on which to lay their eggs involved the use of ten cages of flies with three substances available in each cage. The number of eggs laid on the various substances was as shown in the table below (Source: Mead and Curnow 1983).

	Substances					
Cage	1	2	3	4	5	6
1	452	69	83	–	–	–
2	802	143	–	53	–	–
3	699	–	32	–	4	–
4	1207	–	–	19	–	32
5	958	–	–	–	8	8
6	–	328	147	–	–	53
7	–	314	–	264	223	–
8	–	158	–	–	36	5
9	–	–	117	14	115	–
10	–	–	23	16	–	2

(a) How many times did each treatment pair occur together in the same block?
(b) What is the efficiency factor for this design?
(c) Analyze the data after a suitable transformation and obtain the standard errors between two levels of substances.

Chapter 17
Quantal Bioassay

17.1 Introduction

An assay can be described as the comparative analysis of the estimation of the strength of a drug on animals, animal tissues, etc. with that of a standard drug. An indirect qualitative assay considers groups of animals or experimental units subjected to different levels of some stimulus and the proportions of animals responding to the stimulus is observed and this proportion will be related to the level of the stimulus. A quantal bioassay involves studying the relationship between dosage and response proportions (or percentages). They are usually characterized by studies in which a dose of a drug is applied to n experimental units and r of them are observed to respond to the drug dose and hence $n - r$ of them do not respond. The main objective of a quantal bioassay is to determine what level of the dose would be necessary to bring about a response in a certain percentage of the experimental units in the population. For any individual experimental unit (say, an animal) there is a threshold below which the animal will not respond to the stimulus. In such cases, the amount of the stimulus required to bring about a minimal (tolerance) response in the animal is often referred to as the *individual effective dose* (IED). Usually, we are most interested in the level of the stimulus that would result in 50 % response in a given sample of subjects. We thus have the following classifications of this measure:

- LD_{50}: *median lethal dose*
- ED_{50}: *median effective dose*
- LC_{50}: *median lethal concentration* and
- EC_{50}: *median effective concentration*

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_17,

The model that will be employed here is the nonlinear logistic regression model. The logistic model is based on the response variable having a binary outcome, yes or no; alive or dead, cured or not cured, etc. For instance, in a bioassay experiment, an insect is classified upon the application of treatment as either dead or alive. If p_i is the proportion of insects killed at treatment level i (usually, dosage level), then, the linear logistic model assumes the form

$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}. \tag{17.1}$$

The above leads to the logit model

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i. \tag{17.2}$$

The expression in (17.2) can sometimes be written as

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_i} \tag{17.3}$$

where

- $e = 2.71828$, is the base of the natural logarithm.
- $\frac{p_i}{1-p_i}$ is the odds ratio.
- $\ln\left(\frac{p_i}{1-p_i}\right)$ is the log odds ratio or simply the logit, and
- β_0 and β_1 are the parameters of the model to be estimated.

The logistic regression in (17.2) does not assume normality of error terms or homoscedastic of error variances, and in Fig. 17.1, we have the graphs of the logistic function

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$$

for fixed $\beta_1 = 0.1$, varying $\beta_0 = -2.5, -1.5, -1.0, 2.0$ values and values of x ranging from -40 to 80. That is, for β_0 taking values $\{-2.5, -1.5, -1.0, 2.0\}$, we plot the graph of p_i against x for a constant value of β_0 at 0.1. The plot is presented in Fig. 17.1, where

$$p_i = \frac{1}{1+e^{-(\beta_0+0.1x_i)}};\quad x = -40, \ldots, 80.$$

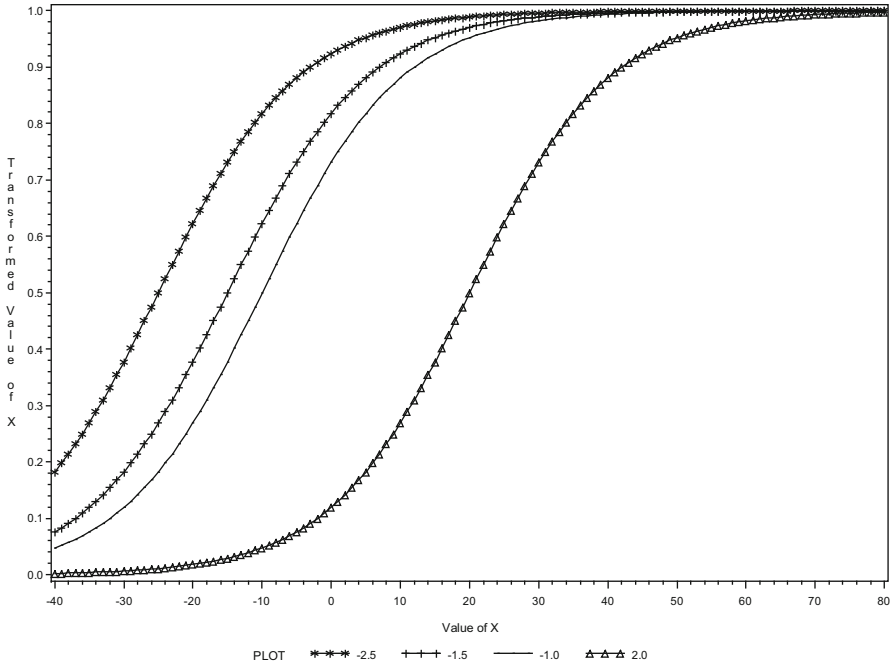


Fig. 17.1 Plot of logistic function

Let us illustrate our discussion with the following example.

Example 17.1.1: A Quantal Assay Example

The data in Table 17.1 give the effect of different concentrations of nicotine sulfate in 10 % saponin solution in *Drosophila melanogaster*.

We usually employ the logarithm of the x_i, and in this case, we chose to use $\log(x_i)$, that is, the log to base 10 of the doses. The table below gives the relevant initial calculations for the data in Table 17.1. While I do not for a moment think that the linear logistic moment should be fitted by what follows in this section, it is nevertheless incorporated here so that students can have a proper understanding of what is really going on from the use of statistical packages.

Table 17.1 Effect of different concentrations of nicotine sulfate on *Drosophila melanogaster*

Nicotine sulfate (g/100 cc) dose	Number killed r_i	Number of insects n_i
0.10	23	137
0.30	95	152
0.50	119	146
0.70	141	154
0.95	144	152

The mortality rates are computed as $p_i = r_i/n_i$. Further, we need to calculate the following:

$$\text{log-dose } x_i = \log_{10}(\text{dose})$$

$$\text{logit } y_i = \log_e\left(\frac{p_i}{1-p_i}\right)$$

$$\text{weighting coefficients } w_i = n_i\, p_i\, (1-p_i).$$

Next, we compute the following:

$$\sum w_i, \quad \sum w_i x_i, \quad \sum w_i y_i$$
$$\sum w_i x_i y_i, \quad \sum w_i x_i^2$$

and

$$S_{xy} = \sum w_i x_i y_i - \frac{(\sum w_i x_i)(\sum w_i y_i)}{\sum w_i}$$
$$S_{xx} = \sum w_i x_i^2 - \frac{(\sum w_i x_i)^2}{\sum w_i}$$

Hence,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$
$$\hat{\beta}_0 = \frac{\sum w_i y_i}{\sum w_i} - \hat{\beta}_1 \frac{\sum w_i x_i}{\sum w_i}. \tag{17.4}$$

An initial estimate of the linear logistic model can be found from solving the equations in (17.5).

$$\hat{\beta}_0 \sum w + \hat{\beta}_1 \sum wx = \sum wy \tag{17.5a}$$
$$\hat{\beta}_1 \sum wx + \hat{\beta}_1 \sum wx^2 = \sum wxy. \tag{17.5b}$$

Thus, substituting the summary statistics from Table 17.2 into the equations in (17.5), we have

$$96.252\,\hat{\beta}_0 - 46.404\,\hat{\beta}_1 = 70.485$$
$$-46.404\,\hat{\beta}_0 + 31.162\,\hat{\beta}_1 = 6.4105.$$

Table 17.2 Initial summary statistics for the data in Table 17.1

n_i	x_i	p_i	y_i	w_i	wx	wy	wxy	wx^2
137	−1.000	0.1679	−1.601	19.1387	−19.1387	−30.6354	30.6354	19.1387
152	−0.523	0.6250	0.511	35.6250	−18.6276	18.1982	−9.5154	9.7400
146	−0.301	0.8151	1.483	22.0068	−6.6247	32.6425	−9.8264	1.9942
154	−0.155	0.9156	2.389	11.9026	−1.8437	28.3735	−4.3951	0.2856
152	−0.022	0.9474	2.883	7.5789	−0.1688	21.9060	−0.4880	0.0038
Total				96.2520	−46.4040	70.485	6.4105	31.162

Solving the above, we have initial estimates of the parameters as

$$\hat{\beta}_0 = 2.9477 \quad \text{and} \quad \hat{\beta}_1 = 4.5952.$$

These initial estimates lead to a revised estimated logit $\hat{y}_{1i}$ and $\hat{p}_{1i}$ in Table 17.3. The subscript 1 refers to first-step of the computation.

Table 17.3 Step 1 summary statistics for the data in Table 17.1

n_i	x_i	$\hat{p}_{1i}$	$\hat{y}_{1i}$	w_i	wx	wy	wxy	wx^2
137	−1.000	0.1614	−1.6007	18.5473	−18.5473	−29.6888	29.6888	18.5473
152	−0.523	0.6330	0.5450	35.3126	−18.4642	18.0386	−9.4320	9.6545
146	−0.301	0.8270	1.5644	20.8898	−6.2885	30.9856	−9.3276	1.8930
154	−0.155	0.9034	2.2359	13.4360	−2.0813	32.0289	−4.9613	0.3224
152	−0.022	0.9451	2.8453	7.8898	−0.1758	22.8044	−0.5080	0.0039
Total				96.076	−45.557	74.1690	5.4599	30.4210

which again leads to the equations

$$96.076\,\hat{\beta}_0 - 45.557\,\hat{\beta}_1 = 74.169$$
$$-45.557\,\hat{\beta}_0 + 30.421\,\hat{\beta}_1 = 5.4599$$

Solving the above, we have second step estimates of the parameters as

$$\hat{\beta}_0 = 2.9570 \quad \text{and} \quad \hat{\beta}_1 = 4.6080.$$

Again, the corresponding estimated logits and proportions, $\hat{y}_{2i}$ and $\hat{p}_{2i}$ respectively, at the second step are presented in Table 17.4.

Table 17.4 Step 2 summary statistics for the data in Table 17.1

n_i	x_i	p_{2i}	$\hat{y}_{2i}$	w_i	wx	wy	wxy	wx^2
137	−1.000	0.1610	−1.6007	18.5034	−18.5034	−29.6185	29.6185	18.5034
152	−0.523	0.6336	0.5108	35.2881	−18.4514	18.0260	−9.4254	9.6478
146	−0.301	0.8278	1.4833	20.8154	−6.2661	30.8753	−9.2944	1.8863
154	−0.155	0.9041	2.3838	13.3568	−2.0690	31.8402	−4.9321	0.3205
152	−0.022	0.9455	2.8904	7.8267	−0.1744	22.6220	−0.5039	0.0039
Total				95.790	−45.464	73.745	5.4626	30.362

which again leads to the equations

$$95.790\,\hat{\beta}_0 - 45.464\,\hat{\beta}_1 = 73.745$$
$$-45.745\,\hat{\beta}_0 + 30.362\,\hat{\beta}_1 = 5.4626.$$

The above can again be repeated. We obtained

$$\hat{\beta}_0 = 2.9799 \quad \text{and} \quad \hat{\beta}_1 = 4.6364.$$

This process continues until we have convergence (that is, when the estimated values between two successive steps are very very close). The above approach is sometimes referred to as the weighted least squares method for fitting the desired model.

17.2 The Logistic Regression Approach

Let p_i be the probability that an insect will be killed with concentration x_i. The linear logistic model fits the logit of p_i, namely, $\ln\left(\frac{p_i}{1-p_i}\right)$ as the dependent function, leading to the model

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \log_{10} x_i, \quad \text{that is,}$$
$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1\, \text{dose}_i$$

where dose_i is the logarithm to base 10 of x_i, the concentrations in the ith group. In Fig. 17.2, we present the plot of the logit of p_i for values of p ranging from $0 \le p \le 1$.

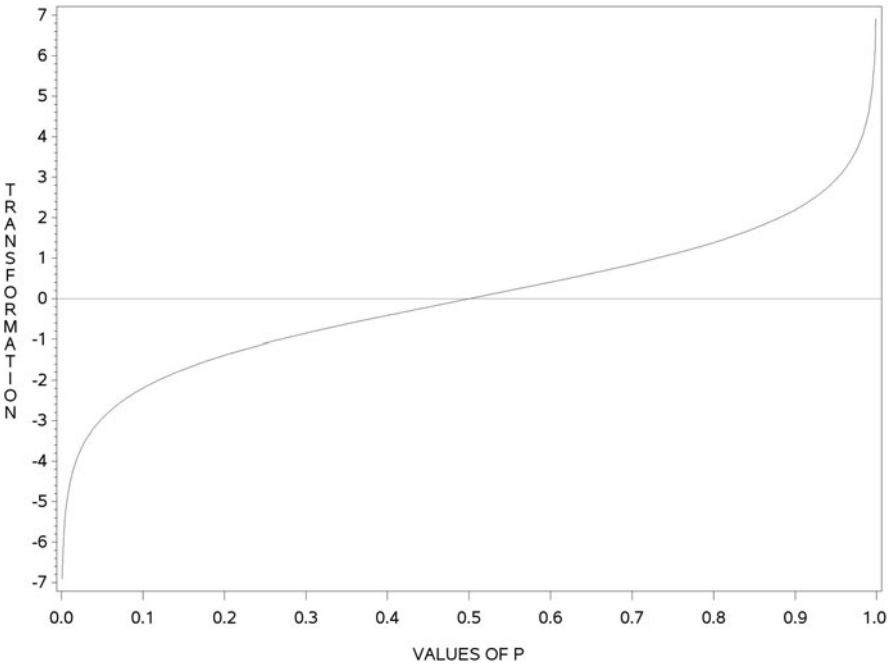


Fig. 17.2 Logit transformation plot

A MINITAB analysis of the data in Table 17.1 which utilizes the linear logistic regression provides the following results:

```
MTB > read c1-c3
DATA> .10 23 137
DATA> .3 95 152
DATA> .5 119 146
DATA> .7 141 154
DATA> .95 144 152
DATA> end

MTB > Let c4 = LOGT(c1)

MTB > Print 'dose'-'ldose'.

Data Display

 Row   dose      r      n     ldose

   1   0.10     23    137  -1.00000
   2   0.30     95    152  -0.52288
   3   0.50    119    146  -0.30103
   4   0.70    141    154  -0.15490
   5   0.95    144    152  -0.02228
```

```
MTB >

MTB > BLogistic 'r' 'n' = ldose;
SUBC>   ST;
SUBC>   Logit;
SUBC>   Brief 2.

Binary Logistic Regression: r, n versus ldose

Link Function:  Logit

Response  Information

Variable  Value         Count
r         Success         522
          Failure         219
n         Total           741

Logistic Regression Table

                                                        Odds       95% CI
Predictor       Coef    SE Coef        Z     P    Ratio   Lower    Upper
Constant      2.9519     0.1897    15.56 0.000
ldose         4.6008     0.3370    13.65 0.000    99.56   51.43   192.75

Log-Likelihood = -308.624
Test that all slopes are zero: G = 282.392, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

Method               Chi-Square    DF     P
Pearson                   0.505     3  0.918
Deviance                  0.512     3  0.916
Hosmer-Lemeshow           0.505     3  0.918

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

                   Group
Value      1     2     3     4     5   Total
Success
  Obs     23    95   119   141   144     522
  Exp   22.1  96.3 120.8 139.2 143.7
Failure
  Obs    114    57    27    13     8     219
  Exp  114.9  55.7  25.2  14.8   8.3

  Total  137   152   146   154   152     741

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs           Number  Percent    Summary Measures
Concordant       89481    78.3%    Somers' D                0.69
Discordant       10602     9.3%    Goodman-Kruskal Gamma    0.79
Ties             14235    12.5%    Kendall's Tau-a          0.29
Total           114318   100.0%
```

MINITAB gives us the parameter estimates as

$$\hat{\beta}_0 = 2.9519 \quad \text{and} \quad \hat{\beta}_1 = 4.6008.$$

The parameter estimates $\hat{\beta}_0 = 2.9519$ and $\hat{\beta}_1 = 4.6008$ are both significantly different from zero, $(p < .0001)$ in both cases.

The odds ratio for the intercept would be $e^{2.9519} = 19.14$. Thus at log (base 10) dosage level (dose = 0), it is almost 19 times most likely that the insect *D. melanogaster* would die than not die. Note that this dose level is equivalent to 1.0 g/100 cc nicotine sulfate concentration. Similarly, with each unit increase in $\log_{10}$ dosage level, the odds of insect dying increases by 99.6 times.

The estimated killing probability as a function of the log dosage is estimated as

$$\hat{p} = \frac{e^{2.9519+4.6008\text{dose}}}{1 + e^{2.9519+4.6008\text{dose}}},$$

which for the first dosage becomes

$$\begin{aligned} &= \frac{e^{2.9519+4.6008(-1)}}{1 + e^{2.9519+4.6008(-1)}} \\ &= \frac{0.1923}{1.1923} \\ &= 0.16128 \end{aligned}$$

The expected number of deaths for this dosage level is $n_i * \hat{p}_i = n_1 * \hat{p}_1 = 137 * 0.16128 = 22.0954$. These and other relevant parameters are displayed below to a better accuracy.

The expected or predicted values are computed as

$$\left[-1.6477, 0.5411, 1.5641, 2.2354, 2.8461\right]$$

with estimated proportions

$$\hat{p}_i = [0.1613, 0.6333, 0.8273, 0.9037, 0.9453].$$

These results from MINITAB are displayed below.

```
MTB > LET C8=C3*C7
MTB > PRINT C2 C3 C7 C8

Data Display
 Row       r      n     phat    expected
  ----------------------------------------
   1      23    137   0.161263    22.093
   2      95    152   0.633271    96.257
   3     119    146   0.827349   120.793
   4     141    154   0.903720   139.173
   5     144    152   0.945290   143.684
```

The leverages and deviance residuals are also presented below. These are very useful for diagnostics.

```
MTB > print c2 c5-c8

Data Display

 Row      r    resid       lev        phat        yhat
 ------------------------------------------------------
   1     23   0.209557   0.774609   0.161263     22.093
   2     95  -0.211261   0.377342   0.633271     96.257
   3    119  -0.389042   0.288564   0.827349    120.793
   4    141   0.508824   0.295026   0.903720    139.173
   5    144   0.113373   0.264459   0.945290    143.684
```

The level of the dosage which would result in a 50 % response by subjects in the population under study is an important parameter in dose–response models. A measure of the potency of the drug is the statistic LD_{50}, *median lethal dose*. In this example, LD_{50} is the lethal dosage at which 50 % of the subjects (insects) are expected to be killed, and in experiments where the response is not death, we refer to the ED_{50}, median effective dose. There is also LC_{50} (*median lethal concentration*) and EC_{50} (*median effective concentration*). In our example, the LD_{50} is computed as

$$\ln\left(\frac{50}{100-50}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_M \Rightarrow \hat{x}_M = -\frac{\hat{\beta}_0}{\hat{\beta}_1} = -\frac{2.9519}{4.6008} = -0.6416.$$

That is, $\log_{10}(\mathrm{LD}_{50}) = -0.6416 \Rightarrow \mathrm{LD}_{50} = 10^{-0.6416} = 0.2282$. That is, the $\mathrm{LD}_{50} = 0.228$ g/100 cc.

Similarly, an LD_{90} is given by

$$10^U, \quad \text{where} \quad U = \frac{(2.1972 - \hat{\beta}_0)}{\hat{\beta}_1} = 0.6854.$$

That is, the LD_{90} is 0.685 g/100 cc. This is the dose level at which we expect 90 % of the insects to be killed in the population.

The logistic model when applied to the data in Table 17.1 fits the data well with the goodness-of-fit test statistic $X^2 = 0.505$ on 3 d.f. (p value = 0.918).

A test of the hypothesis concerning the parameters of the logistic, that is, a test of whether $H_0 : \beta = 0$ against $H_a : \beta \neq 0$, is provided in the MINITAB output below.

```
Test that all slopes are zero: G = 282.392, DF = 1, P-Value =
0.000
```

Clearly, this indicates that the explanatory variable dose_i is important in the logistic model.

17.3 Using the Probit Model

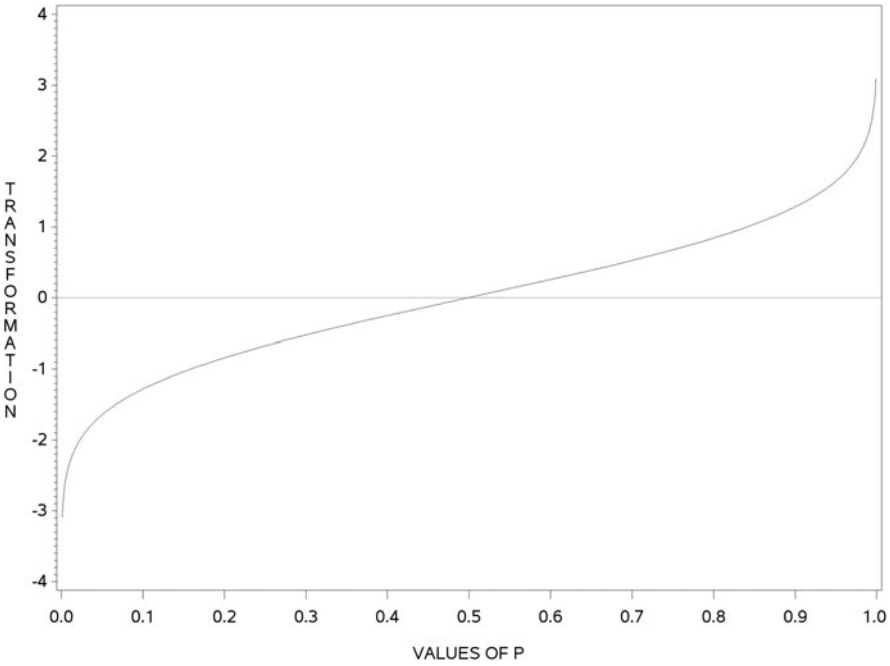


Fig. 17.3 Probit transformation plot

To fit the probit model (a typical probit plot is displayed in Fig. 17.3), we simply specify the link function as NORMIT in the MINITAB sub command statement.

```
Results for: bioassy3.MTW

MTB > BLogistic 'r' 'n' = ldose;
SUBC>   ST;
SUBC>   Normit;
SUBC>   Hi 'HI1';
SUBC>   Eprobability 'EPRO1';
SUBC>   XPWXinverse 'XPWX1';
SUBC>   Loglikelihood 'LOGL1';
SUBC>   Ghdchisquare;
SUBC>   Brief 2.

Binary Logistic Regression: r, n versus ldose


Link Function:  Normit

Response    Information

Variable    Value        Count
r           Success        522
            Failure        219
n           Total          741

Logistic Regression Table
```

```
Predictor       Coef    SE Coef        Z     P
Constant    1.72778    0.09943    17.38 0.000
ldose        2.6914     0.1791    15.03 0.000

Log-Likelihood = -308.526
Test that all slopes are zero: G = 282.587, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

Method           Chi-Square    DF      P
Pearson               0.312     3  0.958
Deviance              0.316     3  0.957
Hosmer-Lemeshow       0.312     3  0.958


Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

                Group
Value      1     2     3     4     5   Total
Success
  Obs     23    95   119   141   144     522
  Exp   23.0  95.1 119.8 139.4 144.8
Failure
  Obs    114    57    27    13     8     219
  Exp    114.0  56.9  26.2  14.6   7.2

  Total   137   152   146   154   152     741

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs           Number   Percent     Summary Measures
Concordant       89481     78.3%     Somers' D                0.69
Discordant       10602      9.3%     Goodman-Kruskal Gamma    0.79
Ties             14235     12.5%     Kendall's Tau-a          0.29
Total           114318    100.0%


 Matrix XPWX1

 0.0098865  0.0145852
 0.0145852  0.0320833

Data Display


   dose      r      n     ldose      lev       prob       yhat
   ------------------------------------------------------------
   0.10     23    137  -1.00000   0.790354   0.167619    22.964
   0.30     95    152  -0.52288   0.317444   0.625709    95.108
   0.50    119    146  -0.30103   0.272876   0.820584   119.805
   0.70    141    154  -0.15490   0.313976   0.905051   139.378
   0.95    144    152  -0.02228   0.305350   0.952325   144.753
```

MINITAB gives parameter estimates from the probit model as:

$$\beta_0 = 1.7278 \quad \text{and} \quad \beta_1 = 2.6914$$

We observe here that the ratio of the logistic parameter model estimates to those of the logit model is

$$(2.9519/1.7278) = (4.6008/2.6914) = 1.708.$$

This ratio is expected to be within the range 1.6 to 1.8. Our ratios are within this range.

				Parameters	
Models	d.f.	X^2	p value	$\hat{\beta}_0$	$\hat{\beta}$
Logistic	3	0.505	0.918	2.9519	4.6008
Probit	3	0.312	0.958	1.7278	2.6914

The two models provide adequate fits of the data, although the probit model seems better both in terms of X^2 values and the standardized residuals (not printed). Both models are, of course, based on 3 degrees of freedom.
The final model based on the logistic regression is given in (17.6), while the estimated logistic regression model is plotted against the concentration levels in Fig. 17.4.

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 2.9519 + 4.6008\,\text{dose}_i, \quad i = 1, 2, \cdots, 5. \tag{17.6}$$

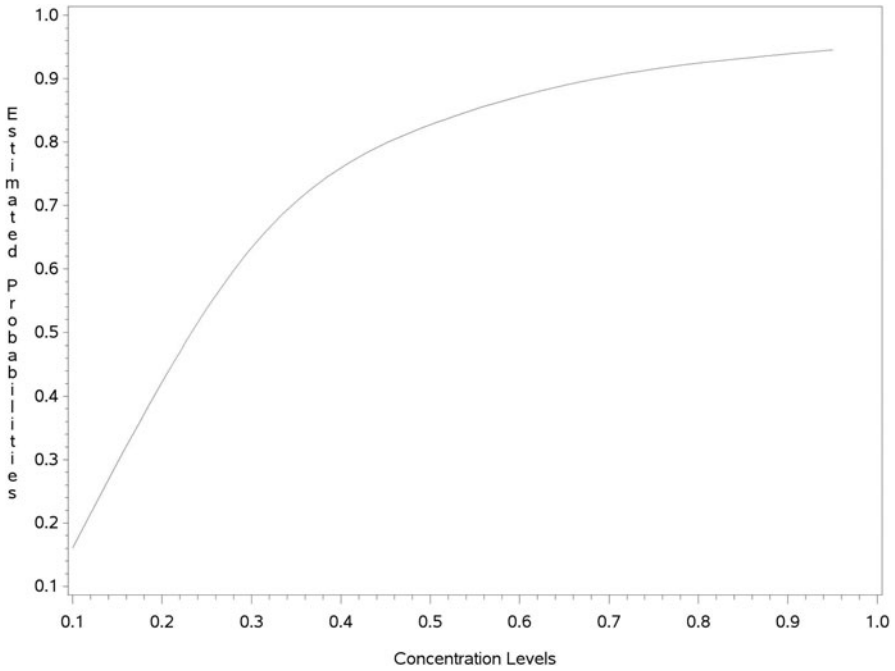


Fig. 17.4 Fitted logistic model

17.4 Parallel-Line Bioassay

For situations when we wish to compare the toxicity of two drugs, for instance, we need to fit individual logistic or probit models to both data and compute the ratio of their LD_{50}s.

As an example, consider the following case: The data in Table 17.5 were adapted from Litchfield and Wilcoxon (1949). The first set of data was from an experiment in which the antihistamine activities (the response) of certain animals to Tagathen (Chlorothen citrate) were studied. The second set of data also result from a similar experiment with triplennamine (Pyribenzamine) as the agent. The dose units are milligrams per kilogram and the number of animals tested at each dose level was eight.

Table 17.5 Data for this example on relative potency

Tagathen			Pyribenzamine		
Dose	n	No. Alive	Dose	n	No. Alive
0.025	8	1	0.175	8	1
0.125	8	4	0.35	8	3
0.25	8	4	0.7	8	5
0.50	8	7	1.4	8	5
1.0	8	8	2.8	8	8

Individual analyses are performed in MINITAB with the split command, after transforming the dose to log to base 10.

```
Data Display

Row    dose     r     n   drug         x

  1   0.025     1     8     TA  -1.60206
  2   0.125     4     8     TA  -0.90309
  3   0.250     4     8     TA  -0.60206
  4   0.500     7     8     TA  -0.30103
  5   1.000     8     8     TA   0.00000
  6   0.175     1     8     PY  -0.75696
  7   0.350     3     8     PY  -0.45593
  8   0.700     5     8     PY  -0.15490
  9   1.400     5     8     PY   0.14613
 10   2.800     8     8     PY   0.44716
```

The logistic regression model applied to both data gives respectively $X^2 = 2.038$ and $X^2 = 2.289$ on 3 d.f. Both models fit the individual data sets very well with respective p values of 0.565 and 0.515. The estimated logistic regression equations are

$$\ln\left(\frac{\hat{p}_{1i}}{1-\hat{p}_{1i}}\right) = 2.6181 + 3.0944\, x_{1i}, \quad i = 1, \cdots, 5$$

$$\ln\left(\frac{\hat{p}_{2i}}{1-\hat{p}_{2i}}\right) = 0.8290 + 3.422\, x_{2i}, \quad i = 1, \cdots, 4$$

$$\text{ED}_{50} = 10^{\left(\frac{-2.6181}{3.0944}\right)} = 0.1425 \quad \text{for drug TA}$$

$$\text{ED}_{50} = 10^{\left(\frac{-0.8290}{3.422}\right)} = 0.5725 \quad \text{for drug PY.}$$

Hence, potency ratio

$$\frac{\text{drug TA}}{\text{drug PY}} = \frac{\text{ED}_{50}^{TA}}{\text{ED}_{50}^{PY}} = 0.1425/0.5725 = 0.2489$$

Thus, drug TA is $\frac{1}{0.2489} = 4.02$ times more active than drug PY. That is, Tagathen is four times more active than Pyribenzamine. The individual analysis from MINITAB are presented below.

```
MTB > LET C5=LOGT(C1)
MTB > Split;
SUBC>   NoMatrices;
SUBC>   NoConstants;
SUBC>   By 'drug'.


Results for: potency.MTW(drug = TA)

MTB > Name c6 = 'EPRO1'
MTB > BLogistic 'r' 'n' = x;
SUBC>   ST;
SUBC>   Logit;
SUBC>   Eprobability 'EPRO1';
SUBC>   Brief 2.

Binary Logistic Regression: r, n versus x


Link Function:  Logit

Response  Information

Variable Value          Count
r        Success           24
         Failure           16
n        Total             40

Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z      P   Ratio    Lower    Upper
Constant      2.6181     0.8060     3.25 0.001
x             3.0944     0.9822     3.15 0.002   22.07     3.22   151.35

Log-Likelihood = -18.363
```

```
Test that all slopes are zero: G = 17.115, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

Method               Chi-Square    DF      P
Pearson                   2.038     3  0.565
Deviance                  2.488     3  0.477
Hosmer-Lemeshow           2.038     3 0.565


Results for: potency.MTW(drug = PY)

MTB > Name c6 = 'EPRO1'
MTB > BLogistic 'r' 'n' = x;
SUBC>   ST;
SUBC>   Logit;
SUBC>   Eprobability 'EPRO1';
SUBC>   Brief 2.

Binary Logistic Regression: r, n versus x


Link Function:  Logit

Response  Information

Variable Value          Count
r        Success           22
         Failure           18
n        Total             40

Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z     P   Ratio    Lower    Upper
Constant      0.8290     0.4480     1.85 0.064
x              3.422      1.083     3.16 0.002   30.64     3.66   256.23

Log-Likelihood = -20.291
Test that all slopes are zero: G = 14.470, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

Method               Chi-Square    DF      P
Pearson                   2.289     3  0.515
Deviance                  2.798     3  0.424
Hosmer-Lemeshow           2.289     3  0.515
```

17.5 Use of Joint Model

Alternatively, we could employ the joint model

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 \,\text{dose}_i + \beta_2 \,\text{drug}_j + \beta_3 \,\text{dose} * \text{drug}_{ij}$$

where p_{ij} is the probability of the animal receiving dose i of drug j if alive. *Dose* represents the dose effect, *drug* represents the effect of drug, and *dose*drug* represents the interaction effect of dose and drug.

```
MTB > BLogistic 'r' 'n' = x drug x*drug;
SUBC>   ST;
SUBC>   Factors 'drug';
SUBC>   Logit;
SUBC>   Brief 2.

Binary Logistic Regression: r, n versus x, drug


Link Function:  Logit

Response  Information

Variable  Value      Count
r         Success       46
          Failure       34
n         Total         80

Logistic Regression Table
                                                    Odds        95% CI
Predictor      Coef    SE Coef        Z      P    Ratio    Lower   Upper
Constant     0.8290     0.4480     1.85 0.064
x             3.422      1.083     3.16 0.002    30.64     3.66  256.23

drug
 TA          1.7891     0.9221     1.94 0.052     5.98     0.98   36.47

drug*x
 TA          -0.328      1.462    -0.22 0.823     0.72     0.04   12.66

Log-Likelihood = -38.653
Test that all slopes are zero: G = 31.790, DF = 3, P-Value = 0.000

Goodness-of-Fit Tests

Method            Chi-Square    DF      P
Pearson                4.327     6  0.633
Deviance               5.286     6  0.508
Hosmer-Lemeshow        4.327     8  0.827
```

The p value for the interaction term (drug $*\, x$) is 0.823, which indicates that the interaction term can be removed from the model. A result that leads to conclude the parallelism of the assays. Hence, a reduced model in (17.7) is given by:

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 \,\text{dose}_i + \beta_2 \,\text{drug}_j \tag{17.7}$$

where the drugs are dichotomized as:

$$\text{drug}_j = \begin{cases} 1 & \text{if drug TA} \\ 0 & \text{if drug PY.} \end{cases}$$

```
MTB > Name c6 = 'EPRO1'
MTB > BLogistic 'r' 'n' = x drug;
SUBC>   ST;
SUBC>   Factors 'drug';
SUBC>   Logit;
SUBC>   Eprobability 'EPRO1';
SUBC>   Brief 2.

Binary Logistic Regression: r, n versus x, drug


Link Function:  Logit

Response  Information

Variable  Value        Count
r         Success         46
          Failure         34
n         Total           80

Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z      P    Ratio    Lower    Upper
Constant      0.7930     0.4103     1.93 0.053
x             3.2485     0.7339     4.43 0.000    25.75     6.11   108.53
drug
 TA           1.9337     0.6730     2.87 0.004     6.91     1.85    25.86

Log-Likelihood = -38.679
Test that all slopes are zero: G = 31.740, DF = 2, P-Value = 0.000

Goodness-of-Fit Tests

Method              Chi-Square    DF      P
Pearson                  4.339     7  0.740
Deviance                 5.336     7  0.619
Hosmer-Lemeshow          4.339     8  0.825

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs          Number  Percent     Summary Measures
Concordant       1269    81.1%     Somers' D               0.69
Discordant        197    12.6%     Goodman-Kruskal Gamma   0.73
Ties               98     6.3%     Kendall's Tau-a         0.34
Total            1564   100.0%

MTB > print c1 c2 c4 c6

Data Display


 Row    dose     r   drug     phat

   1   0.025     1     TA   0.077447
   2   0.125     4     TA   0.448443
   3   0.250     4     TA   0.683722
   4   0.500     7     TA   0.851804
   5   1.000     8     TA   0.938584
   6   0.175     1     PY   0.158964
   7   0.350     3     PY   0.334463
   8   0.700     5     PY   0.571952
   9   1.400     5     PY   0.780351
  10   2.800     8     PY   0.904271
```

The model in (17.7) fits the data very well with $X^2 = 4.339$ on 7 d.f. (p value$= 0.740$). The parameter estimates of the model are

$$\hat{\beta}_0 = 0.7930 \quad \hat{\beta}_1 = 3.2485 \quad \hat{\beta}_2 = 1.9337.$$

Under the above coding scheme for drugs, therefore, the relative potency of drug PY to TA is computed as $10^{\frac{-\hat{\beta}_2}{\hat{\beta}_1}}$. In our case, this equals

$$10^{-1.9337/3.2485} = 10^{-0.5953} = 0.2539.$$

Hence again, drug TA is $\frac{1}{0.2539} = 3.94$ times more active than drug PY. The plot of the parallel models are displayed in Fig. 17.5.

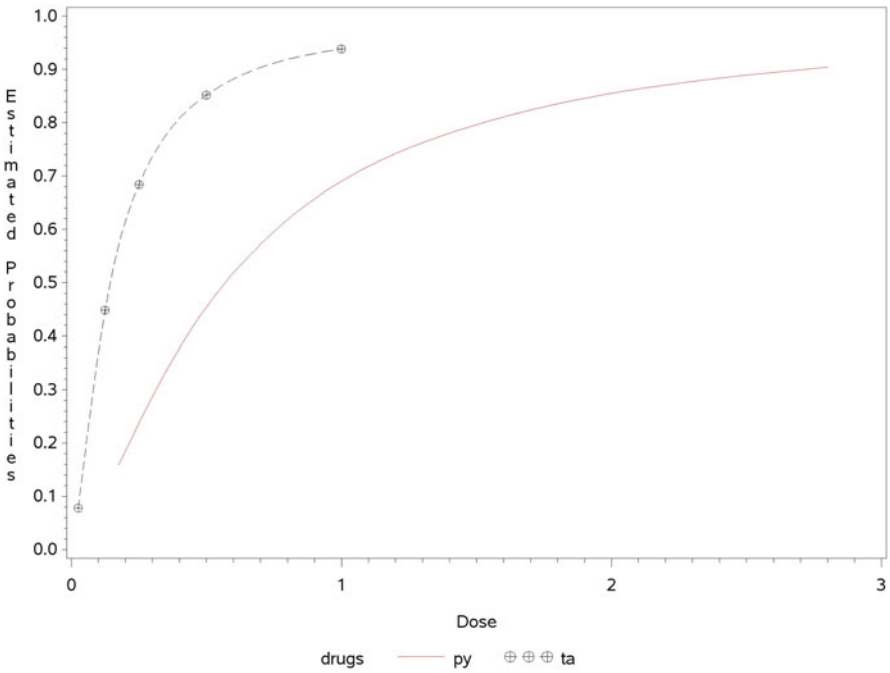


Fig. 17.5 Estimated probabilities plot

17.5.1 Example 17.3

The data in Table 17.6 from Breslow and Day (1980) relate to the occurrence of esophageal cancer in Frenchmen. Potential risk factors related to the occurrence are age and alcohol consumption where any consumption of wine more than 1 L a day is considered high.

(a) Fit a logistic model with the explanatory variables age and alcohol consumption by first considering age as a continuous variable.

Table 17.6 Occurrence of esophageal cancer

	Alcohol	Cancer	
Age group	consumption	Yes	No
25–34	High	1	9
	Low	0	106
35–44	High	4	26
	Low	5	164
45–54	High	25	29
	Low	21	138
55–64	High	42	27
	Low	34	139
65–74	High	19	18
	Low	36	88
75+	High	5	0
	Low	8	31

(b) Consider fitting the interaction term in both situations above. Use the stepwise regression procedure to fit the most parsimonious model. Interpret your results.

A logistic model involving age, x (alcohol consumption, coded 1 for high and 0 for low) with their interaction term gives a goodness-of-fit statistic $X^2 = 27.157$ on 8 d.f. (p value $= 0.0001$), which indicates that the model does not fit at all.

```
Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z      P    Ratio    Lower    Upper
Constant     -5.2879     0.5221   -10.13 0.000
age         0.061368   0.008531     7.19 0.000     1.06     1.05     1.08
x             1.7359     0.9492     1.83 0.067     5.67     0.88    36.47
age*x        0.00078    0.01642     0.05 0.962     1.00     0.97     1.03

Log-Likelihood = -404.905
Test that all slopes are zero: G = 179.678, DF = 3, P-Value = 0.000

Goodness-of-Fit Tests

Method               Chi-Square    DF     P
Pearson                  27.157     8 0.001
Deviance                 31.929     8 0.000
Hosmer-Lemeshow          12.063     5 0.034
```

Next, we introduce the quadratic effect of age into the model, and still retaining the interaction term between x and age with the following partial output from MINITAB.

```
Logistic Regression Table
                                                        Odds        95% CI
Predictor       Coef    SE Coef        Z      P    Ratio    Lower    Upper
Constant     -13.531      2.232    -6.06 0.000
age          0.35744    0.07545     4.74 0.000     1.43     1.23     1.66
age2      -0.0025467  0.0006315    -4.03 0.000     1.00     1.00     1.00
x             2.465      1.086     2.27 0.023    11.76     1.40    98.85
age*x       -0.01366    0.01846    -0.74 0.459     0.99     0.95     1.02

Log-Likelihood = -395.182
Test that all slopes are zero: G = 199.125, DF = 4, P-Value0= 0.0

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                   10.274     7  0.174
Deviance                  12.482     7  0.086
Hosmer-Lemeshow            5.442     5  0.364
```

The model now fits the data with $X^2 = 10.274$ on 7 d.f. (p value $= 0.174$). However, the parameter estimates indicate that the interaction term is not significant and could, thus, be dropped from the model. Further, the quadratic term is highly significant in the model (p value $= 0.000$). Hence, our final model would be the model

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ age}_i^2 + \beta_3\, x \tag{17.8}$$

where the alcohol consumption x is defined as

$$x = \begin{cases} 1 & \text{if high} \\ 0 & \text{if low.} \end{cases}$$

The MINITAB implementation of the model in (17.8) gives the following partial output.

```
MTB > BLogistic 'status' = age age2 x;
SUBC>   Frequency 'freq';
SUBC>   Logit;
SUBC>   Eprobability 'EPRO1';
SUBC>   Brief 2.

Binary Logistic Regression: status versus age, age2, x

Link Function:  Logit

Response Information

Variable Value       Count
status   yes           200  (Event)
         no            775
         Total         975
Frequency:  freq

   22 cases were used
    2 cases contained missing values
      or was a case with zero frequency.
```

```
Logistic Regression Table
                                                        Odds       95% CI
Predictor       Coef    SE Coef        Z      P      Ratio   Lower     Upper
Constant     -13.007      2.084    -6.24  0.000
age          0.34378    0.07249     4.74  0.000     1.41    1.22      1.63
age2      -0.0024665  0.0006195    -3.98  0.000     1.00    1.00      1.00
x             1.6744     0.1897     8.83  0.000     5.34    3.68      7.74

Log-Likelihood = -395.455
Test that all slopes are zero: G = 198.579, DF = 3, P-Value0= 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                   11.135     8  0.194
Deviance                  13.028     8  0.111
Hosmer-Lemeshow            3.580     5  0.611
```

The implementation of the model described in (17.8) gives a $X^2 = 11.135$ on 8 d.f. The model fits the data well. Parameter estimates of β_1, β_2, and β_3 are highly significant. The analysis shows that the odds of an cancer status being yes is $e^{1.6744} = 5.335$ times higher for high-alcohol-consumption individuals than those on low alcohol consumption when the effect of age is controlled.

We also obtain the expected probabilities ($\hat{p}_i$), of having cancer based on the estimated logistic model. The estimated logistic regression is plotted in Fig. 17.6

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -13.007 + 0.3438\,\text{age}_i - 0.00247\,\text{age}_i^2 + 1.6744\,x. \qquad (17.9)$$

```
Data Display

 Row    age     xx  status   freq    x      age2      phat

   1   29.5   high     yes      1    1    870.25   0.03430
   2   29.5   high      no      9    1    870.25   0.03430
   3   29.5    low     yes      0    0    870.25   0.00661
   4   29.5    low      no    106    0    870.25   0.00661
   5   39.5   high     yes      4    1   1560.25   0.16774
   6   39.5   high      no     26    1   1560.25   0.16774
   7   39.5    low     yes      5    0   1560.25   0.03640
   8   39.5    low      no    164    0   1560.25   0.03640
   9   49.5   high     yes     25    1   2450.25   0.41115
  10   49.5   high      no     29    1   2450.25   0.41115
  11   49.5    low     yes     21    0   2450.25   0.11572
  12   49.5    low      no    138    0   2450.25   0.11572
  13   59.5   high     yes     42    1   3540.25   0.59629
  14   59.5   high      no     27    1   3540.25   0.59629
  15   59.5    low     yes     34    0   3540.25   0.21681
  16   59.5    low      no    139    0   3540.25   0.21681
  17   69.5   high     yes     19    1   4830.25   0.65609
  18   69.5   high      no     18    1   4830.25   0.65609
  19   69.5    low     yes     36    0   4830.25   0.26339
  20   69.5    low      no     88    0   4830.25   0.26339
  21   79.5   high     yes      5    1   6320.25   0.60074
  22   79.5   high      no      0    1   6320.25   0.60074
  23   79.5    low     yes      8    0   6320.25   0.21998
  24   79.5    low      no     31    0   6320.25   0.21998
```

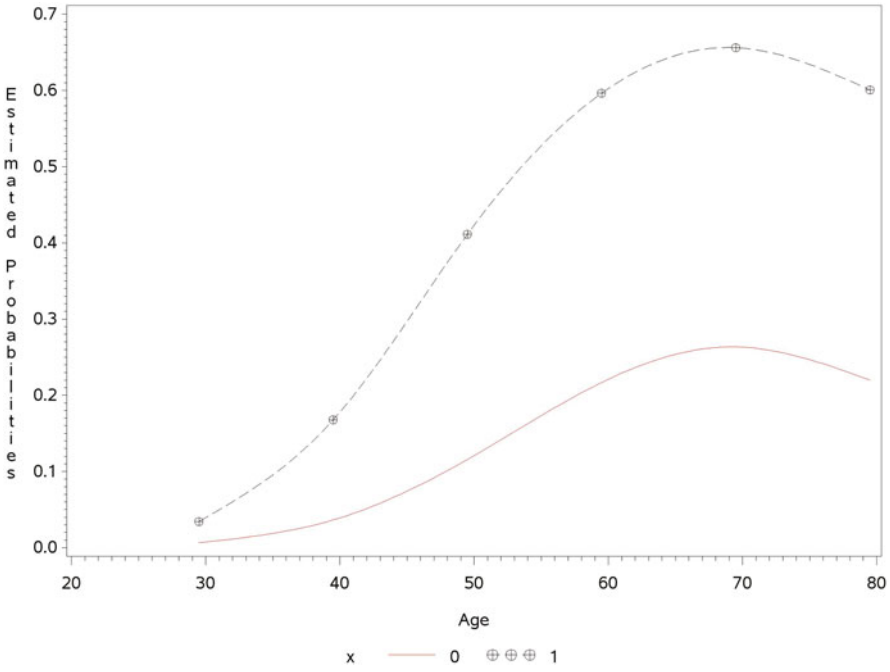


Fig. 17.6 Estimated probabilities plot

17.5.2 HIV Status Data Example

The following data in Table 17.7 come from Schork and Remington (2000). The data represent the outcome variable, HIV status (0 = no, 1 = yes), factor variables IV (intravenous) drug status (0 = no, 1 = yes), and number of sexual partners for 25 men selected from a homeless shelter.

Table 17.7 Data for the HIV status example

ID	STATUS	IVDRUG	SEXPART	ID	STATUS	IVDRUG	SEXPART
1	0	0	4	14	0	0	5
2	0	1	4	15	1	1	9
3	1	1	3	16	1	0	19
4	0	0	2	17	0	0	7
5	0	0	7	18	1	1	10
6	1	0	12	19	0	0	5
7	1	1	8	20	1	1	8
8	0	0	1	21	0	0	14
9	1	0	9	22	0	1	8
10	0	0	5	23	1	0	14
11	0	0	6	24	1	1	9
12	0	1	4	25	1	1	17
13	0	1	2				

Suppose we define the response variable to be Y_i from individual i, then, we have:

$$Y_i = \begin{cases} 1 & \text{if STATUS is 1} \\ 0 & \text{otherwise} \end{cases}$$

If we assume that the probability p_i of yes HIV status depends on the drug, SEXPART, and the interaction between drug and SEXPART, then our logistic model would be as described in (17.10), while the corresponding probability is as defined in (17.11).

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{ drug}_i + \beta_2 \text{ part}_i + \beta_3 \text{ drug} * \text{ part}_i \tag{17.10}$$

$$\text{Prob}(\text{Y}_i = 1|\text{drug}_i, \text{part}_i) = \frac{\exp\left(\beta_0 + \beta_1 \text{drug}_i + \beta_2 \text{part}_i + \beta_3 \text{drug*part}_i\right)}{1 + \exp\left(\beta_0 + \beta_1 \text{drug}_i + \beta_2 \text{part}_i + \beta_3 \text{drug*part}_i\right)} \tag{17.11}$$

where π_i is the probability of the ith individual having the HIV, *drug* represents the IV drug effect, *part* represents the effect of the number of sexual partners, and *drug*part* represents the interaction effect of IV drug and numbers of sexual partners. Also,

$$\text{drug}_i = \begin{cases} 1 & \text{if yes} \\ 0 & \text{if otherwise.} \end{cases}$$

The implementation of the model in (17.10) in MINITAB is carried out with the following codes and the corresponding partial output.

```
MTB > print c1-c4

Data Display

Row  y  x1  x2  x12
  1  0   0   4    0
  2  0   1   4    4
  3  1   1   3    3
  4  0   0   2    0
  5  0   0   7    0
  6  1   0  12    0
  7  1   1   8    8
  8  0   0   1    0
  9  1   0   9    0
 10  0   0   5    0
 11  0   0   6    0
 12  0   1   4    4
 13  0   1   2    2
 14  0   0   5    0
 15  1   1   9    9
 16  1   0  19    0
 17  0   0   7    0
```

```
 18  1   1  10   10
 19  0   0   5    0
 20  1   1   8    8
 21  0   0  14    0
 22  0   1   8    8
 23  1   0  14    0
 24  1   1   9    9
 25  1   1  17   17

MTB > Blogistic 'y' = x1 x2 x12;
SUBC>   Logit;
SUBC>   Reference 'y'  1;
SUBC>   Brief 2.

Binary Logistic Regression: y versus x1, x2, x12

Link Function: Logit


Response Information

Variable  Value  Count
y         1         11  (Event)
          0         14
          Total     25


Logistic Regression Table

                                                    Odds      95% CI
Predictor       Coef   SE Coef      Z      P  Ratio  Lower    Upper

Constant    -5.33038   2.52235  -2.11  0.035
x1           2.51506   3.22805   0.78  0.436  12.37   0.02  6918.35
x2          0.481626  0.241700   1.99  0.046   1.62   1.01     2.60
x12        0.0341946  0.386656   0.09  0.930   1.03   0.48     2.21

Log-Likelihood = -9.095
Test that all slopes are zero: G = 16.107, DF = 3, P-Value = 0.001

Goodness-of-Fit Tests

Method           Chi-Square  DF      P
Pearson             10.4732  13  0.655
Deviance            11.5979  13  0.561
Hosmer-Lemeshow      3.5724   8  0.893


Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs       Number  Percent  Summary Measures
Concordant     139     90.3  Somers' D              0.82
Discordant      12      7.8  Goodman-Kruskal Gamma  0.84
Ties             3      1.9  Kendall's Tau-a        0.42
Total          154    100.0
```

The global hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{vs.}$$
$$H_a : \text{at least one of these } \beta \text{ parameters is not zero} \tag{17.12}$$

is tested with the value of $G = 16.107$ and corresponding p value of 0.001 on 3 d.f. The p value of 0.001 at $\alpha = 0.05$ level of significance, indicates that we would have to reject H_0, and therefore, conclude that at least one of the parameters β_1, β_2, and β_3 is not zero. The model fits the data well with deviance being 211.5979 on 13 degrees of freedom with corresponding (p value = 0.561). The resulting estimated model is, therefore, given by

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -5.3304 + 2.5151\ \text{drug}_i + 0.4816\ \text{part}_i - 0.0342\ \text{drug*part}_i.$$

Analysis of the results, however, indicates that the interaction effect is not significant. This result is obtained from the logistic table in the MINITAB output. Thus, the test of the hypotheses

$$H_0 : \beta_3 | \beta_1, \beta_2 = 0 \quad \text{vs.} \quad H_a : \beta_3 | \beta_1, \beta_2 \neq 0$$

gives a p value of 0.930, which indicates that we would fail to reject H_0. That is, the interaction term $x_{12} = x_1 \times x_2 = \text{drug} * \text{part}_i$ is not important in the model, given that x_1 and x_2 are already in the model. We can, therefore, fit the reduced model involving only x_1 and x_2 as explanatory variables. That is, the model

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1\ \text{drug}_i + \beta_2\ \text{part}_i \tag{17.13}$$

with corresponding probability of HIV status being "1" as

$$\text{Prob}(Y = 1, |x_1, x_2) = \frac{\exp(\beta_0 + \beta_1\ \text{drug}_i + \beta_2\ \text{part}_i)}{1 + \exp(\beta_0 + \beta_1\ \text{drug}_i + \beta_2\ \text{part}_i)}. \tag{17.14}$$

Again, the model in (17.13) is estimated in MINITAB with the following codes and partial output.

```
MTB > Name c5 "EPRO1"
MTB > Blogistic 'y' = x1 x2 ;
SUBC>   Logit;
SUBC>   Reference 'y'  1;
SUBC>   Eprobability 'EPRO1';
SUBC>   Brief 2.

Binary Logistic Regression: y versus x1, x2

Link Function: Logit
```

```
Response Information

Variable  Value  Count
y         1         11 (Event)
          0         14
          Total     25


Logistic Regression Table

                                                 Odds      95% CI
Predictor      Coef   SE Coef      Z      P Ratio Lower    Upper
Constant   -5.46491   2.06826  -2.64  0.008
x1          2.77476   1.38807   2.00  0.046 16.03  1.06   243.55
x2         0.495392  0.190061   2.61  0.009  1.64  1.13     2.38


Log-Likelihood = -9.099
Test that all slopes are zero: G = 16.099, DF = 2, P-Value = 0.000


Goodness-of-Fit Tests

Method           Chi-Square  DF      P
Pearson             10.3988  14  0.732
Deviance            11.6057  14  0.638
Hosmer-Lemeshow      3.5322   8  0.897



Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs       Number  Percent  Summary Measures
Concordant     139     90.3  Somers' D              0.82
Discordant      12      7.8  Goodman-Kruskal Gamma  0.84
Ties             3      1.9  Kendall's Tau-a        0.42
Total          154    100.0
```

The model in (17.13), when implemented, fits the data very well with a deviance value of 11.6057 on 14 degrees of freedom and p value of 0.638. The parameter estimates are very important in the model and the estimated logistic regression equation is

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -5.4649 + 2.7748\,\text{drug}_i + 0.4954\,\text{part}_i, \tag{17.15}$$

with a corresponding estimated probability

$$\hat{p}_i = \frac{\exp(-5.4649 + 2.7748\,\text{drug}_i + 0.4954\,\text{part}_i)}{1 + \exp(-5.4649 + 2.7748\,\text{drug}_i + 0.4954\,\text{part}_i)}. \tag{17.16}$$

17.5.3 Interpretation of Parameters

We have

$$\exp(\beta_0) = \frac{\text{Prob}(Y = 1|\text{drug} = \text{part} = 0)}{\text{Prob}(Y = 0|\text{drug} = \text{part} = 0)}$$
$$= \text{Odds of a positive HIV status}$$
$$\exp(\beta_1) = \frac{\text{Odds of a positive HIV status when drug}_i = 1, \text{part}_i = 0}{\text{Odds of a positive HIV status at baseline}},$$
$$\exp(\beta_2) = \frac{\text{Odds of a positive HIV status when drug}_i = 0, \text{part}_i = 1}{\text{Odds of a positive HIV status at baseline}},$$

where $\exp(\beta_0)$ is often referred to as the odds of a positive HIV status at the baseline. That is, at $(\text{drug}_i = \text{part}_i = 0)$.

The analysis shows that the odds of an HIV status being positive is 16.035 times higher for IV drug users than those not on IV drugs when the effect of sexual partner is controlled. Similarly, the odds increase by 1.641 for a unit increase in the number of sexual partners. The odds increase by $2.693 = e^{2*0.4954} = 1.641^2$ and $4.420 = e^{3*0.4954} = 1.641^3$ for two-unit and three-unit increases in the number of sexual partners, respectively.

In general, the estimated odds of a positive HIV status for any given drug, part is

$$= \exp(\hat{\beta}_0) \times \exp(\hat{\beta}_1 \text{ drug}_i) \times \exp(\hat{\beta}_2 \text{ part}_i)$$
$$= \left\{ \begin{matrix} \text{odds} \\ \text{for} \\ \text{baseline} \end{matrix} \right\} \times \left\{ \begin{matrix} \text{factor} \\ \text{due} \\ \text{to drug}_i \end{matrix} \right\} \times \left\{ \begin{matrix} \text{factor} \\ \text{due} \\ \text{to part}_i \end{matrix} \right\}.$$

These expected probabilities are presented below. They are generated automatically by MINITAB and stored in neutral column. We have produced the estimated probabilities below.

Obs	HIV	DRUG	SEXPART	_LEVEL_	PHAT
1	0	0	4	1	0.02979
2	0	1	4	1	0.32991
3	1	1	3	1	0.23077
4	0	0	2	1	0.01127
5	0	0	7	1	0.11950
6	1	0	12	1	0.61770
7	1	1	8	1	0.78125
8	0	0	1	1	0.00690
9	1	0	9	1	0.26769

10	0	0	5	1	0.04797
11	0	0	6	1	0.07638
12	0	1	4	1	0.32991
13	0	1	2	1	0.15455
14	0	0	5	1	0.04797
15	1	1	9	1	0.85425
16	1	0	19	1	0.98106
17	0	0	7	1	0.11950
18	1	1	10	1	0.90583
19	0	0	5	1	0.04797
20	1	1	8	1	0.78125
21	0	0	14	1	0.81314
22	0	1	8	1	0.78125
23	1	0	14	1	0.81314
24	1	1	9	1	0.85425
25	1	1	17	1	0.99677

Here, for instance, for observation 22, the estimated probability is computed as

$$\begin{aligned}
\text{Prob}(Y = 0, |\text{drug} = 1, \text{part} = 8) &= \frac{\exp(-5.4649 + 2.7748\,\text{drug}_i + 0.4954\,\text{part}_i)}{1 + \exp(-5.4649 + 2.7748\,\text{drug}_i + 0.4954\,\text{part}_i)} \\
&= \frac{\exp(-5.4649 + 2.7748 \times (1) + 0.4954 \times 8)}{1 + \exp(5.4649 + 2.7748 \times (1) + 0.4954 \times 8)} \\
&= \frac{\exp(1.2731)}{1 + \exp(1.2731)} = \frac{3.5719}{4.5719} \\
&= 0.7812
\end{aligned}$$

Others can be similarly computed, but as indicated earlier, these probabilities are automatically generated in MINITAB upon request. The estimated logistic model is displayed in Fig. 17.7.

17.5.4 ROC Curve for the Analysis

As discussed in Chap. 4, the receiver operating characteristic curve for the final model is presented in Fig. 17.8. The area under the curve is estimated to be 0.9123, indicating a very good classification rule.

We present below the sensitivity and specificity of the model as produced by SAS. We also have the positively and negatively classified, the false positives and the false negatives.

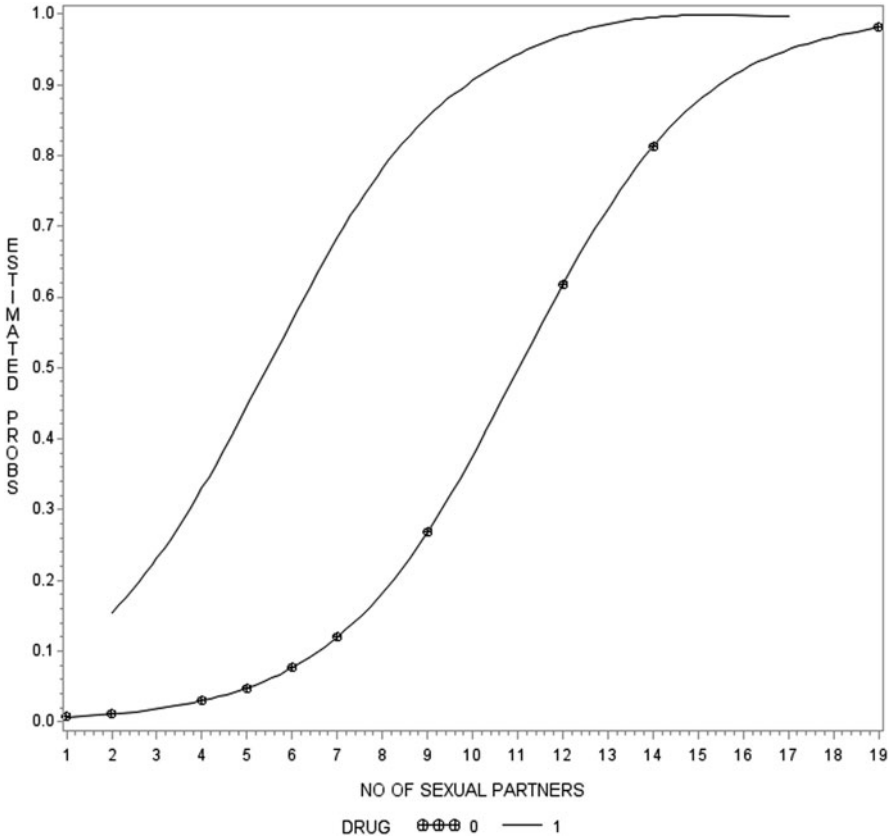


Fig. 17.7 Predicted probability plots from model (17.16)

Obs	_PROB_	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_
1	0.99677	1	14	0	10	0.09091	0.00000
2	0.98106	2	14	0	9	0.18182	0.00000
3	0.90583	3	14	0	8	0.27273	0.00000
4	0.85425	5	14	0	6	0.45455	0.00000
5	0.81314	6	13	1	5	0.54545	0.07143
6	0.78125	8	12	2	3	0.72727	0.14286
7	0.61770	9	12	2	2	0.81818	0.14286
8	0.32991	9	10	4	2	0.81818	0.28571
9	0.26769	10	10	4	1	0.90909	0.28571
10	0.23077	11	10	4	0	1.00000	0.28571
11	0.15455	11	9	5	0	1.00000	0.35714
12	0.11950	11	7	7	0	1.00000	0.50000
13	0.07638	11	6	8	0	1.00000	0.57143
14	0.04797	11	3	11	0	1.00000	0.78571
15	0.02979	11	2	12	0	1.00000	0.85714
16	0.01127	11	1	13	0	1.00000	0.92857
17	0.00690	11	0	14	0	1.00000	1.00000

17.6 Exercises

1. In an acute toxicity testing of Jubi Formula, the following data were obtained after oral administration.

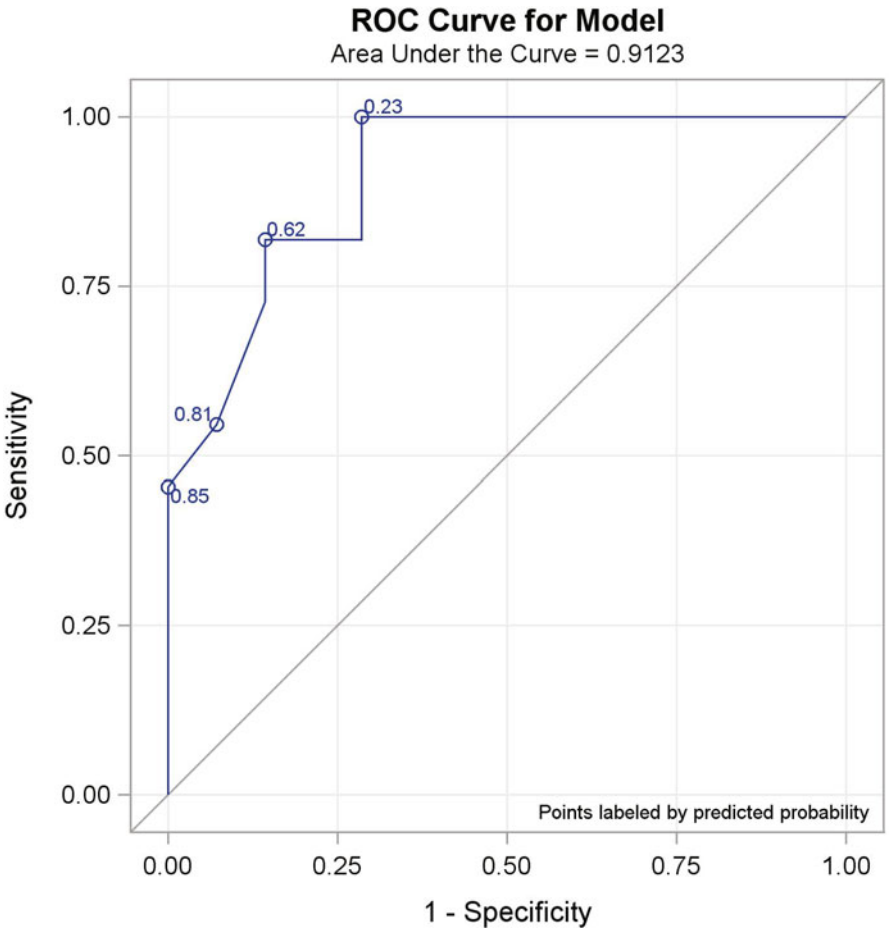


Fig. 17.8 ROC curve for the HIV data

Dose (mg)	Dose (mg/kg)	# mice	# mice that died
0.00	0.00	6	0
0.82	49.85	6	0
1.64	99.70	6	0
2.46	149.54	6	1
3.28	199.39	6	2
4.10	249.24	6	3
5.74	348.95	6	5

Fit a logistic regression model to the above data and compute the LD_{50}. What can you say about the toxicity of the formula?

2. The data in the table below are reported in Woodward et al. (1941) and is reproduced from Christensen (1991). The data examines the relationship between exposure to chloroacetic acid and the death of mice. Ten mice were exposed at each dose level and the doses are measured in grams per kilogram of body weight.

Dose	# dead	# exposed
0.0794	1	10
0.1000	2	10
0.1259	1	10
0.1413	0	10
0.1500	1	10
0.1588	2	10
0.1778	4	10
0.1995	6	10
0.2239	4	10
0.2512	5	10
0.2818	5	10
0.3162	8	10

Fit the logistic regression model to the data and estimate the LD_{50}, LD_{90}, and LD_{999}. Discuss the possible danger of extrapolation to LD_{999}. Determine how well the model fits the data.

3. An antihistaminic drug was used at various doses to protect test animals against a certain lethal dose of histamine, with the results given below.

Dose μg/kg	Alive/ total
1000	8/8
500	7/8
250	4/8
125	4/8
62.5	1/8

Fit the logistic and probit models to the data above and compute the LD_{50} in each case. Comment about your models.

4. Two anticonvulsant drugs were compared by administering them to mice which were then given electric shock under conditions that caused all control mice to convulse. The results of the experiment are displayed in the table below (Goldstein 1965).

   (a) Fit separate regression lines for both drugs, and hence, obtain an estimate of the relative potency from estimates of their LD_{50}s.
   (b) Fit a combined regression line and test for equality slopes. Test whether there is dosage and/or drug effects. Summarize your conclusions.

Drug A	
Dose mg/kg	Convulsed/ Total
10	13/15
30	9/15
90	4/15

Drug B	
Dose mg/kg	Convulsed/ Total
200	12/15
600	6/15
1800	2/15

5. The example below relates to a sample of patients with coronary heart disease (CHD) and a "normal" sample free of CHD (Lunneborg 1994). A 1 indicates the patient has no CHD, while a 2 indicates that the patient has CHD. Three risk factors are being evaluated. The risk factors are systolic blood pressure (sbp), blood-cholesterol level (chol), and age of the patients.

Group	SBP	Chol	Age
No	135	227	45
No	122	228	41
No	130	219	49
No	148	245	52
No	146	223	54
No	129	215	47
No	162	245	60
No	160	262	48
No	144	230	44
No	166	255	64
No	138	222	59
No	152	250	51
No	138	264	54
No	140	271	56
No	134	220	50
Yes	145	238	60
Yes	142	232	64
Yes	135	225	54
Yes	149	230	48
Yes	180	255	43
Yes	150	240	43
Yes	161	253	63
Yes	170	280	63
Yes	152	271	62
Yes	164	260	65

If we define the variable Y to be

$$Y = \begin{cases} 1 & \text{if CHD} \\ 0 & \text{if no CHD.} \end{cases}$$

6. The data below give the effect of different concentrations of nicotine sulfate in a 1 % saponin solution on an insect *Drosophila melanogaster,* the fruit fly.

Nicotine sulfate (g/100 cc) x_i	Number killed r_i	Number of insects n_i
0.10	8	47
0.15	14	53
0.20	24	55
0.30	32	52
0.50	38	46
0.70	50	54
0.95	50	52

Fit a logistic regression to the above data.

(a) What would be the explanatory variable in this model?
(b) Write down the estimated regression equation.
(c) Estimate the LD_{50}. What does this mean?
(d) Estimate the LD_{90}. What does this mean?
(e) Find the estimated proportion of insects *D. melanogaster,* killed when given a saponin solution concentration of 0.20 g/100 cc.

7. Hastie and Tibshirani (1990) described a study to determine risk factors for kyphosis, severe forward flexion of the spine following corrective spinal surgery. The ages in months at the time of the operation for the 18 subjects for whom kyphosis was present were 12, 15, 42, 52, 59, 73, 82, 91, 96, 105, 114, 120, 121, 128, 130, 139, 139, 157; and 22 of the subjects for whom kyphosis was absent were 1, 1, 2, 8, 11, 18, 22, 31, 37, 61, 72, 81, 97, 112, 118, 127, 131, 140, 151, 159, 177, and 206.

   (a) Fit a logistic regression model using age as a predictor of whether kyphosis is present. Test whether age has a significant effect.
   (b) Fit the model $\text{logit}[\pi(x)] = \beta_0 + \beta_1 x + \beta_2 x^2$. Test the significance of the squared age term, plot the fit, and interpret.

8. The following data relate to the outcome of the rate at which blood cells (erythrocytes) settle out of suspension in blood plasma. The response, y, is 1 if erythrocyte sedimentation (ES) exceeds 20 mm/h and values below this characterize healthy individuals. Positive responses are known to be associated with fibrinogen (x_1) and gamma-globulin (x_2). Fit a parsimonious logistic regression model to these data.

x_1	x_2	y	x_1	x_2	y	x_1	x_2	y	x_1	x_2	y
2.52	38	0	3.15	39	0	3.53	46	1	2.67	39	0
2.56	31	0	2.60	41	0	2.68	34	0	2.29	31	0
2.19	33	0	2.29	36	0	2.60	38	0	2.15	31	0
2.18	31	0	2.35	29	0	2.23	37	0	2.54	28	0
3.41	37	0	5.06	37	1	2.88	30	0	3.93	32	1
2.46	36	0	3.34	32	1	2.65	46	0	3.34	30	0
3.22	38	0	2.38	37	1	2.09	44	1	2.99	36	0
2.21	37	0	3.15	36	0	2.28	36	0	3.32	35	0

9. The following data from Guerrero and Johnson (1982) relate to the number of Warsaw girls that have menstruated given 25 groups of ages at menarche of 3918 girls. The total number of girls in each group (n) and the number having experienced menarche (r) are presented in the data *menarche* with the mean age (x) of the group. Fit a logistic model to the data with age as the explanatory variable.

x	n	r	x	n	r	x	n	r	x	n	r
9.21	376	0	11.83	111	17	13.58	105	81	15.33	111	107
10.21	200	0	12.08	100	16	13.83	117	88	15.58	94	92
10.58	93	0	12.33	93	29	14.08	98	79	15.83	114	112
10.83	120	2	12.58	100	39	14.033	97	90	17.58	1049	1049
11.08	90	2	12.83	108	51	14.58	120	113			
11.33	88	5	13.08	99	47	14.83	102	95			
11.58	105	10	13.33	106	67	15.08	122	117			

10. A local health clinic sent fliers to its clients to encourage everyone, especially older people at a high risk of complications, to get flu shots in time for protection against an expected flu epidemic. In a pilot follow-up study, 50 clients were randomly selected and asked whether they actually received a flu shot. In addition, data were collected on their age (x). A client who received a flu shot was coded $y = 1$, and a client who did not receive a flu shot was coded $y = 0$. A simple logistic regression model is fitted to the following data.

y	Age (years)
0	38, 41, 43, 34, 31, 54, 63, 38, 28, 42, 36, 45, 47, 53, 42, 42, 48, 46, 44, 46, 35 40, 40, 64, 34, 38, 56, 45, 33
1	52, 46, 41, 57, 49, 53, 39, 53, 49, 49, 46, 54, 63, 56, 64, 52, 46, 57, 56, 46, 47

(a) Find the maximum likelihood estimates of β_0 and β_1. State the fitted response logistic regression function.
(b) Obtain $\exp(\beta_1)$ and interpret the number.
(c) What is the estimated probability that clients aged 55 will receive a flu shot?
(d) Obtain an approximate 95 % confidence interval for the regression coefficient β_1. Convert this interval into one for the odds ratio.
(e) What is the estimated age at which 60 % of the clients will receive flu shots?
(f) Is the estimated age in (e) reasonable? Explain.
(g) What is the estimated age at which 80 % of the clients will receive flu shots?
(h) Is the estimated age in (g) reasonable? Explain.
(i) Based on your result in (c), assess the success of the fliers.
(j) Using the five-step procedure and $\alpha = 0.05$, test the null hypothesis that the regression coefficient β_1 is nonpositive. State the p value of your test.

11. In an experiment testing the effect of toxic substances, 1500 experimental insects were divided at random into six groups of 25 each. The insects in each group were exposed to a fixed dose of the toxic substance. A day later, each insect was observed. Death from exposure was scored 1 and survival was scored 0. The results are shown in the following table; x_i denotes the dose level received by the insects in group i and r_i denotes the number of insects that died out of 250 (n_i) in the group.

Group	Time					
x_i	1	2	3	4	5	6
r_i	28	53	93	126	172	197
n_i	250	250	250	250	250	250

(a) Fit a logistic regression response to the data.
(b) Obtain $\exp(\beta_1)$ and interpret this number.
(c) What is the estimated probability that an insect dies when the dose level is 3.5?
(d) What is the estimated median lethal dose- that is, the dose for which 50 % of the experimental insects are expected to die?
(e) Obtain an approximate 95 % confidence interval for β_1. Convert this interval into one for the odds ratio. Interpret this latter interval.

12. The data below, reported in Woodward et al. (1941), examined the relationship between exposure to chloracetic acid and the death of mice. Ten mice were exposed at each dose level and the doses are measured in grams per kilogram of the body weight.

Dose	# Dead	# Exposed	Dose	# Dead	# Exposed
0.0794	1	10	0.1778	4	10
0.1000	2	10	0.1995	6	10
0.1259	1	10	0.2239	4	10
0.1413	0	10	0.2512	5	10
0.1500	1	10	0.2818	5	10
0.1588	2	10	0.3162	8	10

(a) Fit an appropriate logistic regression model to the data.
(b) What is the estimated equation? Interpret the estimated parameters of the model.

Chapter 18
Repeated Measures Design

18.1 Introduction

Data collection in experiments usually involves two methods. The first method usually involves administering the treatment to different subjects which are stratified into groups and then measuring the outcome or dependent variable. Thus, different groups of subjects take part in the experimental condition. We could then conduct a one-way ANOVA to test differences in the mean outcome variables across the groups. This method is often referred to as the *between-group* or *between-subjects* design.

The second method is to manipulate the explanatory variable on the same subjects. That is, measure the outcome variable at different points in time on the same subjects. This method is called the *within-subject* or *repeated-measures* design. The method of analysis of the data depends on which method was employed in the collection of data.

While the within-subject designs are certainly more powerful, it does have its drawback: namely, it has the potential for carryover effects. This often happens if, for instance, we perceive that individual outcome scores increased because the subjects have gained practice or familiarity with the treatment. However, within-subject designs require far fewer subjects than the between-subject design.

As an example, consider h independent groups of patients each of whom are subjected to repeated measurements of the same response variable, y, at t time periods. If we let n_i represent the number of patients in group i $(i = 1, 2, \cdots, h)$, then a typical data structure for $h = 3$ is presented in Table 18.1, where each of $\mathbf{x}_{ij}$ and $\mathbf{y}_i$ are t-dimensional.

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_18,

Table 18.1 Typical data structure

		Time			
Group	Patient	1	2	$\cdots$	t
1	101	y_{111}	y_{112}	$\cdots$	y_{11t}
	102	y_{121}	y_{122}	$\cdots$	y_{12t}
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
	n_1	y_{1n_11}	y_{1n_12}	$\cdots$	y_{1n_1t}
2	201	y_{211}	y_{212}	$\cdots$	y_{21t}
	202	y_{221}	y_{222}	$\cdots$	y_{22t}
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
	n_2	y_{2n_21}	y_{2n_22}	$\cdots$	y_{2n_2t}
3	301	y_{311}	y_{312}	$\cdots$	y_{31t}
	302	y_{321}	y_{322}	$\cdots$	y_{32t}
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
	n_3	y_{3n_31}	y_{3n_32}	$\cdots$	y_{3n_3t}

18.2 Single-Factor Experiments with Repeated Measures

We give an example of this case in Example 18.1.1 below.

Example 18.1.1

This example is taken from Winer et al. (1991). The data in Table 18.2 relate to scores on five individuals on four different drugs (1, 2, 3, and 4). We are interested if there are differences in the mean scores among the four drugs.

Table 18.2 Scores on five subjects administered four different drugs

	Drug			
Subject	1	2	3	4
1	30	28	16	34
2	14	18	10	22
3	24	20	18	30
4	38	34	20	44
5	26	28	14	30

The above observations can be viewed as a repeated-measures data set since there are four measurements on each of the subjects. If we consider the data as a one-factor experiment, then the model of interest would be as in (18.1),

$$y_{ij} = \mu + t_j + \varepsilon_{ij} \quad i = 1, \cdots, 5; \quad j = 1, 2, \cdots, 4 \tag{18.1}$$

with the corresponding MINITAB output displayed below:

```
MTB > Oneway 'score' 'drug'.

One-way ANOVA: score versus drug

Source  DF      SS     MS     F      P
drug     3   698.2  232.7  4.69  0.016
Error   16   793.6   49.6
Total   19  1491.8

S = 7.043   R-Sq = 46.80%   R-Sq(adj) = 36.83%
```

Although the F p-value $= 0.016$ which indicates that significant differences exist between the means of the four drugs; however, the assumption of the one-way ANOVA that the observations be independently distributed has been violated here since the observations on each subject are undoubtedly correlated. The analysis would have been valid had we collected the four observations for each drug on four different subjects. However, the advantage of the repeated measures data above is that it allows us to gain a better precision as it allows us to compare drugs within each subject rather than between subjects. To overcome the short fall in the above initial analysis, the appropriate model for this example is, therefore, a two-way ANOVA model with

$$y_{ij} = \mu + b_i + t_j + \varepsilon_{ij} \quad i = 1, \cdots, 5; \quad j = 1, 2, \cdots, 4. \tag{18.2}$$

We observe the model in (18.2) is the familiar randomized complete block design (RCBD) discussed in Chap. 11. Here, however, the blocks are the persons (subjects) and drugs 1 to 4 are the treatments. Thus, the analysis and the hypothesis testing are equivalent to those discussed under the RCBD. Further, we have assumed that the four drug treatments are fixed and that there is no person–drug interaction. Consequently, the hypotheses of interest are

$$H_0: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_a: \quad \text{at least two of these means are unequal.}$$

The corresponding variances and covariances for the four repeated measures are presented in Table 18.3.

Table 18.3 Variances and covariances for four repeated measures

	y_1	y_2	y_3	y_4
y_1	σ_1^2	σ_{12}	σ_{13}	σ_{14}
y_2	σ_{21}	σ_2^2	σ_{23}	σ_{24}
y_3	σ_{31}	σ_{32}	σ_3^2	σ_{34}
y_4	σ_{41}	σ_{42}	σ_{43}	σ_4^2

Under the usual analysis of variance assumptions, we would expect the observations to be independent, and hence, the covariances in this case would all be zero.

The analysis of the data in Table 18.2 is carried out in MINITAB with the following commands.

```
MTB >  set c1
DATA> (1:5)4
DATA> end
MTB > set c2
DATA> 5(1:4)
DATA> end
MTB > set c3
DATA> 30 28 16 34 14 18 10 22
DATA> 24 20 18 30 38 34 20 44
DATA> 26 28 14 30
DATA> END

MTB > print c1-c3

Data Display

 Row  subjects    drug    score

   1         1       1       30
   2         1       2       28
   3         1       3       16
   4         1       4       34
   5         2       1       14
   6         2       2       18
   7         2       3       10
   8         2       4       22
   9         3       1       24
  10         3       2       20
  11         3       3       18
  12         3       4       30
  13         4       1       38
  14         4       2       34
  15         4       3       20
  16         4       4       44
  17         5       1       26
  18         5       2       28
  19         5       3       14
  20         5       4       30

Results for: repeat1.MTW

MTB > GLM 'score' = Subjects drug;
SUBC>   Random 'Subjects';
SUBC>   Brief 2 ;
SUBC>   EMS.
```

```
General Linear Model: score versus Subjects, drug

Factor    Type    Levels  Values
Subjects  random       5  1, 2, 3, 4, 5
drug      fixed        4  1, 2, 3, 4


Analysis of Variance for score, using Adjusted SS for Tests


Source    DF   Seq SS  Adj SS  Adj MS      F      P
Subjects   4   680.80  680.80  170.20  18.11  0.000
drug       3   698.20  698.20  232.73  24.76  0.000
Error     12   112.80  112.80    9.40
Total     19  1491.80


S = 3.06594   R-Sq = 92.44%   R-Sq(adj) = 88.03%


Expected Mean Squares, using Adjusted SS

            Expected Mean Square
   Source   for Each Term
1  Subjects (3) + 4.0000 (1)
2  drug     (3) + Q[2]
3  Error    (3)


Error Terms for Tests, using Adjusted SS

                                 Synthesis
   Source    Error DF  Error MS  of Error MS
1  Subjects     12.00      9.40  (3)
2  drug         12.00      9.40  (3)


Variance Components, using Adjusted SS

         Estimated
Source       Value
Subjects    40.200
Error        9.400
```

The calculated F value of 24.76 is greater than 3.49; hence, there are significant differences in the four drugs' means. Alternatively, since the p value $= 0.000 << 0.05$, we would, therefore, strongly reject the null, leading to the same conclusion. We observe immediately that the residual SS of 793.60 in model (18.1) has been reduced to 112.80 in model (18.2), a reduction of almost 86 %, which is attributable to the subject-to-subject variation.

The mean scores plot for the four drugs are presented in Fig. 18.1. Clearly, drug 4 has the highest mean while drug 3 has the lowest mean.

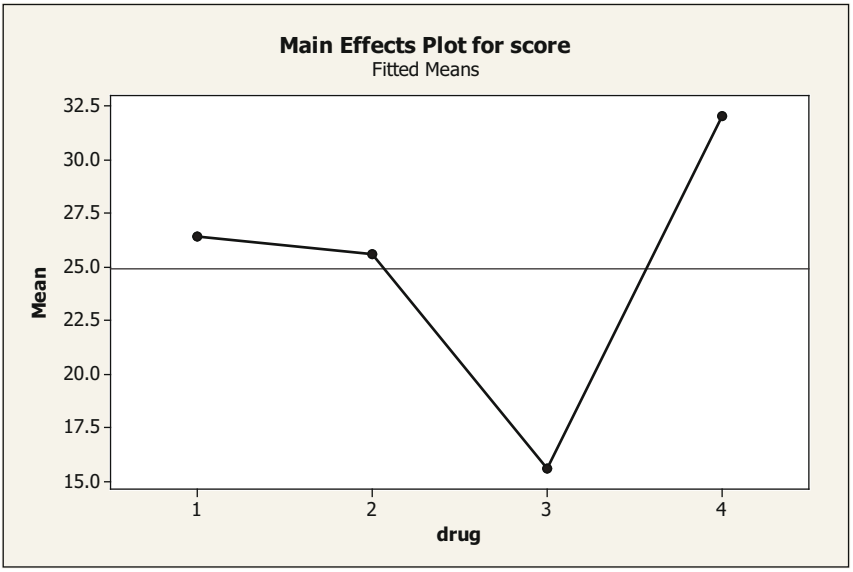


Fig. 18.1 Drugs mean score plots

We present below the implementation of the above analysis using the multivariate approach. The results displayed are equivalent.

```
MTB > ANOVA 'score' = subjects drug;
SUBC>   Random 'subjects';
SUBC>   MANOVA;
SUBC>   NoUnivariate.

ANOVA: score versus subjects, drug

MANOVA for subjects            s =  1    m = 1.0     n =    5.0

Criterion       Test Statistic          F           DF       P
-------------------------------------------------------------------
Wilk's                  0.14214     18.106  (  4,     12)  0.000
Lawley-Hotelling        6.03546     18.106  (  4,     12)  0.000
Pillai's                0.85786     18.106  (  4,     12)  0.000
Roy's                   6.03546


MANOVA for drug                s =  1    m = 0.5     n =    5.0

Criterion       Test Statistic          F           DF       P
-------------------------------------------------------------------
Wilk's                  0.13909     24.759  (  3,     12)  0.000
Lawley-Hotelling        6.18972     24.759  (  3,     12)  0.000
Pillai's                0.86091     24.759  (  3,     12)  0.000
Roy's                   6.18972
```

18.2.1 Correlation Within Subjects

The estimated correlation within subjects can be calculated as

$$\hat{\rho} = \frac{\sigma^2_{\text{subj}}}{\sigma^2_{\text{subj}} + \sigma^2_{\varepsilon}} = \frac{40.20}{40.20 + 9.40} = 0.810 \tag{18.3}$$

From the variance components estimates, we have $\sigma^2_{\text{subj}} = 40.20$ and $\hat{\sigma}^2_{\varepsilon} = 9.40$. Consequently, $\hat{\rho} = 0.810$, which clearly indicates high correlations within subjects. This would be the correlation structure under the compound symmetry assumption where the variance covariance structure now reduces to

Variances and covariances under the compound symmetry

	y_1	y_2	y_3	y_4
y_1	σ^2	$\rho\sigma^2$	$\rho\sigma^2$	$\rho\sigma^2$
y_2	$\rho\sigma^2$	σ^2	$\rho\sigma^2$	$\rho\sigma^2$
y_3	$\rho\sigma^2$	$\rho\sigma^2$	σ^2	$\rho\sigma^2$
y_4	$\rho\sigma^2$	$\rho\sigma^2$	$\rho\sigma^2$	σ^2

18.3 Two Factors with Repeated Measures on One Factor

For a two-factor case, consider the example below to evaluate the effect of a new vaccine on discomfort due to arthritis. Here, multiple measurements (month1, month2, and month3) are taken on each subject, resulting in a two-factor experiment (vaccine and visits) with repeated measures taken over one of the factors (visits).

Example 18.3.1

A pilot study was conducted on eight patients to evaluate the effect of a new vaccine on discomfort due to arthritic outbreaks. Four patients were randomly assigned to receive an active vaccine and four to receive a placebo. Patients were asked to return to the clinic monthly for 3 months and evaluate their comfort level with routine daily chores during the preceding month on a scale of 0 (no discomfort) to 10 (maximum discomfort). Eligibility criteria required patients to have a rating of at least an 8 in the month prior to vaccination. The rating data are displayed in Table 18.4. Is there evidence of a difference in response profiles between the active and placebo vaccines?

Table 18.4 Data for Example 18.1.2

		Visits		
Vaccine	Subjects	Month 1	Month 2	Month 3
Active	101	6	3	0
	103	7	3	1
	104	4	1	2
	107	8	4	3
Placebo	102	6	5	5
	105	9	4	6
	106	5	3	4
	108	6	2	3

The analysis of the above data is similar to a split-plot analysis discussed in Chap. 15. Hence, the model specification is similar. The model for the above data is given by

$$y_{ijk} = \mu + v_i + e_{ik} + m_j + (vm)_{ij} + \epsilon_{ijk}$$
$$i = 1, 2,\ j = 1, 2, 3,\ k = 1, 2, 3 \tag{18.4}$$

where μ is the general mean, v_i is the effect of the ith vaccine, e_{ik} is the random error term for subjects within treatments (vaccines) with variance σ_e^2, m_j is the effect of the jth month, $(vm)_{ij}$ is the interaction between vaccines and and month of visit, and ϵ_{ijk} is the identically normal distributed random error term on repeated measures with variance σ^2.

The table of mean scores is displayed as

	Visiting months			
Vaccines	1	2	3	Mean
Active	6.25	2.75	1.50	3.52
Placebo	6.50	3.50	4.5	4.83
Mean	6.38	3.13	3.00	4.17

The observed profile mean scores for each vaccine at each month of visit are presented in Fig. 18.2.

The MINITAB implementation of the analysis of the data is presented in what follows:

```
MTB > print c1-c4
Data Display

 Row   VAC  subjects  visit  score
 -------------------------------
   1     1       101      1      6
   2     1       101      2      3
   3     1       101      3      0
   4     1       103      1      7
   5     1       103      2      3
   6     1       103      3      1
```

```
 7     1       104      1      4
 8     1       104      2      1
 9     1       104      3      2
10     1       107      1      8
11     1       107      2      4
12     1       107      3      3
13     2       102      1      6
14     2       102      2      5
15     2       102      3      5
16     2       105      1      9
17     2       105      2      4
18     2       105      3      6
19     2       106      1      5
20     2       106      2      3
21     2       106      3      4
22     2       108      1      6
23     2       108      2      2
24     2       108      3      3

MTB > GLM 'score' = VAC subjects(vac) visit vac*visit;
SUBC>   Random 'subjects';
SUBC>   Brief 1 ;
SUBC>   Means VAC visit.

General Linear Model: score versus VAC, visit, subjects


Factor          Type Levels Values
VAC            fixed      2 1 2
subjects(VAC) random      8 101 103 104 107 102 105 106 108
visit          fixed      3 1 2 3

Analysis of Variance for score, using Adjusted SS for Tests

Source          DF     Seq SS     Adj SS     Adj MS       F      P
VAC              1     10.667     10.667     10.667    2.53  0.163
subjects(VAC)    6     25.333     25.333      4.222    4.16  0.017
visit            2     58.583     58.583     29.292   28.89  0.000
VAC*visit        2      8.583      8.583      4.292    4.23  0.041
Error           12     12.167     12.167      1.014
Total           23    115.333

Least Squares Means for score

VAC           Mean
1            3.500
2            4.833
visit
1            6.375
2            3.125
3            3.000
```

Our analysis shows that the interactions between vaccines and visits are significant. Further, it also shows that there are no significant differences between the two vaccines, although vaccine 2 shows a higher discomforting effect on patients. Similarly, for the visits, visit 1 clearly shows the most significant difference in discomforting scores among patients. The second and third visit scores are not significant. These results are displayed in Fig. 18.2.

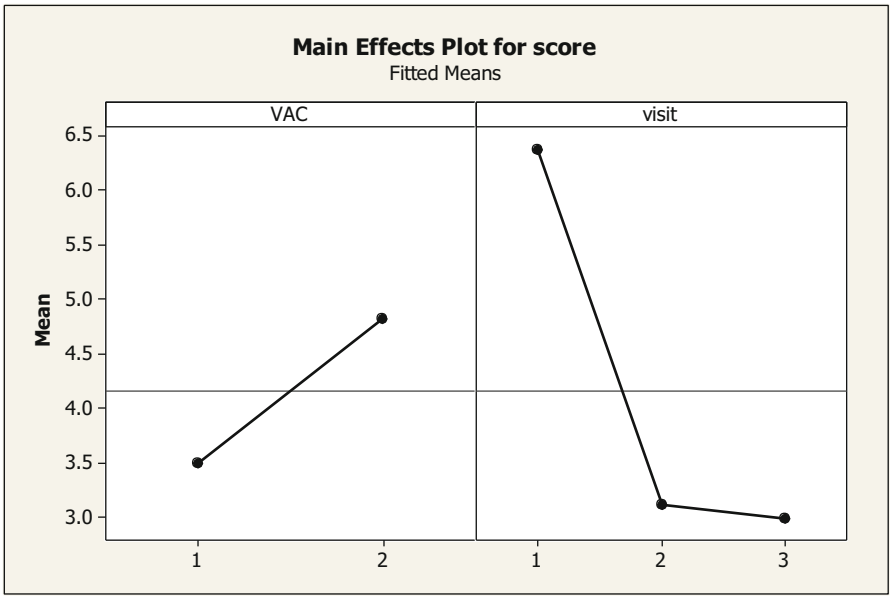


Fig. 18.2 Vaccines and visits main effects means plots

The corresponding interaction plots are also presented in Fig. 18.3.

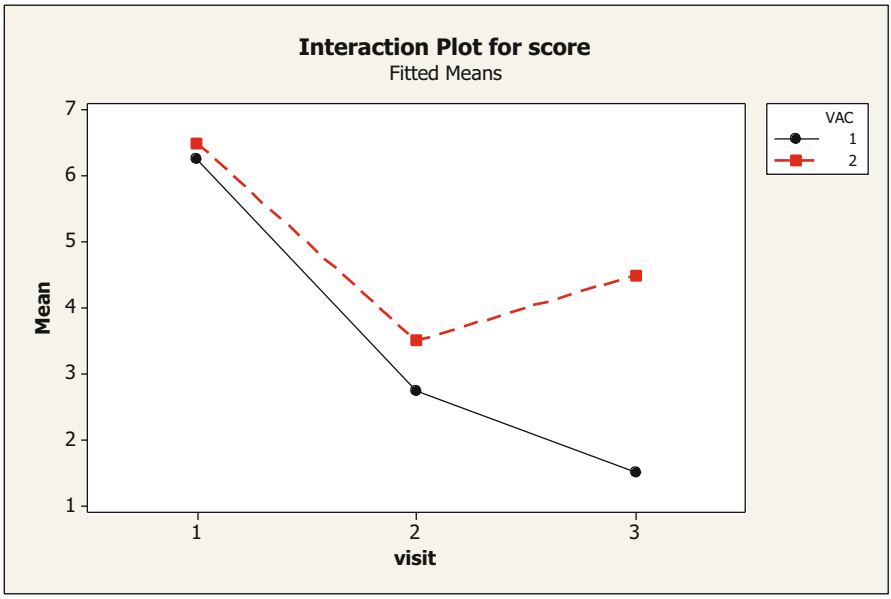


Fig. 18.3 Interaction plot of vaccines and visits

18.3.1 Calculations

The calculations of the sum of squares displayed by MINITAB are computed as follows from Table 18.4. We may note here that the F value computed for vaccine effect is not correct. The correct value is displayed in the calculated ANOVA table below. The analysis is the familiar split-plot design analysis.

$$\text{Total SS} = 6^2 + 3^2 + \cdots + 3^2 - \frac{100^2}{24} = 115.3333$$

$$\text{Subj SS} = \frac{9^2 + 11^2 + \cdots + 11^2}{3} - \frac{100^2}{24} = 25.3333$$

$$\text{Visit SS} = \frac{51^2 + 25^2 + 24^2}{8} - \frac{100^2}{24} = 58.5833$$

$$\text{Vac SS} = \frac{42^2 + 58^2}{12} - \frac{100^2}{24} = 10.6667$$

$$\text{int SS}^* = \frac{25^2 + 11^2 + \cdots + 18^2}{4} - \frac{100^2}{24} = 77.8333$$

Hence,

$$\begin{aligned}\text{Vac} \times \text{Visit SS} &= (\text{int SS}) - \text{Vac SS} - \text{Visit SS} \\ &= 77.8333 - 10.6667 - 58.5833 = 8.5833\end{aligned}$$

Source	d.f.	SS	MS	F
Between subjects	7	25.3333		
Vaccine	1	10.6667	10.6667	2.53 ns
Subj (vac)	6	25.3333	4.2222	
Visit	2	58.5833	29.2917	28.89**
Vac*visit	2	8.5833	4.2917	4.23*
Error	12	12.1667	1.0139	
Total	23	115.3333		

* Significance at 5%
** Significance at 1%

18.3.2 Multivariate Approach

The analysis above is the univariate approach to repeated-measures design data. Sometimes, this approach may not be enough. The univariate analysis assumes what we call *sphericity* assumption. This assumption assumes, in our example here for instance, that the three correlations r_{12}, r_{13}, and r_{23} between the three visits are all about the same in size, and that any differences observed can only be due to sampling variation. The sphericity test is often accomplished with Mauchly's test of sphericity, and if the assumption is violated, a correction factor called *epsilon* is usually applied by most

statistical software to the error degrees of freedom before calculating the F value. Other softwares use the Greenhouse–Geisser epsilon or the Huynh–Feldt epsilon. Values of these close to one indicate that the assumption is not being violated. Smaller values indicate that the assumption is being violated and adjustments need to be made to the error degrees of freedom. In this example, both tests give $\hat{\epsilon} = 0.9567$ and 1.3941 respectively.

To implement the multivariate approach to repeated measures data, we first transform the data in this example to the format below together with the relevant MINITAB command to perform the analysis.

```
MTB > print c1-c5

Data Display

Row  vac  subj  y1  y2  y3
  1    1   101   6   3   0
  2    1   103   7   3   1
  3    1   104   4   1   2
  4    1   107   8   4   3
  5    2   102   6   5   5
  6    2   105   9   4   6
  7    2   106   5   3   4
  8    2   108   6   2   3

MTB > GLM 'y1' 'y2' 'y3' = vac;
SUBC>   MANOVA vac / Error;
SUBC>   NoUnivariate;
SUBC>   Means vac.

General Linear Model: y1, y2, y3 versus vac

MANOVA for vac
s = 1    m =  0.5   n = 1.0

                      Test             DF
Criterion         Statistic      F  Num  Denom      P
Wilks'              0.30969  2.972    3      4  0.160
Lawley-Hotelling    2.22906  2.972    3      4  0.160
Pillai's            0.69031  2.972    3      4  0.160
Roy's               2.22906
```

The results presented indicate, by examining the appropriate p values that there is no significant difference between the vaccines scores. The p value is similar to the earlier analysis, which is often referred to as "between-subjects" analysis.

Example 18.3.2

A study was conducted on human subjects to measure the effects of three foods on serum glucose levels. Each of the three foods was randomly assigned to four subjects. The serum glucose mass was measured for each of the

subjects at 15, 30, and 45 min after the food was ingested. The data are displayed in Table 18.5.

We can analyze the data in MINITAB by first entering the data and noting that the levels of time are equally spaced; thus, we can fit a quadratic model to both main effect and interaction terms in the model. We present the MINITAB statements below.

Table 18.5 Serum glucose levels for this example

		Time (min)		
Diets	Subjects	15	30	45
1	1	28	34	32
	2	15	29	27
	3	12	33	28
	4	21	44	39
2	5	22	18	12
	6	23	22	10
	7	18	16	9
	8	25	24	15
3	9	31	30	39
	10	28	27	36
	11	24	26	36
	12	21	26	32

```
MTB > set c1
DATA> (1:3)12
DATA> end
MTB > set c2
DATA> (1:12)3
DATA> end
MTB > set c3
DATA> 12(1:3)
DATA> end
MTB > set c4
DATA> 28 34 32 15 29 27 12 33 28 21 44 39
DATA> 22 18 12 23 22 10 18 16 9 25 24 15
DATA> 31 30 39 28 27 36 24 26 36 21 26 32
DATA> end
MTB > print c1-c4

Data Display


 Row   Diet  subjects   Time   mass

   1      1         1      1     28
   2      1         1      2     34
   3      1         1      3     32
   4      1         2      1     15
   5      1         2      2     29
   6      1         2      3     27
```

```
  ..............................
 27     3         9      3     39
 28     3        10      1     28
 29     3        10      2     27
 30     3        10      3     36
 31     3        11      1     24
 32     3        11      2     26
 33     3        11      3     36
 34     3        12      1     21
 35     3        12      2     26
 36     3        12      3     32

MTB > GLM 'mass' = Diet subjects(Diet) Time Diet*Time;
SUBC>   Random 'subjects';
SUBC>   Brief 2 ;
SUBC>   Means Diet Time Diet*Time;
SUBC>   Pairwise Diet Time;
SUBC>     Tukey;
SUBC>     NoCI.

General Linear Model: mass versus Diet, Time, subjects

Factor           Type Levels Values
Diet            fixed      3 1 2 3
subjects(Diet) random     12  1  2  3  4  5  6  7  8  9 10 11 12
Time            fixed      3 1 2 3

Analysis of Variance for mass, using Adjusted SS for Tests

Source          DF     Seq SS     Adj SS     Adj MS       F      P
Diet             2    1020.67    1020.67     510.33   11.11  0.004
subjects(Diet)   9     413.33     413.33      45.93    6.45  0.000
Time             2     170.17     170.17      85.08   11.95  0.000
Diet*Time        4     869.67     869.67     217.42   30.53  0.000
Error           18     128.17     128.17       7.12
Total           35    2602.00

Variance Components, using Adjusted SS

              Estimated
Source            Value
subjects(Diet)   12.935
Error             7.120

Least Squares Means for mass

Diet        Mean
1          28.50
2          17.83
3          29.67

Time
1          22.33
2          27.42
3          26.25
```

```
Diet*Time
1    1          19.00
1    2          35.00
1    3          31.50
2    1          22.00
2    2          20.00
2    3          11.50
3    1          26.00
3    2          27.25
3    3          35.75


MTB > GLM  'mass' = Diet subjects( Diet)   Time Diet* Time;
SUBC>   Random 'subjects';
SUBC>   Brief 2 ;
SUBC>   EMS;
SUBC>   Means Diet Time Diet* Time.

General Linear Model: mass versus Diet, Time, subjects

Factor          Type    Levels  Values
Diet            fixed        3  1, 2, 3
subjects(Diet)  random      12  1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
Time            fixed        3  1, 2, 3


Analysis of Variance for mass, using Adjusted SS for Tests

Source          DF   Seq SS   Adj SS  Adj MS      F      P
Diet             2  1020.67  1020.67  510.33  11.11  0.004
subjects(Diet)   9   413.33   413.33   45.93   6.45  0.000
Time             2   170.17   170.17   85.08  11.95  0.000
  Tl             1    92.04    92.04   92.04  12.93  0.002
  Tq             1    78.12    78.12   78.12  10.97  0.004
Diet*Time        4   869.67   869.67  217.42  30.53  0.000
  Diet*Tl        2   631.08   631.08  315.54  44.32  0.000
  Diet*Tq        2   238.58   238.58  119.29  16.75  0.000
Error           18   128.17   128.17    7.12
Total           35  2602.00


S = 2.66840   R-Sq = 95.07%   R-Sq(adj) = 90.42%


Variance Components, using Adjusted SS

                Estimated
Source              Value
subjects(Diet)     12.935
Error               7.120
```

The table of mean scores for the interaction is displayed in the following:

	Time			
Diets	1	2	3	Mean
1	19.00	35.00	31.50	28.50
2	22.00	20.00	11.50	17.83
3	26.00	27.25	35.75	29.67
Mean	22.33	27.42	26.25	25.33

We observe that the levels for time are equally spaced; hence, we can partition the effects of time into both linear and quadratic components, each based on 1 d.f. Similar partitioning can be made for the Diet*Time interaction leading

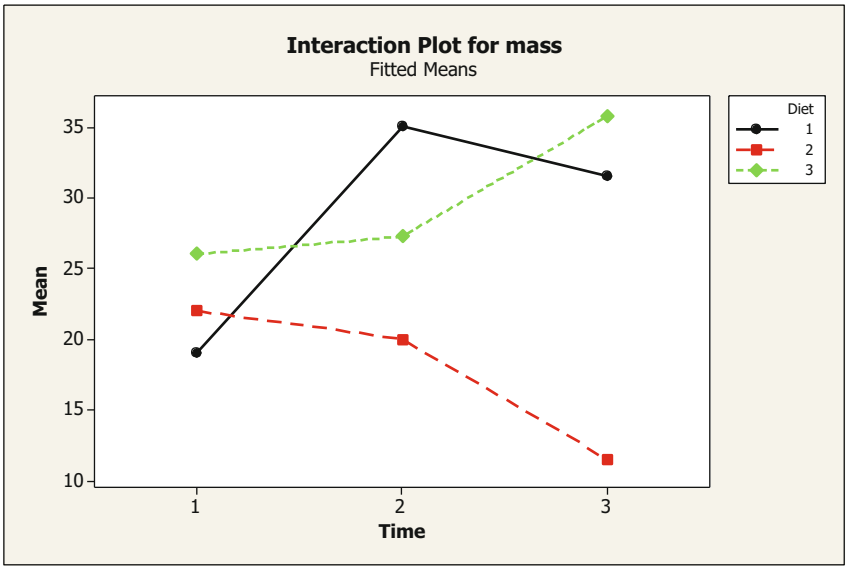


Fig. 18.4 Diet and time interaction plot

to 2 d.f. each for these components. The components are designated TL, TQ, D*TL, and D*TQ, respectively in the above MINITAB ANOVA table, and are generated with the following statements in MINITAB.

```
MTB > code (1) -1 (2) 0 (3) 1 c3 c5 -----TL (linear)
MTB > code (1) 1 (2) -2 (3) 1 c3 c6 -----TQ (quadratic)

MTB > GLM 'mass' = Diet subjects( Diet) Tl Tq Diet* Tl Diet * Tq;
SUBC>   Covariates 'Tl' 'Tq';
SUBC>   Random 'subjects';
SUBC>   Brief 2 .
```

Results

(a) For Diet 1: There is an increase in mass from time 1 to time 2, and then a drop in mass between time 2 and time 3.
(b) For Diet 2: There is a decrease from time 1 to time 2, and then from time 2 to time 3. Thus, there are significant mass losses in succession from time 1 to time 3.
(c) For Diet 3: There are significant increases from time 1 to time 2 and again from time 2 to time 3 (Fig. 18.4).

The relationship between the interaction between diet and time is best described by a quadratic model.

Example 18.3.3

This example is data on arthritis from a study involving a two-factor type joint problem (hip or shoulder), dose (drug administered) at three levels, and motion measurements recorded for each subject at four different times. The author unfortunately lost the original source for this data but is, hereby, duly acknowledged. The responses are motion scores recorded over a period of 10 h. The data are presented in Table 18.6.

We use MINITAB to analyze the data, but we present a partial output for the data to indicate how the data was read in.

Table 18.6 Arthritis pain data repeated over over 12 combinations of dose and time

			Time			
Joint	Dose	Subject	2	4	6	10
Hip	1.0	1	27	32	39	28
		2	29	31	36	21
		3	37	44	47	33
	2.0	4	38	44	53	43
		5	31	34	41	35
		6	53	55	58	44
	3.0	7	53	55	60	49
		8	42	47	48	43
		9	64	64	69	62
Shoulder	1.0	10	23	31	33	19
		11	17	28	31	20
		12	27	37	40	27
	2.0	13	33	41	48	43
		14	26	30	37	32
		15	38	44	49	33
	3.0	16	47	50	48	53
		17	43	42	45	47
		18	58	56	60	61

```
Data Display

Row   Joint   Dose   Subj   Time   y
  1       1      1      1      2  27
  2       1      1      1      4  32
  3       1      1      1      6  39
  4       1      1      1     10  28
  5       1      1      2      2  29
  6       1      1      2      4  31
  7       1      1      2      6  36
  8       1      1      2     10  21
  9       1      1      3      2  37
 10       1      1      3      4  44
 11       1      1      3      6  47
 12       1      1      3     10  33
 ------------------------------------
 61       2      4     16      2  47
 62       2      4     16      4  50
 63       2      4     16      6  48
 64       2      4     16     10  53
 65       2      4     17      2  43
 66       2      4     17      4  42
 67       2      4     17      6  45
 68       2      4     17     10  47
 69       2      4     18      2  58
 70       2      4     18      4  56
 71       2      4     18      6  60
 72       2      4     18     10  61

MTB > GLM 'y' = Joint Dose Joint* Dose Subj( Joint Dose) Time Joint* Time &
CONT>      Dose* Time Joint* Dose* Time;
SUBC>   Random 'Subj';
SUBC>   Brief 2 ;
SUBC>   EMS;
SUBC>   Means Joint* Time Dose* Time.

General Linear Model: y versus Joint, Dose, Time, Subj

Factor           Type    Levels  Values
Joint            fixed        2  1, 2
Dose             fixed        3  1, 2, 4
Subj(Joint Dose) rando       18  1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,
                                 14, 15, 16, 17, 18
Time             fixed        4  2, 4, 6, 10

Analysis of Variance for y, using Adjusted SS for Tests

Source            DF    Seq SS   Adj SS   Adj MS       F      P
Joint              1    512.00   512.00   512.00    2.46  0.143
Dose               2   5839.53  5839.53  2919.76   14.01  0.001
Joint*Dose         2     20.58    20.58    10.29    0.05  0.952
Subj(Joint Dose) 12   2501.33  2501.33   208.44   32.34  0.000
Time               3    888.06   888.06   296.02   45.93  0.000
Joint*Time         3     53.67    53.67    17.89    2.78  0.055
Dose*Time          6    256.36   256.36    42.73    6.63  0.000
Joint*Dose*Time    6     66.42    66.42    11.07    1.72  0.145
Error             36    232.00   232.00     6.44
Total             71  10369.94
```

```
S = 2.53859   R-Sq = 97.76%   R-Sq(adj) = 95.59%

Analysis of Variance for y, using Adjusted SS for Tests

Source           DF    Seq SS   Adj SS   Adj MS      F      P
Joint             1    512.00   512.00   512.00   2.46  0.143
Dose              2   5839.53  5839.53  2919.76  14.01  0.001
Joint*Dose        2     20.58    20.58    10.29   0.05  0.952
Subj(Joint Dose) 12   2501.33  2501.33   208.44  32.34  0.000
Time              3    888.06   888.06   296.02  45.93  0.000
Joint*Time        3     53.67    53.67    17.89   2.78  0.055
Dose*Time         6    256.36   256.36    42.73   6.63  0.000
Joint*Dose*Time   6     66.42    66.42    11.07   1.72  0.145
Error            36    232.00   232.00     6.44
Total            71  10369.94

S = 2.53859   R-Sq = 97.76%   R-Sq(adj) = 95.59%


Variance Components, using Adjusted SS

                Estimated
Source              Value
Subj(Joint Dose)   50.500
Error               6.444


Least Squares Means for y

Joint*Time   Mean
1      2    41.56
1      4    45.11
1      6    50.11
1     10    39.78
2      2    34.67
2      4    39.89
2      6    43.44
2     10    37.22
Dose*Time
1      2    26.67
1      4    33.83
1      6    37.67
1     10    24.67
2      2    36.50
2      4    41.33
2      6    47.67
2     10    38.33
4      2    51.17
4      4    52.33
4      6    55.00
4     10    52.50
```

Results

1. The analysis of variance table for the data is presented above in the MINITAB output.
2. The joint type effects are not significant.

3. The dose effects are significant, and so is the interaction between dose and time at 5 % level of significance.
4. The time effects are also significantly different.
5. The dose, time, and dose*time interaction effects can be partitioned into various components since their levels are chosen for the appropriate utilization of orthogonal polynomials.
6. However, the interaction display in Fig. 18.5 shows that the motion increases for each dose level. For dose levels 1 and 2, motions increase steadily until about the sixth hour and then drop sharply for both levels of drugs.
7. The third level of dose sees steady increase, and the drop is not very sharp.

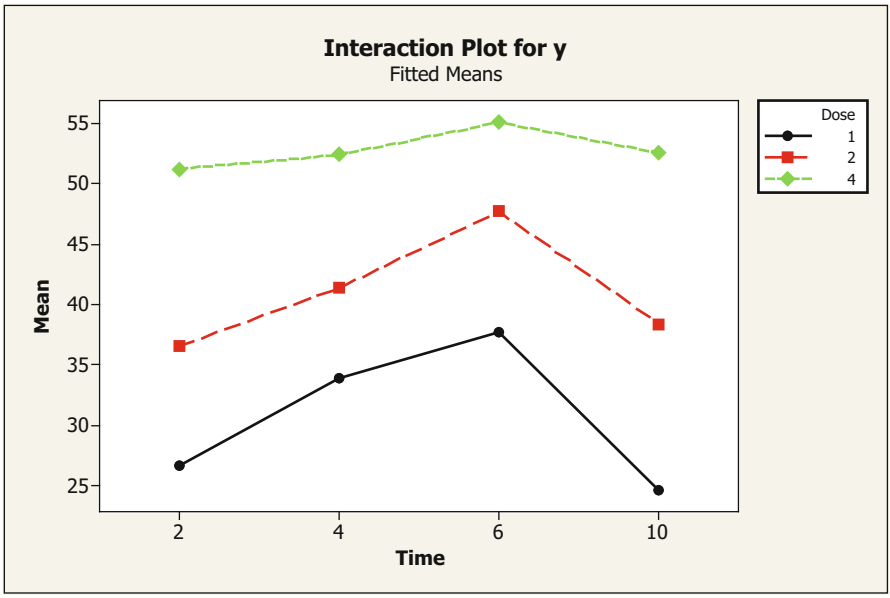


Fig. 18.5 Interaction plot of dose and time

18.4 Exercises

1. The following exercise is adapted and relates to measured pulses of subjects at three intensity levels: (1) taking at the warm-up exercising trial, (2) pulse measurement taken after running. The factors of interest are the variable diet, which denotes dietary preference, with values of 1 signifying meat eaters and 2 signifying vegetarians and variable exertype, which is the type of exercise assigned to the subjects, with 1 signifying aerobic stairs, 2 signifying racquet ball, and 3 signifying weight training.

			Intensity		
Exertype	Diet	Subject	1	2	3
1	1	1	112	166	215
		2	111	166	225
		3	89	132	189
	2	4	95	134	186
		5	66	109	150
		6	69	119	177
2	1	7	125	177	241
		8	85	117	186
		9	97	137	185
	2	10	93	151	217
		11	77	122	178
		12	78	119	173
3	1	13	81	134	205
		14	88	133	180
		15	88	157	224
	2	16	58	99	131
		17	85	132	186
		18	78	110	164

The observations are, thus, repeated three times. Analyze the above data as a repeated-measures design using the univariate approach. Draw your conclusions.

2. The following is taken with permission from Rao (1998) and relates to the study of a new drug on total cholesterol of subjects measured at six 4-week periods. (Hirotsu 1993).

		Period					
Treatment	Subject	1	2	3	4	5	6
Drug	1	317	280	275	270	274	266
	2	186	189	190	135	197	205
	3	377	395	368	334	338	334
	4	229	258	282	272	264	265
	5	276	310	306	309	300	264
	6	272	250	250	255	228	250
	7	219	210	236	239	242	221
	8	260	245	264	268	317	314
	9	284	256	241	242	243	241
	10	365	304	294	287	311	302
	11	298	321	341	342	357	335
	12	274	245	262	263	235	246

Placebo	13	232	205	244	197	218	233
	14	367	354	358	333	338	355
	15	253	256	247	228	237	235
	16	230	218	245	215	230	207
	17	190	188	212	201	169	179
	18	290	263	291	312	299	279
	19	337	337	383	318	361	341
	20	283	279	277	264	269	271
	21	325	257	288	326	293	275
	22	266	258	253	284	245	263
	23	338	343	307	274	262	309

(a) Why would the data above be considered as a repeated-measure study?
(b) Analyze the data and draw your conclusions.

Chapter 19
Survival Analysis

19.1 Introduction

In the health sciences, survival analysis is often used to model the time duration until the occurrence of an event—usually death (often referred to as the *survival time*). These survival time durations arise as a result of subjects being followed over time until they reach a specified endpoint or the event of interest occurs. An example of this is time to death of females who are diagnosed with breast cancer. Here, the event is death. Another example is the length of time a particular disease in humans remains in remission. The distribution of survival times is often skewed to the right and analysis often focuses on the probability that the individual survives for a given length of time. In time to event studies, subjects often leave the study either through death or are lost through follow-up or willingly leave the study. In other situations, some patients are not followed until death because of the expiration of the study at a specified time. Censoring occurs when an event of interest (e.g., remission, death, recovery) has not occurred by the time observations were made, so that all we knew at that point in time is that, the individual has survived at least up to some time). Thus censoring can not be glossed over as they carry important information about the factor of interest.

Survival analysis has also found applications in engineering where the time to failure (often described as the accelerated failure time model) of a component is of interest. Another area that survival analysis has found use is in the behavioral sciences, particularly in the study of recidivism, involving the duration (in months or weeks) when prisoners are released and rearrested.

19.2 Censoring

Censoring introduces complications to the statistical analysis of survival data. It is, however, important to distinguish even if we are not going to treat the various censoring schemes here. We describe briefly these schemes below.

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_19,

(i) **Right Censoring:** It occurs when we do not yet observe the event of interest at the end of the study (time t) but we do know that the event occurs after time t.
(ii) **Left Censoring:** This is a situation when we do know that the event of interest occurs at some time t_0 which is less than t. This often occurs when observations are obtained on patients at fixed appointment times (say, every 3 months), and that only at the next appointment do we realize that the event (such as death) has occurred some time between the last visit and the current visit. So, survival time is left than the observation time t.
(iii) **Interval Censoring:** This occurs when the event is known to have occurred during an interval.

19.3 Describing Event Times

If we use T to denote the survival time, then the survival function, designated as $S(t)$ is defined as the probability that an individual survives past time t. That is,

$$S(t) = \Pr(T > t) = 1 - F(t), \tag{19.1}$$

where

$$F(t) = \Pr(T \leq t),$$

is the cumulative distribution function. The graph of $S(t)$ against t is called the the *survivor curve*.

19.4 Estimating the Survival Function $S(t)$

From (19.1), we have $S(t) = 1 - F(t)$, hence, $0 < S(t) < 1$. That is, $S(t)$ is a decreasing function of t. For situations involving censored data, survival functions or curves can be estimated by the method of *product limit* or the *Kaplan–Meier (KM) estimator* and the *life table method*.

19.4.1 The Kaplan–Meier Method

We illustrate the Kaplan–Meier method with the following example taken from Sedmak et al. (1989) which relates to female breast cancer patients originally classified as lymph node-negative by standard light microscopy

(SLM). The data in Table 19.1 give the times to death in months of a random sample of 45 female breast cancer patients with a minimum of 10 years follow-up from the Ohio State University Hospitals Cancer Registry. Of these 45 patients, 36 were immunoperoxidase-negative while the remaining 9 were positive. A status of 0 denotes a censored observation, that is, patients lost to follow-up, or patient was withdrawn alive.

Table 19.1 Times to death for 45 breast cancer patients

Immunoperoxidase-negative									Immunoperoxidase-positive		
Sub.	Time	Status	Sub.	Time	Status	Sub.	Time	Status	Sub.	Time	Status
1	19	1	13	67	1	25	143	0	37	22	1
2	25	1	14	74	1	26	148	0	38	23	1
3	30	1	15	78	1	27	151	0	39	38	1
4	34	1	16	86	1	28	152	0	40	42	1
5	37	1	17	122	0	29	153	0	41	73	1
6	46	1	18	123	0	30	154	0	42	77	1
7	47	1	19	130	0	31	156	0	43	89	1
8	51	1	20	130	0	32	162	0	44	115	1
9	56	1	21	133	0	33	164	0	45	144	0
10	57	1	22	134	0	34	165	0			
11	61	1	23	136	0	35	182	0			
12	66	1	24	141	0	36	189	0			

The KM method can be implemented in MINITAB by utilizing the nonparametric KM method. This is accomplished with the following menu selection in MINITAB.

Stat > Reliability/Survival > Distribution Analysis (Right Censoring) > Nonparametric Distribution Analysis

We present the corresponding MINITAB statements and partial output for the analysis of the data in Table 19.1. We note here that we changed the days for patient 20 from 130 to 131 to avoid ties and our initial analysis ignores the treatment effects.

We present a partial MINITAB output for the analysis of the data.

```
\scriptsize
\begin{verbatim}
```

Obs	tsurv	censor	trt
1	19	1	0
2	25	1	0
3	30	1	0
4	34	1	0
5	37	1	0
6	46	1	0
7	47	1	0
8	51	1	0
9	56	1	0

10	57	1	0
11	61	1	0
12	66	1	0
13	67	1	0
14	74	1	0
15	78	1	0
16	86	1	0
17	122	0	0
18	123	0	0
19	130	0	0
20	131	0	0
21	133	0	0
22	134	0	0
23	136	0	0
24	141	0	0
25	143	0	0
26	148	0	0
27	151	0	0
28	152	0	0
29	153	0	0
30	154	0	0
31	156	0	0
32	162	0	0
33	164	0	0
34	165	0	0
35	182	0	0
36	189	0	0
37	22	1	1
38	23	1	1
39	38	1	1
40	42	1	1
41	73	1	1
42	77	1	1
43	89	1	1
44	115	1	1
45	144	0	1

```
MTB > Ltest 'surv';
SUBC>   Noparametric;
SUBC>   Splot;
SUBC>   CFPlot;
SUBC>   Hplot;
SUBC>   Xminimum 0;
SUBC>   Brief 2;
SUBC>   KMEstimates;
SUBC>   Confidence 95.0;
SUBC>   TwoCI;
SUBC>   Censor 'censor';
SUBC>   Cvalue 0.

Distribution Analysis: surv

Variable: surv

Censoring Information  Count
Uncensored value          24
Right censored value      21

Censoring value: censor = 0


Nonparametric Estimates


Characteristics of Variable

           Standard   95.0% Normal CI
Mean(MTTF)     Error    Lower    Upper
   117.378   10.5414  96.7170  138.039

Median = 89
IQR = *  Q1 = 51  Q3 = *
```

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95.0% Normal CI Lower	Upper
19	45	1	0.977778	0.0219739	0.934710	1.00000
22	44	1	0.955556	0.0307207	0.895344	1.00000
23	43	1	0.933333	0.0371849	0.860452	1.00000
25	42	1	0.911111	0.0424232	0.827963	0.99426
30	41	1	0.888889	0.0468486	0.797067	0.98071
34	40	1	0.866667	0.0506745	0.767347	0.96599
37	39	1	0.844444	0.0540284	0.738551	0.95034
38	38	1	0.822222	0.0569937	0.710517	0.93393
42	37	1	0.800000	0.0596285	0.683130	0.91687
46	36	1	0.777778	0.0619748	0.656309	0.89925
47	35	1	0.755556	0.0640644	0.629992	0.88112
51	34	1	0.733333	0.0659218	0.604129	0.86254
56	33	1	0.711111	0.0675660	0.578684	0.84354
57	32	1	0.688889	0.0690122	0.553627	0.82415
61	31	1	0.666667	0.0702728	0.528934	0.80440
66	30	1	0.644444	0.0713576	0.504586	0.78430
67	29	1	0.622222	0.0722744	0.480567	0.76388
73	28	1	0.600000	0.0730297	0.456864	0.74314
74	27	1	0.577778	0.0736283	0.433469	0.72209
77	26	1	0.555556	0.0740741	0.410373	0.70074
78	25	1	0.533333	0.0743698	0.387571	0.67910
86	24	1	0.511111	0.0745172	0.365060	0.65716
89	23	1	0.488889	0.0745172	0.342838	0.63494
115	22	1	0.466667	0.0743698	0.320905	0.61243

From the above KM estimates, at 56 months, for instance, the survival probability is 0.7111, indicating that the estimated probability that a patient will survive for 56 or more months is 0.7111. Similarly, the estimated survival probability for any time from 56 months up to (but not including) 57 months is 0.6889. We note that after 115 months, the largest censoring time, the KM estimate is undefined. The median death time, provided by the 50th percentile (labeled quantiles) is 89 months.

19.4.2 Computing Survival Probabilities

We observe that patient 1 lived for 19 months before she died. In general, a survival time of t months implies that a patient survived until time t. Thus, one of the patients died at 19 months, and the proportion of patients dying in this period is estimated as

$$1q_{19} = \frac{1}{45} = 0.0222$$

The proportion who survived in this period is, therefore

$$\hat{S}(19) = 1 - 1q_{19} = 1 - 0.0222 = 0.9778$$

The estimated $\hat{S}(19)$ and $\hat{S}(22)$ are obtained from the multiplicative rule of probability, where

$$P(A \cap B) = P(A)P(B|A)$$

where $P(A)$ represents the probability that the patient is alive during the period 0–18 months and $P(B)$ the probability that the patient survives at time t. The probability, therefore, that patients survive longer than 18 months is $P(A \cap B)$. We may note here that both $\hat{S}(0) = \hat{S}(18) = 1$.

One of the remaining 44 patients died at 22 months, and therefore

$$1q_{22} = \frac{1}{44} = 0.0227$$

The probability of living longer than 22 months is, therefore, estimated as:

$$\hat{S}(22) = \hat{S}(19)[1 - 1q_{22}] = (0.9778)(0.9773) = 0.9556.$$

Similarly, one of the remaining 43 patients died at 23 months, and therefore

$$1q_{23} = \frac{1}{43} = 0.0233.$$

The probability of living longer than 23 months is therefore estimated as

$$\hat{S}(23) = \hat{S}(22)[1 - 1q_{23}] = (0.9556)(0.9767) = 0.9333.$$

The other surviving probabilities are computed similarly. These estimated probabilities are presented in the column labeled "survival probability" in the MINITAB output above.

In the above MINITAB statements, we request that plots be made of *S-Splot* and *CF-CFplot*, the estimated survival and cumulative survival functions, respectively, against time. These are displayed in Fig. 19.1. The LS (the negative log of the estimated survival functions against time) gives us an idea as to whether the distribution of days follows an exponential or a Weibull distribution.

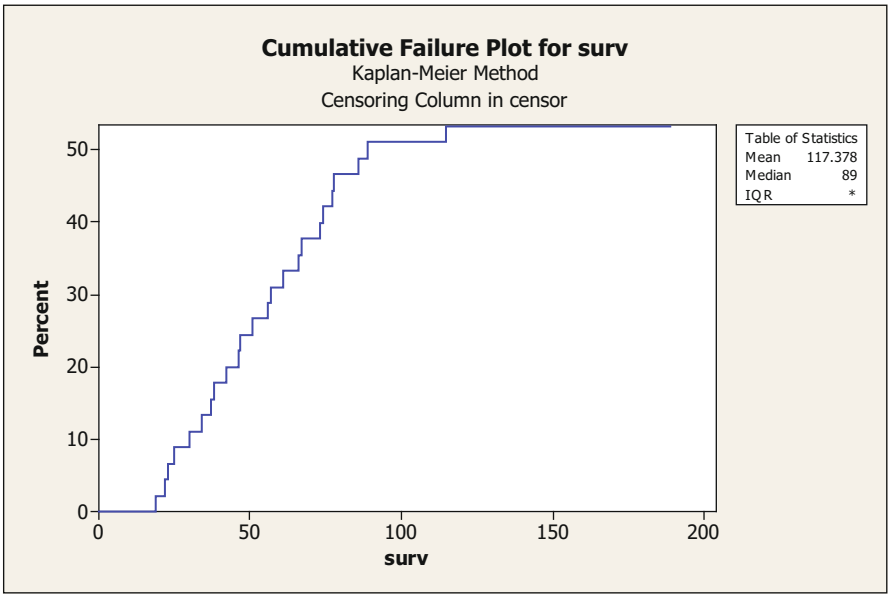
19.4.3 The Life Table Method

This method is usually suitable for large data sets if event times are precisely measured. The method is also often referred to as the *actuarial method* and is implemented as follows in MINITAB.

Stat > Reliability/Survival > Distribution Analysis (Right Censoring) > Nonparametric Distribution Analysis > Estimate (actuarial-and specify beginning and end points)

The latter part of the above statement is the *Intby 200 20;* in the MINITAB statements below.

```
MTB > Ltest 'surv';
SUBC>   Noparametric;
SUBC>   Splot;
SUBC>   CFPlot;
SUBC>   Xminimum 0;
SUBC>   Brief 2;
SUBC>   Intby 200 20;
SUBC>   Confidence 95.0;
SUBC>   TwoCI;
SUBC>   Censor 'censor';
SUBC>   Cvalue 0.

Distribution Analysis: surv

Variable: surv

Censoring Information  Count
Uncensored value          24
Right censored value      21

Censoring value: censor = 0


Nonparametric Estimates


Characteristics of Variable

       Standard   95.0% Normal CI
Median     Error    Lower    Upper
    95  33.5410  29.2608  160.739


Additional Time from Time T until 50% of Running Units Fail

       Proportion
       of Running  Additional  Standard  95.0% Normal CI
Time T      Units        Time     Error   Lower    Upper
    20   0.977778          80  33.1662 14.9953  145.005


Actuarial Table

                                          Conditional
  Interval     Number  Number     Number  Probability   Standard
Lower  Upper Entering  Failed   Censored  of Failure      Error
    0     20       45       1          0    0.022222  0.0219739
   20     40       44       7          0    0.159091  0.0551405
   40     60       37       6          0    0.162162  0.0605974
   60     80       31       7          0    0.225806  0.0750952
   80    100       24       2          0    0.083333  0.0564169
  100    120       22       1          0    0.045455  0.0444095
  120    140       21       0          7    0.000000  0.0000000
  140    160       14       0          9    0.000000  0.0000000
  160    180        5       0          3    0.000000  0.0000000
  180    200        2       0          2    0.000000  0.0000000


       Survival   Standard   95.0% Normal CI
Time Probability      Error     Lower    Upper
  20    0.977778  0.0219739  0.934710  1.00000
  40    0.822222  0.0569937  0.710517  0.93393
  60    0.688889  0.0690122  0.553627  0.82415
  80    0.533333  0.0743698  0.387571  0.67910
 100    0.488889  0.0745172  0.342838  0.63494
 120    0.466667  0.0743698  0.320905  0.61243
 140    0.466667  0.0743698  0.320905  0.61243
 160    0.466667  0.0743698  0.320905  0.61243
 180    0.466667  0.0743698  0.320905  0.61243
 200    0.466667  0.0743698  0.320905  0.61243
```



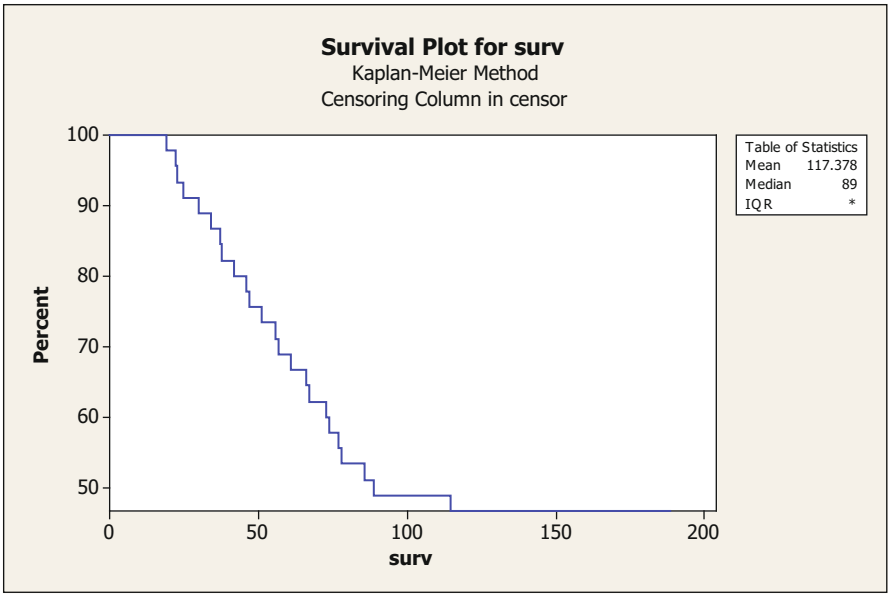


Fig. 19.1 Plots of estimated survival and cumulative functions under the Kaplan–Meier method

As can be seen from this method, the time is first categorized into classes with equal intervals. In this case, we employ 11 categories (0, 20) up to (200, .) with 20-month intervals. For instance, the 60-month survival rate is 0.6889 with a standard error of 0.0690. On the other hand, the estimated median residual lifetime, initially, was 95 months (s.e. = 33.5410) and has been dropping with it being 80 months at the beginning of the 20 months. The survival probability is the probability that the subject survives past the lower limit of that interval. We present the corresponding Survival and cumulative survival plots under this method in Fig. 19.2.

From the MINITAB output, the conditional probability of failure, which is an estimate of the probability that a patient will survive in the given interval, given that he/she made it to the start of the interval is computed as: (number failed/effective sample size). Thus, for instance, for the intervals $[0, 20), [20, 40), [40, 60)$, the computation is as follows: In the first interval, only one of the 45 patients died between 20 (inclusive) and 40 months, and therefore

$$1q_{20} = \frac{1}{45} = 0.0222;$$

$$\text{Hence, } \hat{S}(20) = 1 - 0.0222 = 0.9778$$

Seven of the remaining 44 patients died between 40 and less than 60 months, and therefore

$$1q_{40} = \frac{7}{44} = 0.1591.$$

The probability of living longer than 40 months is, therefore, estimated (using the multiplicative rule defined earlier) as

$$\hat{S}(40) = \hat{S}(20)[1 - 1q_{40}] = (0.9778)(0.8409) = 0.8222.$$

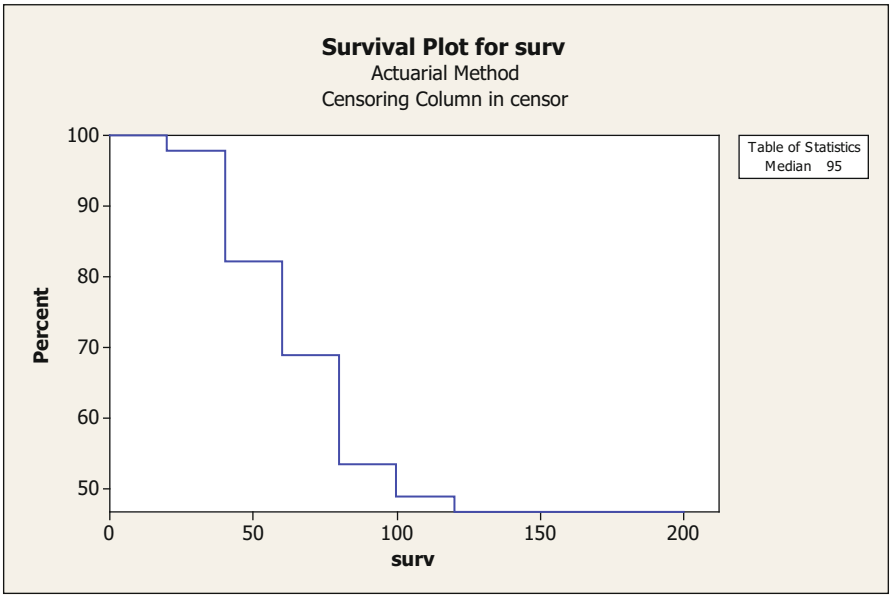
Similarly, 6 of the remaining 37 patients died between 60 (inclusive) and less than 80 months, and therefore

$$1q_{60} = \frac{6}{37} = 0.1622.$$

The probability of living longer than 60 months is, therefore, estimated as

$$\hat{S}(60) = \hat{S}(40)[1 - 1q_{60}] = (0.8222)(0.8378) = 0.6888.$$

The other surviving probabilities are computed similarly. These estimated probabilities are presented in the column labeled "survival probability" in the MINITAB output above.



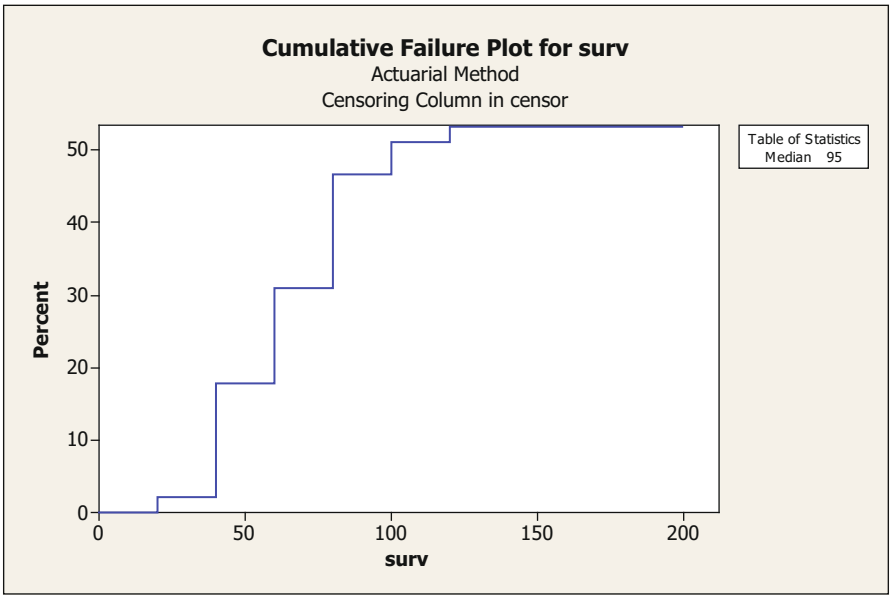


Fig. 19.2 Plot of estimated S and CS function under the life table method

19.4.4 Another Example

The data in Table 19.2 give the remission times in months for ten patients with tumors (adapted from Friendly).

Table 19.2 Remission times in months for ten tumor patients

Patient	Time	Censor
1	3.0	1
2	4.0	0
3	5.7	0
4	6.5	1
5	6.5	1
6	8.4	0
7	10.0	1
8	10.1	0
9	12.0	1
10	15.0	1

For the data above, six patients relapsed (censor = 1) after 3.0, 6.5, 6.5, 10.0, 12.0, and 15.0 months. One is lost to follow-up at 8.4 months, while three are still in remission at the end of study at 4.0, 5.7, and 10.1 months. The KM method applied to the data gives the following results:

```
MTB > print c1-c3

Data Display

Row  pat  time  censor
  1    1   3.0       1
  2    2   4.0       0
  3    3   5.7       0
  4    4   6.5       1
  5    5   6.5       1
  6    6   8.4       0
  7    7  10.0       1
  8    8  10.1       0
  9    9  12.0       1
 10   10  15.0       1

MTB > Ltest 'time';
SUBC>   Noparametric;
SUBC>   Splot;
SUBC>   Xminimum 0;
SUBC>   Brief 3;
SUBC>   KMEstimates;
SUBC>   Confidence 95.0;
SUBC>   TwoCI;
SUBC>   Censor 'censor';
SUBC>   Cvalue 0.

Distribution Analysis: time

Variable: time

Censoring Information  Count
Uncensored value           6
Right censored value       4

Censoring value: censor = 0
```

```
Nonparametric Estimates


Characteristics of Variable

            Standard   95.0%  Normal CI
Mean(MTTF)     Error    Lower    Upper
   10.0875   1.52692  7.09479  13.0802

Median = 10
IQR = 5.5  Q1 = 6.5  Q3 = 12


Kaplan-Meier Estimates

       Number  Number    Survival   Standard   95.0% Normal CI
Time  at Risk  Failed Probability      Error     Lower    Upper
 3.0       10       1    0.900000   0.094868  0.714061  1.00000
 6.5        7       2    0.642857   0.167949  0.313682  0.97203
10.0        4       1    0.482143   0.187719  0.114221  0.85006
12.0        2       1    0.241071   0.194595  0.000000  0.62247
15.0        1       1    0.000000   0.000000  0.000000  0.00000


Empirical Hazard Function

         Hazard
Time  Estimates
 3.0    0.10000
 6.5    0.16667
10.0    0.25000
12.0    0.50000
15.0    1.00000
```

The survival probabilities computations are similar to the earlier example, except that we have a tie here at 6.5 months. The survival probability plot is displayed in Fig. 19.3.

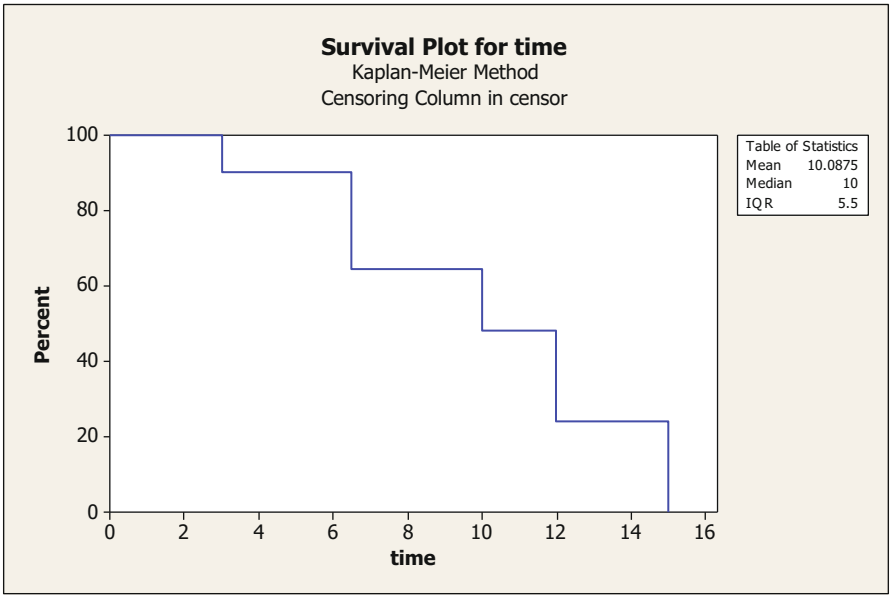


Fig. 19.3 Plot of estimated survival function based on the Kaplan–Meier method

Similarly, the actuarial method applied to the data gives the following results:

```
Nonparametric Survival Plot for time

MTB > Ltest 'time';
SUBC>   Noparametric;
SUBC>   Splot;
SUBC>   Xminimum 0;
SUBC>   Brief 3;
SUBC>   Intby 18 3;
SUBC>   Confidence 95.0;
SUBC>   TwoCI;
SUBC>   Censor 'censor';
SUBC>   Cvalue 0.

Distribution Analysis: time

Variable: time

Censoring Information   Count
Uncensored value            6
Right censored value        4

Censoring value: censor = 0

Nonparametric Estimates

Characteristics of Variable

          Standard    95.0% Normal CI
 Median      Error     Lower     Upper
10.9687   2.69782   5.68112   16.2564

Additional Time from Time T until 50% of Running Units Fail

        Proportion
        of Running  Additional   Standard    95.0% Normal CI
TimeT        Units        Time      Error     Lower     Upper
    3      1.00000     7.96875    2.84375   2.39510   13.5424
    6      0.88889     5.91667    2.97443   0.08689   11.7464
    9      0.61538     4.80000    2.24499   0.39989    9.2001
   12      0.43956     3.00000    2.12132   0.00000    7.1577

Actuarial Table

                                                 Conditional
  Interval       Number  Number    Number  Probability  Standard
Lower  Upper  Entering  Failed  Censored   of Failure     Error
    0      3        10       0         0      0.00000  0.000000
    3      6        10       1         2      0.11111  0.104757
    6      9         7       2         1      0.30769  0.181030
    9     12         4       1         1      0.28571  0.241473
   12     15         2       1         0      0.50000  0.353553
   15     18         1       1         0      1.00000  0.000000

          Survival  Standard 95.0% Normal  CI
Time  Probability     Error    Lower    Upper
   3      1.00000  0.000000  1.00000  1.00000
   6      0.88889  0.104757  0.68357  1.00000
   9      0.61538  0.176504  0.26944  0.96133
  12      0.43956  0.194875  0.05761  0.82151
  15      0.21978  0.183428  0.00000  0.57929
```

Time	Hazard Estimates	Standard Error	Density Estimates	Standard Error
1.5	0.000000	*	0.0000000	*
4.5	0.039216	0.039148	0.0370370	0.0349189
7.5	0.121212	0.084281	0.0911681	0.0547041
10.5	0.111111	0.109557	0.0586081	0.0523075
13.5	0.222222	0.209513	0.0732601	0.0611426
16.5	0.666667	0.000000	0.0732601	0.0611426

The survival probabilities as well as the survival probability $S(t)$ plot are displayed in Fig. 19.4.

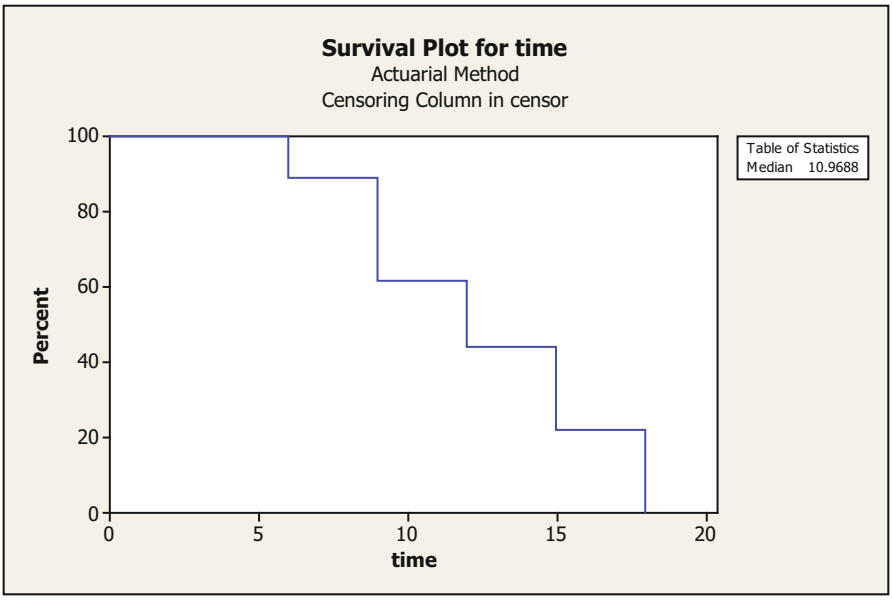


Fig. 19.4 Plot of estimated survival function based on the life table method

19.5 Testing Survival Times Between Two Groups

What we have done in the previous sections relating to the data in Table 19.1 is to characterize the survival times for a single group (we had ignored that there are two groups in the study, corresponding to the two treatments) of patients. However, we would like to compare the distributions of survival times for the two different groups or populations. We would want to know whether survival differs significantly between the two treatment groups. The following MINITAB statements and partial out are used to accomplish this test.

```
MTB > Ltest 'surv';
SUBC>   By 'trt';
SUBC>   Noparametric;
SUBC>   Splot;
SUBC>   Xminimum 0;
SUBC>   Brief 2;
SUBC>   KMEstimates;
SUBC>   Confidence 95.0;
SUBC>   TwoCI;
SUBC>   Censor 'censor';
SUBC>   Cvalue 0.

Distribution Analysis: surv by trt

Variable: surv
trt = 0

Censoring Information  Count
Uncensored value          16
Right censored value      20

Censoring value: censor = 0


Nonparametric Estimates


Characteristics of Variable

            Standard   95.0% Normal CI
Mean(MTTF)     Error    Lower    Upper
   128.167   11.8853  104.872  151.461

Median = *
IQR = *  Q1 = 56  Q3 = *


Kaplan-Meier Estimates

       Number  Number     Survival   Standard   95.0% Normal CI
Time  at Risk  Failed  Probability      Error     Lower    Upper
  19       36       1     0.972222  0.0273893  0.918540  1.00000
  25       35       1     0.944444  0.0381769  0.869619  1.00000
  30       34       1     0.916667  0.0460642  0.826382  1.00000
  34       33       1     0.888889  0.0523783  0.786229  0.99155
  37       32       1     0.861111  0.0576384  0.748142  0.97408
  46       31       1     0.833333  0.0621130  0.711594  0.95507
  47       30       1     0.805556  0.0659621  0.676272  0.93484
  51       29       1     0.777778  0.0692900  0.641972  0.91358
  56       28       1     0.750000  0.0721688  0.608552  0.89145
  57       27       1     0.722222  0.0746505  0.575910  0.86853
  61       26       1     0.694444  0.0767737  0.543971  0.84492
  66       25       1     0.666667  0.0785674  0.512677  0.82066
  67       24       1     0.638889  0.0800538  0.481986  0.79579
  74       23       1     0.611111  0.0812497  0.451865  0.77036
  78       22       1     0.583333  0.0821678  0.422287  0.74438
  86       21       1     0.555556  0.0828173  0.393237  0.71787

Distribution Analysis: surv by trt

Variable: surv
trt = 1

Censoring Information  Count
Uncensored value           8
Right censored value       1
```

```
Censoring value: censor = 0


Nonparametric Estimates


Characteristics of Variable

            Standard   95.0% Normal CI
Mean(MTTF)     Error    Lower    Upper
   69.2222   14.0612  41.6629  96.7816

Median = 73
IQR = 51  Q1 = 38  Q3 = 89


Kaplan-Meier Estimates

       Number  Number     Survival  Standard   95.0% Normal CI
Time  at Risk  Failed  Probability     Error     Lower    Upper
  22        9       1     0.888889  0.104757  0.683570  1.00000
  23        8       1     0.777778  0.138580  0.506166  1.00000
  38        7       1     0.666667  0.157135  0.358688  0.97465
  42        6       1     0.555556  0.165635  0.230918  0.88019
  73        5       1     0.444444  0.165635  0.119806  0.76908
  77        4       1     0.333333  0.157135  0.025355  0.64131
  89        3       1     0.222222  0.138580  0.000000  0.49383
 115        2       1     0.111111  0.104757  0.000000  0.31643


Distribution Analysis: surv by trt

Comparison of Survival Curves


Test Statistics

Method    Chi-Square  DF  P-Value
Log-Rank     5.49427   1    0.019
Wilcoxon     4.35118   1    0.037
```

The product limit method and the life table method give the same results for the test of the homogeneity of survival curves for the two groups. The two tests (log-rank and Wicoxon) indicate that the null hypothesis of homogeneity curves for the two groups is not tenable (p value $< .05$). Hence, we can say that the survival curves of the two groups significantly differ, with group 1 (trt = 1) more likely to survive longer than group 2 (trt = 0), with mean survival times being 70.94 and 66 days, respectively (note that these estimates are underestimated because the largest observations in both groups are censored). We present in Fig. 19.5 the survival curves for both groups under both methods of estimation.

The following MINITAB statements will also implement the life table method for the test of homogeneity of the two treatment groups. Results from this approach are identical to those from the KM method and we do not reproduce results from the application of this method here.

```
MTB > Ltest 'surv';
SUBC>   By 'trt';
SUBC>   Noparametric;
SUBC>   Splot;
SUBC>   Xminimum 0;
SUBC>   Brief 2;
SUBC>   Intby 200 20;
SUBC>   Confidence 95.0;
SUBC>   TwoCI;
SUBC>   Censor 'censor';
SUBC>   Cvalue 0.
```

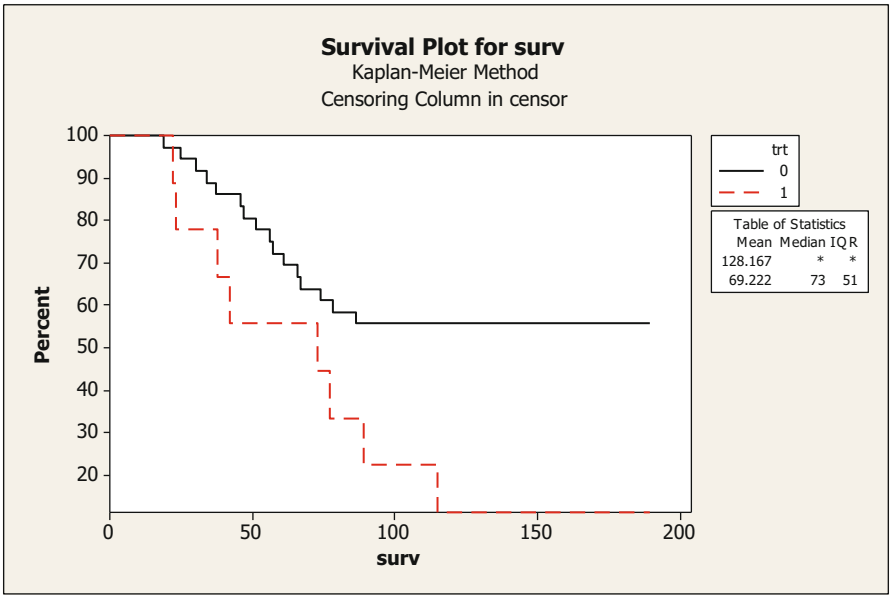


Fig. 19.5 Plot of estimated survival curves under both Kaplan–Meier methods

19.6 Hazard Function

If we let T be a random variable denoting the time of event occurrence, then, the hazard function, denoted by $h(t)$, is the probability that a subject experiences the event of interest in a small interval Δt, given that the subject has survived to the beginning of this interval. In other words, the hazard function is the conditional probability of experiencing the event between time t and $t+\Delta t$, given that the subject survived to at least time t. We can express this mathematically as

$$h(t) = \lim_{\Delta t \to 0} \frac{W(t, t + \Delta t)}{\Delta t}, \tag{19.2}$$

where

$$W(t, t + \Delta t) = \Pr(t < T < t + \Delta t \mid T \geq t).$$

The hazard function is sometimes referred to as the *conditional failure rate* in reliability, the instantaneous failure rate, or the age-specific failure rate in epidemiology, the force of mortality in demography, or simply as the hazard rate function.

Some Basic Properties of $h(t)$

- $h(t) \geq 0$ and can be greater than 1. Thus, $h(t)$ is not a probability.
- If $h(t)$ is constant over time t, then $E(T) = 1/h(t)$. For instance, if $h(t) = 0.32$ for all t, and time is measured in months, then $1/0.32 = 3.125$ days is the expected length of time until the event occurs.
- $h(t) =$ (No. of occurrence of events)/(Unit of time), where unit of time could be days, weeks, months, or years.

19.6.1 Types of Hazard Functions

The hazard function is related to the survival function with the following expression:

$$S(t) = \exp\left(-\int_0^t h(u)\, du\right). \tag{19.3}$$

With $S(t)$ defined above, we have,

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\ln[S(t)].$$

Thus,

$$f(t) = h(t)\exp\left(-\int_0^t h(u)\, du\right).$$

Similarly, the cumulative hazard function $H(t)$ is defined as

$$H(t) = \int_0^t h(u)\, du = -\ln[S(t)].$$

Thus,

$$S(t) = \exp[-H(t)].$$

The integral part in (19.3) is referred to as the *integrated hazard*. We now consider briefly, two of the most widely used hazard functions: the exponential and the Weibull models, although we will also employ both the lognormal and the loglogistic models in our applications.

(a) **The Exponential Model**
The Exponential or *constant hazard* model has the random variable T being distributed exponentially. Thus, if $h(t) = \lambda$ for all t, then, the probability distribution for T is given by

$$f(t) = \lambda \exp(-\lambda t). \tag{19.4}$$

(b) **The Weibull Model**
Here, the model becomes

$$\ln h(t) = \mu + \alpha \ln t, \quad \text{where } \alpha > -1, \tag{19.5}$$

and therefore

$$h(t) = \lambda_0 t^{\alpha}, \quad \text{with } \lambda_0 = e^{\mu}.$$

Suppose we now have explanatory variables $x_1, x_2, \ldots, x_p$, then, including the explanatory variables in each of the above models, we would have

$$\ln h(t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \quad \text{(Exponential)} \tag{19.6a}$$

$$\ln h(t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \alpha \ln t. \quad \text{(Weibull)} \tag{19.6b}$$

The expressions in (19.5) are often described as *accelerated failure time* (AFT) models and are members of the family of models known as the *proportional hazard* models. In general, an accelerated failure time model with covariates can be written in the form

$$\ln T = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}_{\text{covariate parameters}} + \sigma\varepsilon \tag{19.7a}$$

$$= \beta X + \sigma\varepsilon, \tag{19.7b}$$

where σ is the shape parameter and ε is the error distribution. Other AFT models are the Gompertz, the lognormal, the loglogistic, and the generalized Gamma models (these are not discussed in this book). The accelerated failure time models for the data in Table 19.1 has a model of the form

$$\ln T = \beta_0 + \beta_1 \text{trt}_i + \sigma\varepsilon. \tag{19.8}$$

The MINITAB program and the partial output for implementing the model in (19.8) for the case when ε has the exponential distribution is displayed below.

In the program, we specify in the model statement the dependent time variable, the censoring variable with the corresponding level, in this case, the zeros. On the right hand side of the model statement, we have specified the covariate (trt) with the relevant distribution. In this case, we are employing the exponential model. Similar results can be obtained for fitting the other AFT models by simply specifying the distribution in the model statement. For instance, the following will fit the Weibull models to the data. We present in Table 19.3 the log-likelihood for the four AFT models.

```
MTB > Ltest'surv';
SUBC>   By'trt';
SUBC>   Exponential;
SUBC>   Splot;
SUBC>   Brief 1;
SUBC>   LSXY;
SUBC>   Confidence 95.0;
SUBC>   TwoCI;
SUBC>   TESL;
SUBC>   Censor 'censor' ;
SUBC>   Cvalue 0.

Distribution Analysis: surv by trt

Variable: surv
trt = 0

Censoring Information   Count
Uncensored value           16
Right censored value       20

Censoring value: censor = 0

Estimation Method: Least Squares (failure time(X) on rank(Y))

Distribution:    Exponential


Parameter Estimates

                      Standard   95.0% Normal CI
Parameter   Estimate     Error    Lower    Upper
Mean         171.415   36.3923  113.066  259.874

Log-Likelihood = -104.491

Goodness-of-Fit
Anderson-Darling (adjusted) = 133.711


Distribution Analysis: surv by trt

Variable: surv
trt = 1

Censoring Information   Count
Uncensored value            8
Right censored value        1

Censoring value: censor = 0

Estimation Method: Least Squares (failure time(X) on rank(Y))

Distribution:    Exponential
```

```
Parameter Estimates

                     Standard   95.0% Normal CI
Parameter  Estimate     Error    Lower    Upper
Mean        75.3476   26.2036  38.1108  148.967

Log-Likelihood = -42.845

Goodness-of-Fit
Anderson-Darling (adjusted) = 7.010


Distribution Analysis: surv by trt

Test for Equal Scale Parameters

Chi-Square  DF      P
   4.06971   1  0.044
Scale parameter for surv by trt = 0 divided into:

trt   Lower  Estimate   Upper  --+---------+---------+---------+----
1    0.1978    0.4396  0.9769  (---------*---------------------)
                               --+---------+---------+---------+----
                               0.25      0.50      0.75      1.00


MTB > Lregression 'surv' = trt;
SUBC>   Factors 'trt';
SUBC>   Exponential;
SUBC>   CI;
SUBC>   Brief 2;
SUBC>   Confidence 95.0;
SUBC>   TwoCI;
SUBC>   Censor 'censor';
SUBC>   Cvalue 0.

Regression with Life Data: surv versus trt

Response Variable: surv

Censoring Information  Count
Uncensored value          24
Right censored value      21

Censoring value: censor = 0

Estimation Method: Maximum Likelihood

Distribution:   Exponential


Regression Table

                     Standard                 95.0% Normal CI
Predictor      Coef     Error      Z      P     Lower      Upper
Intercept   5.47096      0.25  21.88  0.000   4.98097    5.96095
trt
 1         -1.11585  0.433013  -2.58  0.010  -1.96454  -0.267163
Shape             1

Log-Likelihood = -146.376
```

We present graphically in Fig. 19.6, the probability plots of the applying the accelerated failure time models—exponential, weibull, lognormal, and the loglogistic—both on the full data regarding the covariate (trt) for the data in Table 19.1 and on the bottom graph, the plots by the two treatment groups.

```
Distribution ID Plot for surv

MTB > RDIdentification 'surv';
SUBC>   Weibull;
SUBC>   Exponential;
SUBC>   LNormal;
SUBC>   LLogistic;
SUBC>   Censor 'censor';
SUBC>   Cvalue 0;
SUBC>   LSXY;
SUBC>   Ptiles 1 5 10 50;
SUBC>   Allpts.

Distribution ID Plot:  surv

Goodness-of-Fit

             Anderson-Darling  Correlation
Distribution           (adj)  Coefficient
Weibull              140.453        0.966
Exponential          140.640            *
Lognormal            140.287        0.987
Loglogistic          140.328        0.975
```

From Table 19.3, we notice that the lognormal model provides the best fit in terms of smallest log-likelihood. For instance, since the Weibull is always the default for AFT models, a test of the hypotheses

$$H_0 : \sigma = 1$$
$$H_a : \sigma \neq 1,$$

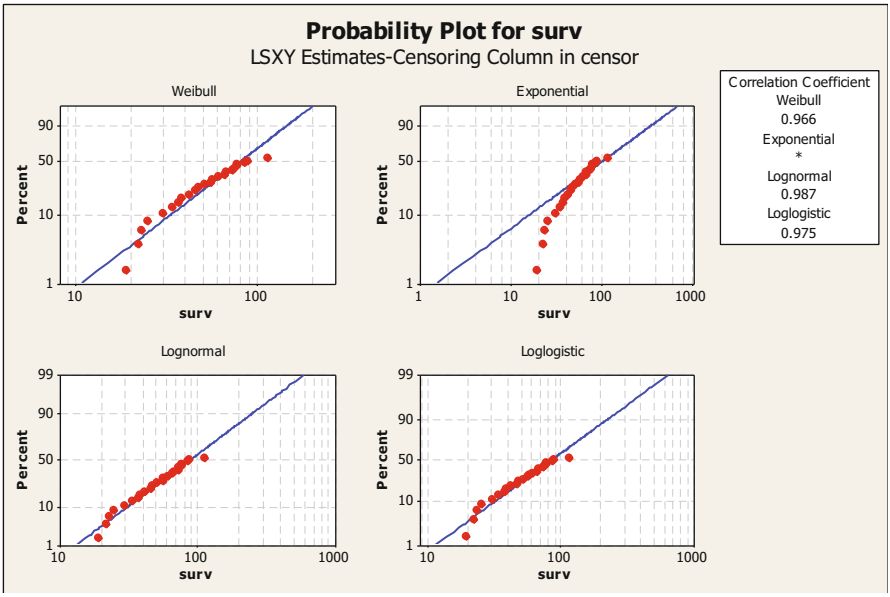
which tests whether the Weibull scale equals 1 or not for the exponential can be tested by calculating

$$-2(\text{log-likelihood}_{\text{Weibull}} - \text{log-likelihood}_{\text{expo}}) = 2(-145.785 + 146.376) = 1.182.$$

This statistic is distributed as χ^2 with 1 degree of freedom. The null hypothesis is rejected if $1.182 > \chi^2_{\alpha,1}$, the tabulated $\chi^2_{\alpha,1}$ value in Table 3 of the Appendix. Since $\chi^2_{0.05,1} = 3.841$, H_0 cannot be rejected. The best estimated AFT model for our data would, therefore, be

$$\widehat{\log T} = 4.95561 - 0.84084\,\text{trt}. \tag{19.9}$$

The percent change in a unit increase in the covariate is often computed as $100(e^{\hat{\beta}} - 1)$. Thus, in our example, the percentage change is: $100(e^{-0.84084} - 1) = -56.87\,\%$. In other words, immunoperoxidase-negative patients can expect their survival times to be decreased by about 56.87 %, or put in another way, immunoperoxidase-negative subjects are $e^{-.84084} = 0.43$ times more likely than immunoperoxidase-positive ones to survive longer. We



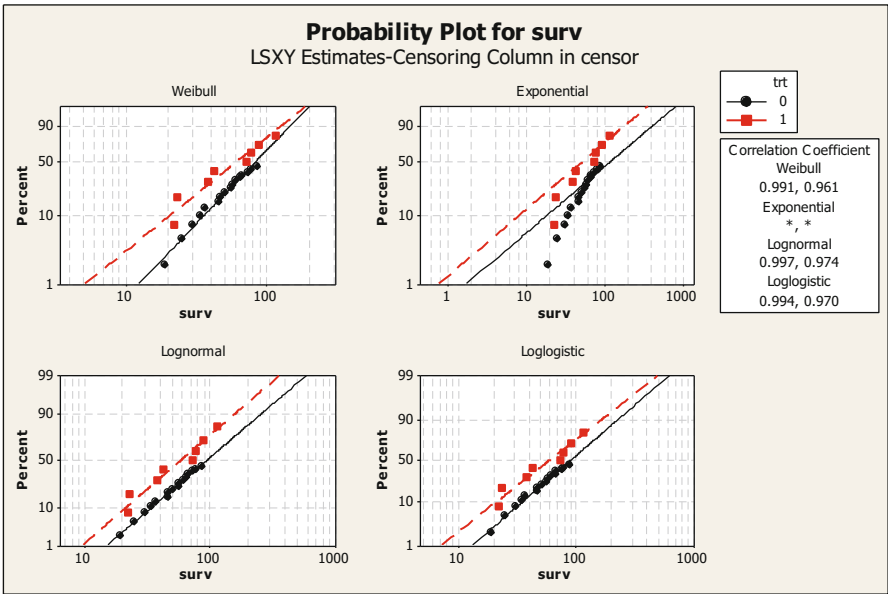


Fig. 19.6 Estimated probability plots for four accelerated failure time (AFT) models

Table 19.3 Results of fitting four accelerated failure time (AFT) models to our data

Model	Log-likelihood	Parameter estimates	
		$\hat{\alpha}$	$\hat{\beta}$
Exponential	−146.376	5.47096	−1.11585
Weibull	−145.785	5.34996	−0.980158
Lognormal	−143.886	4.95561	−0.84084
Loglogistic	−144.932	−0.851647	0.642838

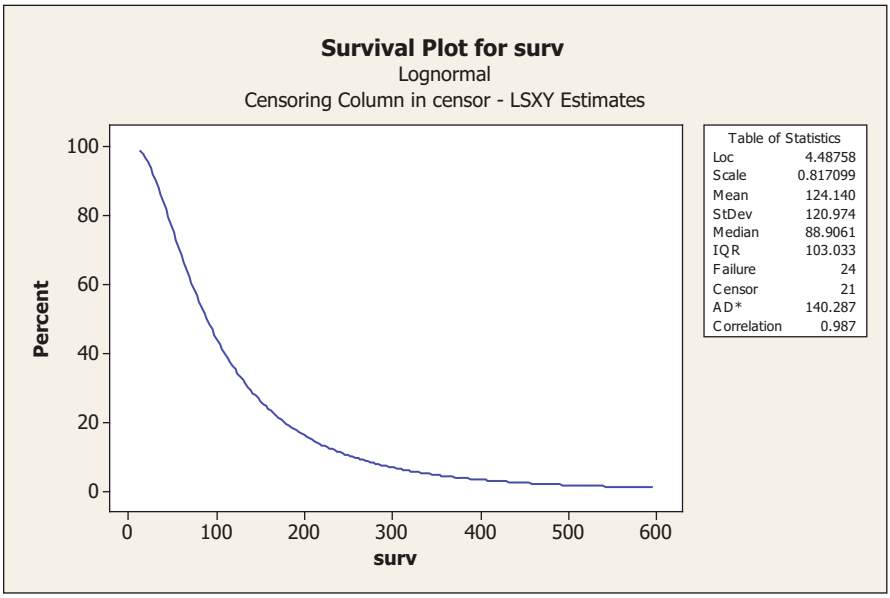


Fig. 19.7 Estimated survival plot under the lognormal accelerated failure time (AFT) model

present in Fig. 19.7 the plot of the estimated survival times for the data in Table 19.1.

We also present in Figs. 19.8, 19.9, and 19.10 respectively, the hazard, survival, and cumulative failure plots from the log-normal AFT models for the data in Table 19.1 for the two levels of the covariate trt.

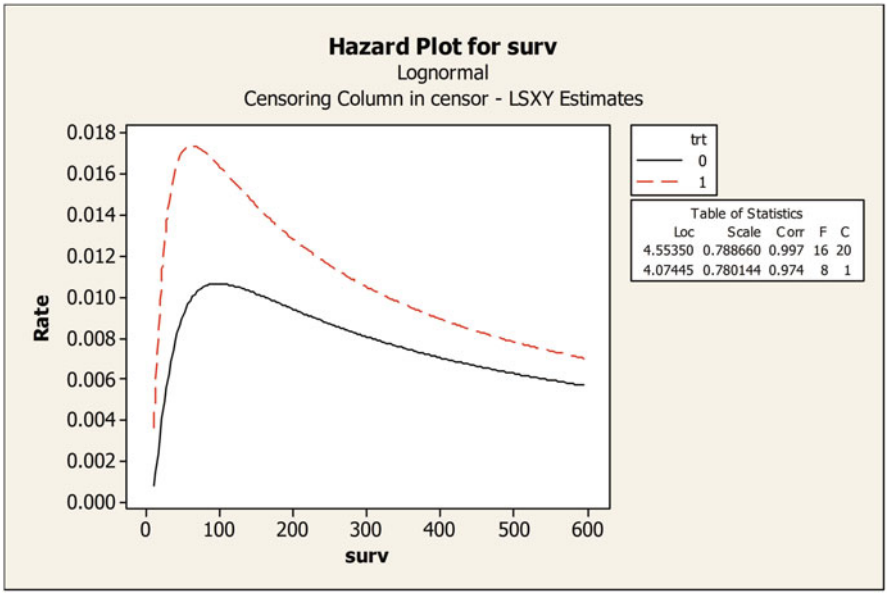


Fig. 19.8 Estimated hazard function from the lognormal Model

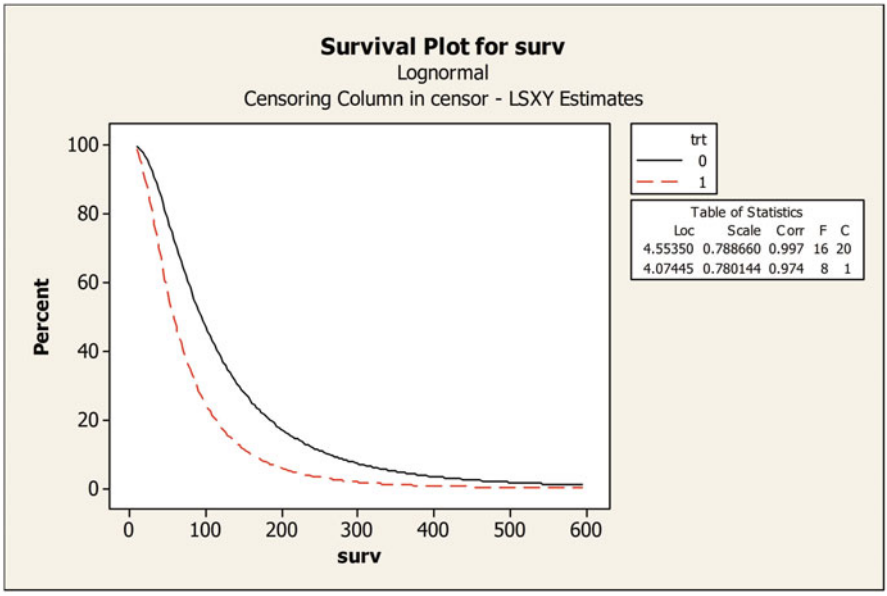


Fig. 19.9 Estimated survival function from the lognormal Model

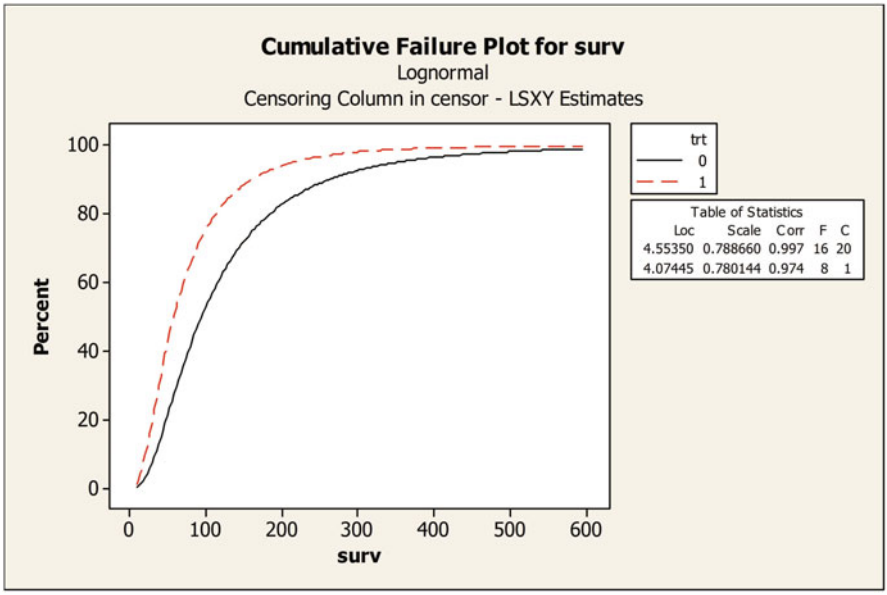


Fig. 19.10 Estimated cumulative Failure time Model

19.7 Proportional Hazards Model

Cox (1972) proposed a general method for modeling the hazard function $h(t)$ which unlike the AFT models discussed earlier, allows time-dependent covariates. Cox general method can be written as

$$\log h(t) = \log \alpha(t) + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \tag{19.10}$$

where $\alpha(t)$ is any function of t.

Clearly, $\alpha(t)$ can be assumed as the baseline hazard function when all the explanatory variables are zero. The expression in (19.10) is called the proportional hazard model because for any two hazard functions, the proportion $\dfrac{h_1(t)}{h_2(t)}$ does not depend on time and is therefore constant. MINITAB fits the Cox proportional hazard model for each of distributions. We now apply this model to the data in Table 19.1 using the Weibull distribution as our choice. The results are presented below.

```
MTB > Lregression 'surv' = trt;
SUBC>   Weibull;
SUBC>   EPplot;
SUBC>   Brief 2;
SUBC>   Confidence 95.0;
SUBC>   TwoCI;
SUBC>   Censor 'censor';
SUBC>   Cvalue 0.

Regression with Life Data: surv versus trt

Response Variable: surv

Censoring Information  Count
Uncensored value          24
Right censored value      21

Censoring value: censor = 0

Estimation Method: Maximum Likelihood

Distribution:   Weibull

Relationship with accelerating variable(s):   Linear


Regression Table

                      Standard                  95.0% Normal CI
Predictor       Coef     Error      Z      P     Lower      Upper
Intercept    5.34996  0.224016  23.88  0.000   4.91090    5.78902
trt        -0.980158  0.369932  -2.65  0.008  -1.70521  -0.255106
Shape        1.22545  0.219684                0.862391    1.74137

Log-Likelihood = -145.785


Anderson-Darling (adjusted) Goodness-of-Fit

Cox-Snell Residuals = 20.389
```

The parameter estimates for trt is -0.9802. We note here that MINITAB models $tt = 0$. Thus a patient with a negative treatment would die at about $e^{0.9802} = 2.67$ the rate of an individual with the positive treatment outcome.

19.8 Exercises

1. Suppose one is interested in examining the survival times of individuals with leukemia. The following data are the times, in months, to remission of 20 such patients.

1.50, 1.50, 1.50, 1.50, 1.75,
2.25, 2.50, 2.50, 2.75, 3.25,
4.00, 4.25, 4.75, 5.00, 5.50,
5.75, 6.25, 8.00, 8.00, 8.50

(a) What is the median survival time?
(b) For fixed intervals of length 2 months, use the life table method to estimate the survival function $s(t)$.
(c) Is the life table cross-sectional or longitudinal? Explain.
(d) Construct a survival curve for this sample of patients.

2. The following data represent survival times, in months, for 11 lymphoma patients. Values with asterisks (*) denote censored observations: 1*, 3, 4*, 5, 5, 6*, 7, 7, 7*, 8*, 8.

(a) What is the modal survival time?
(b) Use the Kaplan–Meier method to estimate the survival function $s(t)$.
(c) Construct a graph for the product-limit curve.
(d) Use the life table method to estimate the survival function $s(t)$.

3. Consider a clinical trial in which ten patients are observed to have the following survival pattern (in months). The plus (+) values are patients who are lost to follow-up. The values are as follows: 1, 2, 3, 3+, 4, 4+, 5, 5+, 8, 9+.

(a) What is the median survival time?
(b) Use the product-limit method to estimate the survival function $s(t)$.
(c) Construct a survival curve for this sample of patients.
(d) Use the life table method to estimate the survival function $s(t)$.

Group	Time
1	143, 164, 188, 188, 190, 192, 206, 209, 213, 216, 216*, 220, 227, 230, 234, 244*, 246, 265, 304
2	142, 156, 163, 198, 204*, 205, 232, 232, 233, 233, 233, 233, 240, 261, 280, 280, 296, 296, 323, 344*

4. Two groups of rats with different pretreatment regimes were exposed to a certain type of carcinogen. The time to mortality from cancer in the two groups was recorded and asterisk (*) denotes censored observation.

(a) For each group, use the product-limit method to estimate the survival function $s(t)$.
(b) For each group, construct a graph for the product-limit curve.
(c) Carry out a test to compare the distributions of survival times for the two groups.
(d) For each group, use the life table method to estimate the survival function $s(t)$.

5. The data in the following table are on two samples of 21 patients each, sample 1 was given an experimental drug and sample 2 was given a placebo.

The times to remission of leukemia patients are given in weeks and values with asterisks (*) denote censored observations.

Sample	Time
1	6*, 6, 6, 6, 7, 9*, 10*, 10, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
2	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

(a) What is the median survival time in each sample?
(b) Use the product-limit method to estimate the survival function $s(t)$ for the two sets of patients.
(c) Use the log-rank test to evaluate the null hypothesis that the distributions of survival times are identical in the two groups.
(d) For each set of patients, use the life table method to estimate the survival function $s(t)$.

6. The example in the following table relates to the survival times of 25 patients diagnosed with myelomatosis (Peto et al.). The patients were randomly assigned to two drug treatments. The variables of interest are:

- DUR is the time in days from the point of randomization to either death or censoring (which can be due to loss to follow-up or termination of the observation)
- STATUS has a value of 1 if dead and a value of 0 if censored.
- TRT takes the value of 1 or 2 to correspond to the two treatments.
- RENAL has a value of 1 if renal functioning was normal at the time of randomization; it has a value of 0 for impaired functioning.

Patients	DUR	STATUS	TRT	RENAL	Patients	DUR	STATUS	TRT	RENAL
1	8	1	1	1	14	1990	0	2	0
2	180	1	2	0	15	1976	0	1	0
3	632	1	2	0	16	18	1	2	1
4	852	0	1	0	17	700	1	2	9
5	52	1	1	0	18	1296	0	1	0
6	2240	0	2	0	19	1460	0	1	0
7	220	1	1	0	20	210	1	2	0
8	63	1	1	1	21	63	1	1	1
9	195	1	2	0	22	1328	0	1	0
10	76	1	2	0	23	1296	1	2	0
11	70	1	2	0	24	365	0	1	0
12	8	1	1	0	25	23	1	2	1
13	13	1	2	1					

7. The data in this exercise presented below give the effects of the drug ganciclovir on AIDS patients suffering from disseminated cytomegalovirus infection. Eighteen patients were treated with the drug and 11 were not (control group). The patients were followed and survival in months after diagnosis is labeled as time in the table, the censoring status (0 indicates censoring) and (1 denotes that death occurred). The treatments are designated (1 for for the drug) and (0 for no drug).

Time	Status	Trt
11	1	1
26	1	1
35	1	1
60	1	1
89	1	1
101	1	1
126	1	1
142	1	1
149	1	1
191	1	1
204	1	1
213	1	1
229	1	1
261	1	1
362	1	1
368	0	1
387	0	1
400	0	1
1	1	2
1	1	2
1	1	2
1	1	2
16	1	2
47	1	2
61	1	2
82	1	2
90	1	2
121	1	2
162	1	2

Chapter 20
Combined Analysis of Experimental Data

20.1 Introduction

Crop performance in field experiments depend on a number of factors, namely, experimental factors introduced by the researcher (e.g., fertilizer levels, spacing, plant density, pest control, weed control, etc.), environmental factors (such as soil fertility, seasonal variation, amount of sunshine, amount of rainfall, humidity, etc.), and the species properties such as the genotype of the crop. While experimental factors can be controlled by the researcher the environmental factors, which are often fixed, cannot be controlled by the researcher, nor can the consequent interaction effects of these environmental factors with experimental factors. In some experiments the effects of these environmental factors and their associated interactions may be even more pronounced and important than the effects of the research factors.

An experiment conducted at a particular site may therefore not lend itself to controllable environmental factors even though we can control the controllable factors nor to a changing seasonal effect on the performance of crops. Nor does an experiment conducted in one period in time lend itself to a generalization of results of such an experiment to other periods. Because these uncontrollable environmental factors are subject to changes with sites and seasons, and because these changes are measurable, the effects of these changes on crops therefore can be quantified or evaluated. This is why researchers repeat experiments at several sites and over several crop seasons. Gomez and Gomez (1984) identified four categories of this kind of experimentation. These are:

(a) Screening experiments which are meant to identify superior strains from a very large pool of strains—that is strains that consistently perform better than others in the pool.
(b) Experiments designed to evaluate the adaptation of the selected high performance strains on the sites where they were developed to further select the best-performing strains.

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8_20,

(c) Experiments designed to evaluate the range of geographical adaptability of the few selected strains earlier identified in our preliminary screening experiments.
(d) Experiments designed to evaluate the long-term effect of a group of strains and their sustainability.

20.1.1 General Analysis of Series of Experiments

For experiments conducted over several seasons or years, the analysis are often carried out in stages as follows:

1. The data thus obtained on the individual year or seasons or sites is analyzed.
2. Then a combined analysis is performed on the entire data.

However, the combined analysis is predicated on the assumption of homogeneity of variances at the various sites, years, or seasons. Bartlett's test of Homogeneity of variances is often employed to test this. However, if this assumption is violated then the validity of the treatments × sites interaction test may be suspect and a pooled error of the sites should be used instead of the overall pooled error. Further complications may arise if the treatments × sites interactions are not homogeneous, giving rise to some sites producing a better estimate of the difference than others. In such a situation, we would employ a transformation of the reciprocal of each site variance (that is, $1/s_i^2$ as a weighting function). We give an example of this in this chapter.

20.2 Analysis of Experiments Over Seasons

In most tropical parts of Africa for instance, maize is sometimes grown two or three times a year. The farmers know the planting dates and seasonal variation based on past experiences. Thus an experiment on a given crop conducted over several seasons will be analyzed first on individual season basis and then a combined analysis recognizing the season effect is *fixed*. The following example, from Gomez and Gomez (1984) is a fertilizer trial experiment over two seasons on rice designed as an RCB. The yields are displayed in Table 20.1. The levels of nitrogen have been modified from that presented by the original authors to simplify calculations and better understanding by our readers.

Table 20.1 Grain yield of rice with five nitrogen rates

Nitrogen rate	Replications		
	Rep I	Rep II	Rep III
Dry season			
0 (N0)	4.891	2.577	4.541
30 (N1)	6.009	6.625	5.672
60 (N2)	6.712	6.693	6.799
90 (N3)	6.458	6.675	6.636
120 (N4)	5.683	6.868	5.692
Wet season			
0 (N0)	4.999	3.503	5.356
30 (N1)	6.351	6.316	6.582
60 (N2)	6.071	5.969	5.893
90 (N3)	4.818	4.024	5.813
120 (N4)	3.436	4.047	3.740

20.2.1 Analysis

We first analyze the data for each season. We will therefore get two ANOVA tables which enable us to test separately whether there are significant differences among the means of the treatments for each of the seasons. These ANOVA tables are presented in Table 20.2. We also present the MINITAB statements for implementing these analyses. The partial output indicates significant differences in the mean responses for the five nitrogen levels. The significance is more pronounced in the wet season experiment as indicated by the computed p values from the ANOVA tables.

```
MTB > GLM 'Y' = REP N;
SUBC>   Brief 2 ;
SUBC>   Means N.

Least Squares Means for Y

N     Mean   SE Mean
  0  4.003   0.4341
 30  6.102   0.4341
 60  6.735   0.4341
 90  6.590   0.4341
120  6.081   0.4341

WET SEASON

MTB > GLM 'Y' = REP N;
SUBC>   Brief 2 ;
SUBC>   Means N.

Least Squares Means for Y

N     Mean   SE Mean
  0  4.619   0.3254
 30  6.416   0.3254
 60  5.978   0.3254
 90  4.885   0.3254
120  3.741   0.3254
```

Table 20.2 Separate ANOVA tables for the two seasons

Source	d.f.	SS	MS	F	p value
		Dry season			
REP	2	0.0186	0.0093	0.203	0.984
N	4	14.5334	3.6333	6.43	0.013
Error	8	4.5222	0.5653		
Total	14	19.0742			

Source	d.f	SS	MS	F	p-value
		Wet season			
REP	2	1.2429	0.6215	1.96	0.203
N	4	13.8699	3.4675	10.91	0.003
Error	8	2.5415	0.3177		
Total	14	17.6543			

20.2.2 Combined Seasonal Analysis

To conduct the combined analysis, we need to first test for the homogeneity of the seasons' variances (since we have two estimates of error variances now). That is, test the hypotheses:

$$\begin{aligned} H_0: &\quad \sigma_1^2 = \sigma_2^2 \\ H_a: &\quad \sigma_1^2 \neq \sigma_2^2 \end{aligned} \tag{20.1}$$

For more than two seasons or sites, we would employ Bartlett's test of homogeneity for this. This is explained in the next example. However, in this particular case, since there are only two seasons, we can accomplish the above test with the F test by computing:

$$F = \frac{\text{Larger MS}}{\text{Lower MS}} = \frac{0.5653}{0.3177} = 1.77$$

The p value for this test is 0.2177. Since p value > 0.05, we would therefore fail to reject H_0. That is, the two variances are homogeneous. We may therefore combine the data and run a combined analysis with the ANOVA table from MINITAB presented in Table 20.3. We present the MINITAB statement for implementing this analysis after the data have been combined (see the data display) which displays the first three and last three observations in the combined data set.

```
Row    N  REP  S    Y
  1    0    1  1  4.891
  2    0    2  1  2.577
  3    0    3  1  4.541
  ..................
 28  120    1  2  3.436
 29  120    2  2  4.047
 30  120    3  2  3.740

MTB > GLM 'Y' = SEASON REP(SEASON) N N* SEASON;
SUBC>   Brief 2 ;
SUBC>   Means N.

General Linear Model: Y versus SEASON, N, REP

Least Squares Means for Y

N     Mean  SE Mean
  0  4.311   0.2713
 30  6.259   0.2713
 60  6.356   0.2713
 90  5.737   0.2713
120  4.911   0.2713
```

Table 20.3 Combined analysis for the two seasons

Source	d.f.	SS	MS	F	p value
S	1	4.4954	4.4954	10.18	a
REP(S)	4	1.2616	0.3154	0.71	0.594
N	4	18.7488	4.6872	10.62	0.000
S*N	4	9.6544	2.4136	5.47	0.006
Error	16	7.0636	0.4415		
Total	29	41.2239			

The F tests for the various effects are now computed as follows:

$$\text{S} = \frac{\text{S MS}}{\text{REP(S) MS}}$$

$$\text{N} = \frac{\text{N MS}}{\text{Error MS}}$$

$$\text{S} \times \text{N} = \frac{\text{N} \times \text{S}}{\text{Error MS}}$$

MINITAB already computed these for us as: For factor N, 10.62 and $S \times N$ as 5.47, both of which are highly significant at $\alpha = 0.05$ level of significance. Notice that we did not compute the S F value because we do not have enough degrees of freedom for the replicate within seasons sum of squares (SS).

20.2.3 Partitioning the Interaction SS

Now that we have established that the SN interaction is significant, we would now partition the $S \times N$ SS into four orthogonal components using the principle of orthogonal polynomials since the levels of nitrogen are equally spaced from 0 to 120 in steps of 30. We employ the coefficients from the table of orthogonal polynomials in the appendix. These are reproduced below for our convenience.

	Factor levels				
	1	2	3	4	5
Linear	−2	−1	0	1	2
Quadratic	2	−1	−2	−1	2
Cubic	−1	2	0	−2	1
Quartic	1	−4	6	−4	1

To implement the partitioning in MINITAB, we first create coded columns based on the orthogonal polynomial coefficients using the "code" statement in MINITAB. The first coding relates to the linear effect of nitrogen denoted by "LR," the next three columns are for the quadratic, cubic, and quartic components of the nitrogen factor (N). Columns 9, 10, 11, and 12 respectively relate to the linear, quadratic, cubic, and quartic significant interaction terms of the $N \times S$ term. Note how the column of S are multiplied in turns with columns corresponding to the components of the N factor. We present the first and last six observations of all these in the MINITAB display below.

```
MTB > code (0) -2 (30) -1 (60) 0 (90) 1 (120) 2 c1 c5
MTB > code (0)  2 (30) -1 (60) -2 (90) -1 (120) 2 c1 c6
MTB > code (0) -1 (30) 2 (60) 0 (90) -2 (120) 1 c1 c7
MTB > code (0) 1 (30) -4 (60) 6 (90) -4 (120) 1 c1 c8
MTB > let c9=c3*c5
MTB > let c10=c3*c6
MTB > let c11=c3*c7
MTB > let c12=c3*c8

MTB > print c1-c12

Data Display

Row    N  REP  S      Y  LR  QR  CQ  QT  SL  SQ  SC  SQT
  1    0    1  1  4.891  -2   2  -1   1  -2   2  -1    1
  2    0    2  1  2.577  -2   2  -1   1  -2   2  -1    1
  3    0    3  1  4.541  -2   2  -1   1  -2   2  -1    1
  4   30    1  1  6.009  -1  -1   2  -4  -1  -1   2   -4
  5   30    2  1  6.625  -1  -1   2  -4  -1  -1   2   -4
  6   30    3  1  5.672  -1  -1   2  -4  -1  -1   2   -4
  ...................................................
 25   90    1  2  4.818   1  -1  -2  -4   2  -2  -4   -8
 26   90    2  2  4.024   1  -1  -2  -4   2  -2  -4   -8
 27   90    3  2  5.813   1  -1  -2  -4   2  -2  -4   -8
 28  120    1  2  3.436   2   2   1   1   4   4   2    2
 29  120    2  2  4.047   2   2   1   1   4   4   2    2
 30  120    3  2  3.740   2   2   1   1   4   4   2    2
```

Now that we have generated all the necessary components in MINITAB, we now fit the appropriate model using the GLM procedure in MINITAB. We observe here that the components while appearing in the GLM line, are further declared as covariates so that they are not considered as categorical variables.

```
MTB > Name c15 "COEF3"
MTB > GLM 'Y' = S REP( S) LR QR CQ QT SL SQ SC SQT;
SUBC>   Covariates 'LR' 'QR' 'CQ' 'QT' 'SL' 'SQ' 'SC' 'SQT';
SUBC>   Random 'S';
SUBC>   SSquares 1;
SUBC>   Brief 2 ;
SUBC>   Coefficients 'COEF3'.

General Linear Model: Y versus S, REP

Factor  Type    Levels  Values
S       random       2  1, 2
REP(S)  random       6  1, 2, 3, 1, 2, 3


Analysis of Variance for Y, using Sequential SS for Tests

Source  DF   Seq SS  Adj SS   Seq MS      F      P
S        1   4.4954  4.4954   4.4954  14.25  0.020
REP(S)   4   1.2616  1.2616   0.3154   0.71  0.594
LR       1   0.2757  9.4883   0.2757   0.62  0.441
QR       1  16.8188  1.2730  16.8188  38.10  0.000
CQ       1   1.6207  0.0000   1.6207   3.67  0.073
QT       1   0.0337  0.0016   0.0337   0.08  0.786
SL       1   9.4367  9.4367   9.4367  21.38  0.000
SQ       1   0.0316  0.0316   0.0316   0.07  0.793
SC       1   0.1755  0.1755   0.1755   0.40  0.537
SQT      1   0.0106  0.0106   0.0106   0.02  0.879
Error   16   7.0636  7.0636   0.4415
Total   29  41.2239


S = 0.664437   R-Sq = 82.87%   R-Sq(adj) = 68.94%


Term          Coef  SE Coef      T      P
Constant    5.5150   0.1213  45.46  0.000
LR          1.2575   0.2713   4.64  0.000
QR         -0.3893   0.2293  -1.70  0.109
CQ          0.0021   0.2713   0.01  0.994
QT          0.0061   0.1025   0.06  0.953
SLR        -0.7932   0.1716  -4.62  0.000
SQ         -0.0388   0.1450  -0.27  0.793
SC          0.1082   0.1716   0.63  0.537
SQT      -0.01006  0.06484  -0.16  0.879
```

Our results here indicate that only the quadratic (QR) effect of nitrogen is significant (p value = 0.000). Similarly, only the linear component of the SN interaction term here denoted as (SLR) is significant, indicating that

only the linear part of the response function is significant. However, since the quadratic component of the nitrogen factor is significant, we now fit the quadratic models to the data for each season. These estimated models are presented for each season below and in Fig. 20.1 are presented the plots of the response of rice to varying levels of nitrogen. The starred plot relates to the dry season response in this experiment.

$$\text{Dry Season: } \hat{Y} = 4.11719 + 0.0726N - 0.000476N^2$$

$$\text{Wet Season: } \hat{Y} = 4.85175 + 0.0513N - 0.000519N^2$$

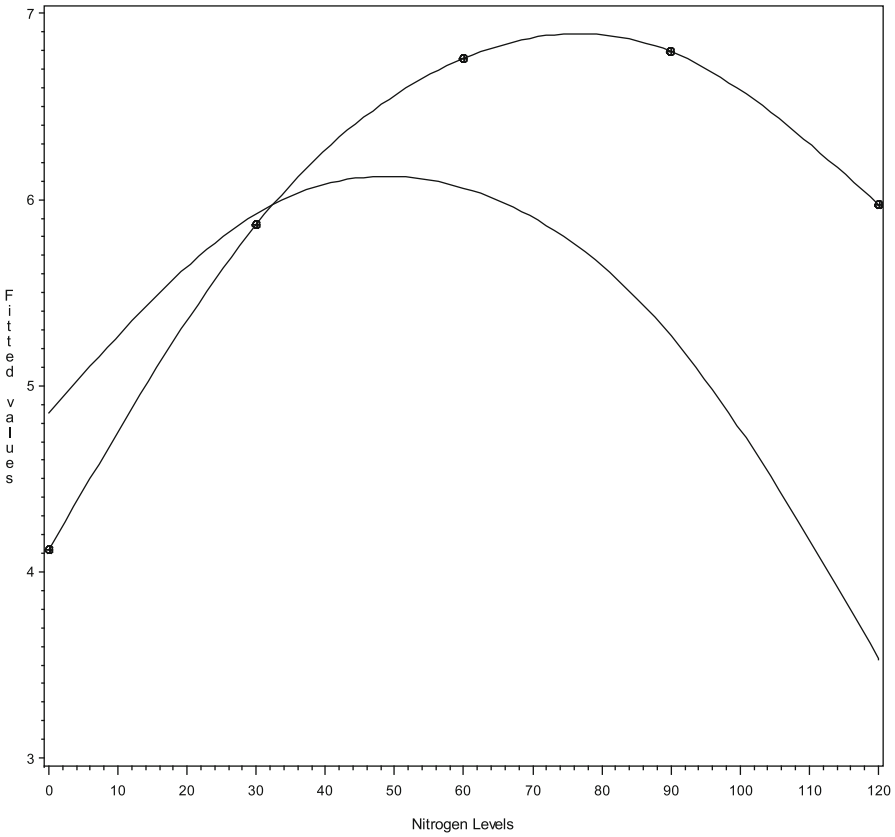


Fig. 20.1 Plots of effects of nitrogen

Based on the above, the optimum yield during the dry season will be achieved at:

$$\frac{0.0726}{2 \times 0.000476} = 76.26 \text{ Kg/ha}$$

Similarly for the wet season, this becomes:

$$\frac{0.0513}{2 \times 0.000519} = 49.42.26 \text{ Kg/ha}$$

Obviously, the cost of procuring nitrogen fertilizer for producing a kilogram of rice during the wet season is 64 % of the cost of producing the same during the dry season.

20.2.4 Effect of Failure of Homogeneity Assumption

If the error variances are not homogeneous, then the combined analysis we conducted above is suspect. An appropriate analysis in this case would be to run a weighted least squares analysis where the weights are the reciprocals of the root mean square errors. That is, $\omega_i = \frac{1}{s_i}$, $i = 1, 2$. In our case therefore, the weights would be $\omega_1 = \frac{1}{\sqrt{0.5653}} = 1.3300$ and $\omega_2 = \frac{1}{\sqrt{0.3137}} =$ 1.1.78544.

Hence, the response variable "yield" will be modified so that for the dry season, each yield value will first be multiplied by ω_1 and similarly, each wet season yield will be multiplied by ω_2 to create a new variable, say (yy) from the original response variable (y). That is,

$$yy = \begin{cases} y \times \omega_1 & \text{if wet season} \\ y \times \omega_2 & \text{if dry season} \end{cases}$$

Our combined analysis will now be performed on this new transformed variable (yy).

20.2.5 Example with Homogeneity Assumption Violated

The following data is reproduced by permission from Hashmand (1994) and relates to a randomized complete block experiment with four replications conducted to determine the effects of planting season using five nitrogen levels. The data in yields per acre are presented in Table 20.4.

Table 20.4 Yield of wheat from spring and winter planting

Nitrogen rate	Replications			
	Rep I	Rep II	Rep III	Rep IV
Spring planting				
0	27.8	24.6	28.2	26.9
50	30.0	29.2	30.1	28.9
100	29.9	28.3	29.7	30.0
150	31.4	32.0	31.7	31.8
200	30.8	31.3	29.9	32.0
250	30.5	31.2	33.0	31.8

Nitrogen rate	Replications			
	Rep I	Rep II	Rep III	Rep IV
Winter planting				
0	25.1	24.0	26.2	24.2
50	24.4	29.2	28.1	26.9
100	30.4	26.8	28.2	29.5
150	30.3	34.3	32.1	36.2
200	31.5	33.6	35.8	32.9
250	34.2	35.4	33.6	31.2

20.2.6 Combined Seasonal Analysis

Assuming the variances are equal, the combined analysis ANOVA table is displayed in Table 20.6 with the following accompanying MINITAB statements.

```
MTB > GLM 'Y' = S Rep( S) N S* N;
SUBC>   Brief 1 .
```

The values are computed with the pooled error mean square of 2.226 that is based on 30 degrees of freedom. Clearly there are significant differences in the means of nitrogen as well as in the interaction means between seasons and nitrogen, with both having p values that are very much less than $\alpha = 0.05$. However, we must be cautious here since the above F tests are predicated on the assumption that the seasons' variances are homogeneous. The test of homogeneity is conducted in this case with the usual F test (rather than with Bartlett's test of homogeneity) since there are only two variances. Hence the F test is computed using pooled estimates from the ANOVA table in Table 20.5 as:

Table 20.5 Separate ANOVA tables for the two seasons

Source	d.f.	SS	MS	F	p value
		Spring season			
REP	3	3.3650	1.1217	1.27	0.321
N	5	67.3983	13.4797	15.25	0.000
Error	15	13.2550	0.8837		
Total	23	84.0183			
		Winter season			
Source	d.f	SS	MS	F	p-value
REP	3	7.0502	2.501	0.70	0.566
N	5	275.043	55.009	15.42	0.000
Error	15	53.513	3.568		
Total	23	336.056			

Table 20.6 Combined analysis for the two seasons

Source	d.f.	SS	MS	F	p value
S	1	0.333	0.333	0.15	a
REP(S)	6	10.867	1.811	0.81	0.568
N	5	299.522	59.904	26.92	0.000
S*N	5	42.919	8.584	3.86	0.008
Error	30	66.768	2.226		
Total	47	420.410			

$$F = \frac{\text{Larger MS}}{\text{Lower MS}} = \frac{3.5680}{0.8837} = 4.04$$

The p value for this test is 0.0052. Since p value ≤ 0.05, we would therefore reject the null that the variances are homogeneous. Thus, we cannot use the pooled variance for the F tests for N and SN. We would need to partition the SN interaction term into a set of orthogonal contrast. We notice again here that the nitrogen levels are equally spaced, hence we can partition the nitrogen SS as well as the SN interaction SS into orthogonal components. The orthogonal coefficients from the appendix for $k = 6$ for the linear and quadratic components are (we would pool the other components together as "rest"):

	Factor levels					
	0	50	100	150	200	250
Linear	−5	−3	−1	1	3	5
Quadratic	5	−1	−4	−4	1	5

Recoding the levels of nitrogen in MINITAB as follows yield:

```
MTB > code (0) -5 (50) -3 (100) -1 (150) 1 (200) 3 (250) 5 c1 c5
MTB >  code (0) 5 (50) -1 (100) -4 (150) -4 (200) -1 (250) 5 c1 c6
MTB > let c7=c2*c5
MTB > let c8=c2*c6

MTB > GLM 'Y' = S Rep(S) Nl Nq SNl SNq;
SUBC>   Covariates 'Nl' 'Nq' 'SNl' 'SNq';
SUBC>   Brief 2 .
```

Table 20.7 Combined analysis for the two seasons

Source	d.f.	SS	MS	F	p value
S	1	0.333	0.333	0.15	a
REP(S)	6	10.867	1.811	0.81	0.568
N	(5)	(299.522)	59.904	26.92	0.000
N_l	1	264.688	264.688	118.91	0.000
N_q	1	18.335	18.355	8.25	0.007
N_{rest}	3	16.499	5.499	2.47	0.081
S*N	(5)	(42.919)	8.584	3.86	0.008
$N_l \times S$	1	36.109	36.109	31.87	0.001
$N_q \times S$	1	0.100	0.100	0.03	0.862
$N_{rest} \times S$	3	6.710	2.237	0.98	0.424
Error	(30)	(66.768)	2.226		
Reps within $N_l \times S$	6	6.797	1.133		
Reps within $N_q \times S$	6	18.384	3.064		
Reps within $N_{rest} \times S$	18	41.587	2.310		
Total	47	420.410			

We now have the comprehensive ANOVA table results for the analysis in Table 20.7

We have done the following in Table 20.7:

(i) The nitrogen SS was partitioned into three orthogonal components (N_l, N_q, N_{rest}) on 1, 1, and 3 d.f., respectively.
(ii) The pooled error SS that is based on 30 d.f. was partition into three components.
(iii) The F values for the orthogonal components of N were computed using as denominator the pooled error value of 2.226 on 30 d.f.
(iv) The F values for the interaction components were computed with the corresponding Reps within $N \times S$ SS. For instance,

$$\mathrm{N}_l \times S \text{ F- value} = \frac{\mathrm{N}_l \times S\ MS}{\text{Reps within } \mathrm{N}_l \times S\ MS} = \frac{36.109}{1.133} = 31.87$$

$$\mathrm{N}_q \times S \text{ F- value} = \frac{\mathrm{N}_q \times S\ MS}{\text{Reps within } \mathrm{N}_q \times S\ MS} = \frac{0.100}{3.064} = 0.03$$

$$\mathrm{N}_{rest} \times S \text{ F- value} = \frac{\mathrm{N}_{rest} \times S\ MS}{\text{Reps within } \mathrm{N}_{rest} \times S\ MS} = \frac{2.237}{2.310} = 0.98$$

Results in Table 20.7 indicate that only the linear component of the nitrogen $\times$ season is significant which indicates that responses to nitrogen rates is

linear across different seasons. Thus, different nitrogen rates need to be used for each of the seasons.

Alternatively, if we choose to use the transformational approach, then the observations for spring and winter seasons will be weighted by the reciprocal of their individual variances. That is by $\frac{1}{0.8837}$ and $\frac{1}{3.560}$, respectively, from Table 20.5. The initial analysis with this approach gives the following MINITAB output. The results here are similar to those we presented in Table 20.6.

```
MTB > GLM 'yy' = S Rep( S) N N* S;
SUBC>   Brief 2 .

Analysis of Variance for yy, using Adjusted SS for Tests

Source  DF    Seq SS   Adj SS   Adj MS         F      P
S        1  7820.67  7820.67  7820.67  11079.03  0.000
Rep(S)   6     4.90     4.90     0.82      1.16  0.355
N        5    94.65    94.65    18.93     26.92  0.000
S*N      5    13.26    13.26     2.65      3.76  0.009
Error   30    21.18    21.18     0.71
Total   47  7954.66
```

Again decomposing the nitrogen and the N*S interaction into orthogonal contrasts we have the results presented in Table 20.8.

Table 20.8 Combined weighted analysis for the two seasons

Source	d.f.	SS	MS	F	p value
S	1	7820.67	7820.67	8996.87	a
REP(S)	6	4.90	0.82	0.94	0.479
N	(5)	(94.65)	18.93	26.92	0.000
N_l	1	79.69	79.69	112.24	0.000
N_q	1	8.34	8.34	11.75	0.002
N_{rest}	3	6.62	2.21	3.47	0.028
S*N	(5)	(13.26)	2.65	3.74	0.009
$N_l \times S$	1	7.20	7.20	10.14	0.003
$N_q \times S$	1	2.56	2.56	3.61	0.067
$N_{rest} \times S$	3	3.50	1.167	1.64	0.201
Error	(30)	(21.18)	0.71		
Total	47	7954.66			

We observe here that the three nitrogen components are significant at the 5 % point as well as the N*S linear interaction. Since this is significant, our inference would therefore be based on this interaction term, and the conclusions to be drawn here are exactly the same as the ones presented earlier.

In the table below are the mean yields for each level of nitrogen by season.

	Nitrogen rates					
Season	0	50	100	150	200	250
Spring	26.88	29.55	29.48	31.75	31.00	31.63
Winter	25.10	27.15	28.73	33.23	33.45	33.60

Nitrogen level means for the two seasons

Since the Season*Nitrogen interaction is significant, we therefore model the nitrogen response by season using the Mitscherlich exponential fertilizer response model

$$y_i = a\left[1 - e^{-b(n_i+c)}\right]$$

where n_i are the levels of nitrogen and y is the mean yield. The model was previously discussed in Chap. 7. The model when implemented gives the following estimated regression Eqs. (20.2a) and (20.2b) for spring and winter yields, respectively:

$$\hat{y}_i = 31.8031\left[1 - e^{-0.012(n_i+159.50)}\right] \tag{20.2a}$$

$$\hat{y}_i = 38.6935\left[1 - e^{-0.005(n_i+221.60)}\right] \tag{20.2b}$$

The mean predicted regression equations are plotted in Fig. 20.2. Clearly, the initial mean yield for winter season is very low for low levels of nitrogen rates, but as the nitrogen rates increase, there is clearly a significant rise in the mean yield for winter over spring. Thus if interest centers on lower rates for nitrogen levels, then spring application will provide higher yield; however, if higher yields are desired, then winter application of more than 125 kg/ha would begin to give higher yields in winter than for spring season. Of course, the cost of procuring fertilizer would have to be taken into consideration in this case.

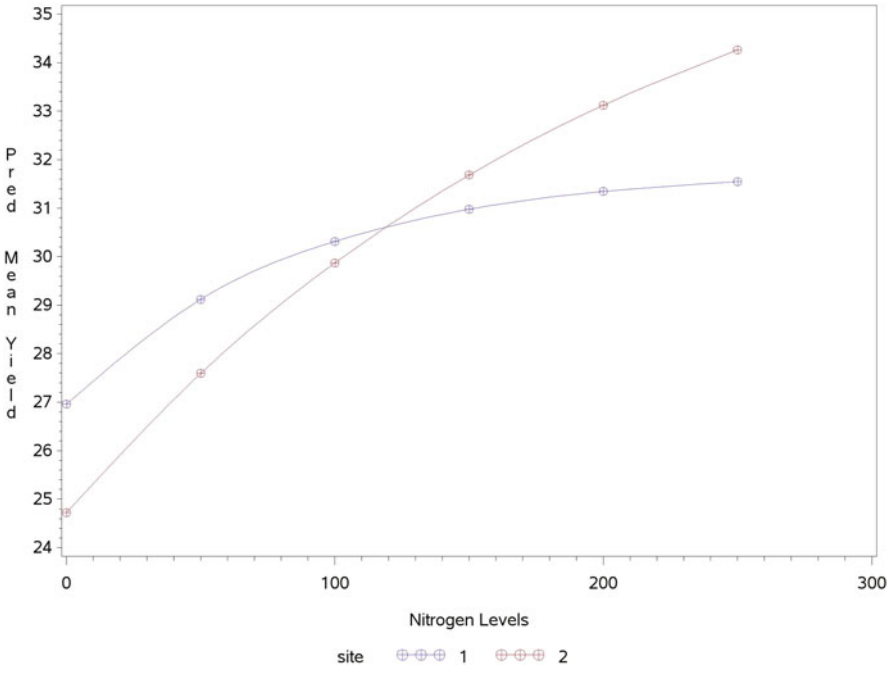


Fig. 20.2 Estimated plots for both seasons

20.3 Experiment at Several Sites

The objective of this type of experiment is to ascertain the adaptability of already selected variety strains of a crop or fertilizer levels. The experiment is often conducted in randomized complete block design or in split-plot design. For factorial experiments, the number of levels of the factors are often very small, usually two level per factor. We present examples of both the randomized block design and split-plot design in the new examples.

20.3.1 RCB Design Example

A screening varietal trial on maize was conducted at three different locations in Nigeria (Samaru, Ibadan, and Umudike). The study concerns sixteen new varieties and four control strains (s-230, s238, ib10, um234). The experiments were conducted in randomized complete block of three replications each. The data is presented in Table 20.9. The concern here is to see the adaptability of the 16 strains over the three sites.

Again our initial analysis is based on analyzing separately the data for each of the sites as a randomized block design having 20 treatments each. We may add here that the treatments are randomized within blocks at each site. We present a MINITAB statement for implementing one of these and in Table 20.10 are presented in the analysis of variance table results for each of the sites.

Table 20.9 Synthetic Data for three Sites

Site	SAMARU			IBADAN			UMUDIKE		
TRT	REP1	REP2	REP3	REP1	REP2	REP3	REP1	REP2	REP3
1	8.14	7.78	7.10	9.22	8.24	8.30	9.28	9.01	9.74
2	7.55	7.55	7.45	8.51	8.79	8.72	8.60	9.70	9.68
3	8.04	7.78	8.07	8.73	7.59	8.34	8.77	9.45	9.59
4	8.21	7.72	8.41	8.33	9.01	8.04	9.62	10.18	9.66
5	8.17	8.66	7.52	8.47	8.62	8.41	9.46	10.60	10.02
6	8.12	7.39	7.24	8.13	8.79	8.76	10.11	9.54	10.16
7	7.82	8.23	8.12	8.02	8.43	8.31	9.85	9.97	9.00
8	8.39	7.30	7.80	8.32	8.63	8.73	10.55	10.11	8.71
9	7.72	7.46	7.36	8.50	8.74	8.64	9.04	9.83	9.30
10	8.57	8.38	7.28	8.39	8.58	8.37	9.79	10.29	9.71
11	8.02	9.09	8.40	8.41	9.08	8.68	10.33	8.76	9.62
12	7.57	8.49	7.67	8.43	8.11	7.93	8.90	8.95	9.46
13	8.16	7.94	7.21	8.17	8.74	8.46	9.48	10.67	9.58
14	7.36	7.39	8.21	9.32	8.19	8.84	9.55	9.92	9.37
15	7.47	7.45	7.78	8.74	8.38	8.22	9.46	9.77	8.83
16	8.07	8.34	7.29	9.08	8.40	8.29	10.19	8.91	10.38
17	3.71	3.71	3.75	4.87	5.02	4.74	4.91	4.97	4.96
18	4.13	4.17	4.19	4.74	4.78	4.32	5.11	4.77	5.10
19	3.91	4.42	3.38	4.60	4.76	4.82	4.81	5.20	5.13
20	4.23	3.93	4.19	4.49	4.55	4.93	5.09	5.23	4.93

Table 20.10 Separate ANOVA Tables for the Three Sites

Samaru

Source	D.F.	SS	MS	F	p-value
REP	2	0.7849	0.3924	2.40	0.105
TRT	19	148.4751	7.8145	47.72	0.000
Error	38	6.2224	0.1637		
Total	59	155.4824			

Ibadan

Source	d.f	SS	MS	F	p-value
REP	2	0.0854	0.0427	0.38	0.685
TRT	19	139.0019	7.3159	65.52	0.000
Error	38	4.2434	0.1117		
Total	59	143.3307			

Umudike

Source	d.f	SS	MS	F	p-value
REP	2	0.2833	0.1416	0.56	0.575
TRT	19	206.5256	10.8698	43.12	0.000
Error	38	9.5788	0.2521		
Total	59	216.3877			

The three individual ANOVA tables above indicate that there are highly significant differences among the means of the twenty varieties at a 5 % significance level. However, we would like to combine data from these various sites into a combined analysis to inform on the adaptability of these varieties at different site environments. We discuss doing this in the next section.

20.4 Combined Analysis

In order to combine the site analysis, we need to ascertain that the variances in each sites are homogeneous. That is, we need to test the hypotheses:

$$\begin{aligned} H_0 &: \quad \sigma_1^2 = \sigma_2^2 = \sigma_3^2 \\ H_a &: \quad \text{at least two of these are unequal} \end{aligned} \tag{20.3}$$

For our individual site analysis, we have the following information from the analysis of variance tables.

	Error d.f.	
Site	f_i	EMS
1	38	0.1637
2	38	0.1117
3	38	0.2521

To accomplish the above hypotheses, we will employ Bartlett's homogeneity test, which is based on the following computations:

(1) Compute the pooled variance for the three sites as

$$S_P^2 = \frac{\sum_i f_i S_i^2}{\sum_i f_i}, \quad i = 1, 2, \cdots, k.$$

(2) Compute

$$q = \left(\sum_{i=1}^{2} f_i\right) \log_{10} S_P^2 - \sum_{i=1}^{k} f_i \log_{10} S_i^2.$$

(3) Compute

$$c = 1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^{k} f_i^{-1} - \left(\sum_{i=1}^{k} f_i\right)^{-1}\right].$$

(4) Then

$$\chi_0^2 = 2.3026 \frac{q}{c} \sim \chi_{k-1}^2.$$

For our data, based on the summary statistics in the table above,

$$S_P^2 = \frac{38(0.1637 + 0.1117 + 0.2521)}{114} = 0.1758;\ q = 2.7352 \text{ and } c = 1.0117.$$

Consequently, $\chi_0^2 = \dfrac{2.7352}{1.0117} = 2.276$. When we compare this with a $\chi_{(.05,2)}^2 = 5.99$, clearly we will fail to reject H_0. In other words, the sites' variances are homogeneous. We are now in a position to combine the entire data set for the three sites and the analysis which is now a random effects model (because the sites are just a sample of several possible sites for the experiment). The ANOVA table and the variance components are presented in the MINITAB output displayed below. First we combine the data and code in the appropriate values for S, REP, and TRT. We present the data display for the first five treatments and the last five treatments. There are here now 180 observations. Since the sites are assumed *fixed*, the GLM analysis of variance for this experiment has the ANOVA table results presented in Table 20.11.

```
MTB > print c1-c4

Data Display

Row  S  REP  TRT      Y
  1  1    1    1   8.14
  2  1    1    2   7.55
  3  1    1    3   8.04
  4  1    1    4   8.21
  5  1    1    5   8.17
  6  1    1    6   8.12
  7  1    1    7   7.82
  8  1    1    8   8.39
  9  1    1    9   7.72
 10  1    1   10   8.57
 ....................
170  3    3   10   9.71
171  3    3   11   9.62
172  3    3   12   9.46
173  3    3   13   9.58
174  3    3   14   9.37
175  3    3   15   8.83
```

```
176  3    3   16  10.38
177  3    3   17   4.96
178  3    3   18   5.10
179  3    3   19   5.13
180  3    3   20   4.93

MTB > GLM 'Y' = S REP( S) TRT S* TRT;
SUBC>   Brief 2 ;
SUBC>   EMS;
SUBC>   Means S TRT S* TRT.

General Linear Model: Y versus S, TRT, REP
```

Table 20.11 Combined ANOVA tables for the three sites

Source	d.f.	SS	MS	F	p value
S	2	78.7382	39.3691	223.91	0.000
REP(S)	6	1.1535	0.1923	1.09	0.371
TRT	19	484.8331	25.5175	145.13	0.000
S*TRT	38	9.1695	0.2413	1.37	0.103
Error	114	20.0445	0.1758		
Total	179	593.9389			

Clearly, the mean yields in the sites are clearly significant as well as the variety means. However, there does not seem to be much significant interaction effect of site and treatment means. The S.E. for each of the means are computed as follows:

(i) The S.E. of site mean is:

$$\sqrt{\frac{EMS}{60}} = \sqrt{\frac{0.1758}{60}} = 0.0541.$$

Thus, the S.E. for comparing any two sites means is:

$$\sqrt{\frac{2\,EMS}{60}} = \sqrt{\frac{2(0.1758)}{60}} = 0.0766.$$

(ii) Similarly, the S.E.s for a variety mean and corresponding S.E. for comparing two variety means are, respectively:

$$\sqrt{\frac{EMS}{9}} = \sqrt{\frac{(0.1758)}{9}} = 0.1398 \quad \text{and} \quad \sqrt{\frac{2\,EMS}{9}} = \sqrt{\frac{2(0.1758)}{9}} = 0.1977.$$

(iii) The S.E. for an interaction mean and comparing any two interaction means are computed, respectively, as:

$$\sqrt{\frac{EMS}{3}} = \sqrt{\frac{(0.1758)}{3}} = 0.2421 \quad \text{and} \quad \sqrt{\frac{2\,EMS}{3}} = \sqrt{\frac{2(0.1758)}{3}} = 0.3423.$$

Variety	Sites			Mean
	Samaru	Ibadan	Umudike	
1	7.673	8.587	9.343	8.534
2	7.517	8.673	9.327	8.506
3	7.963	8.220	9.270	8.484
4	8.113	8.460	9.820	8.798
5	8.117	8.500	10.027	8.881
6	7.583	8.560	9.937	8.693
7	8.057	8.253	9.607	8.639
8	7.830	8.560	9.790	8.727
9	7.513	8.627	9.390	8.510
10	8.077	8.447	9.930	8.818
11	8.503	8.723	9.570	8.932
12	7.910	8.157	9.103	8.390
13	7.770	8.457	9.910	8.712
14	7.653	8.783	9.613	8.683
15	7.567	8.447	9.353	8.456
16	7.900	8.590	9.827	8.772
17	3.723	4.877	4.947	4.516
18	4.163	4.613	4.993	4.590
19	3.903	4.727	5.047	4.559
20	4.117	4.657	5.083	4.619
Mean	7.083	7.746	8.694	7.841

Table of $T \times S$ means

Since there are significant differences among the variety means, we can partition the treatment SS into three components:

1. Between the 16 new varieties with SS computed from the means as: on 15 d.f.

$$\begin{aligned} BTSS = 9\big[&8.534^2 + 8.506^2 + 8.484^2 + 8.798^2 + 8.881^2 + 8.693^2 + 8.639^2 \\ &+ 8.727^2 + 8.510^2 + 8.818^2 + 8.932^2 + 8.390^2 + 8.712^2 + 8.683^2 \\ &+ 8.456^2 + 8.772^2\big] - \frac{9\big[138.535^2\big]}{16} = 3.566. \end{aligned}$$

2. Between control varieties SS: On 3 d.f.

$$BCSS = 9\big[4.516^2 + 4.590^2 + 4.559^2 + 4.619^2\big] - \frac{9\big[18.284^2\big]}{4} = 0.0525.$$

3. Control vs. others on 1 d.f.

$$\begin{aligned} \text{C vs. O}SS &= \frac{\big(138.535 \times 9\big)^2}{144} + \frac{\big(18.284 \times 9\big)^2}{36} \\ &= \frac{1246.77^2}{144} + \frac{164.556^2}{36} - \frac{1411.326^2}{180} \\ &\quad 481.0922 \end{aligned}$$

The variety (or TRT SS of 484.8331 has been partitioned into three orthogonal components. These results are presented in Table 20.12. Notice that the partitioning SS add up (give or take a few computational errors) to the TRT SS as do the degrees of freedom. Our analysis therefore indicate that while there are no significant differences between the means of the 16 new varieties, nor between the means of the four control varieties, there is however significant difference between the means of the controls and the new variety. Clearly, the new variety strains are much more superior to the control strains.

Table 20.12 New ANOVA table for partitioned SS

Source	d.f.	SS	MS	F
TRT	19	484.8331	25.5175	145.13
Others	15	3.566	0.2377	1.35
Between controls	3	0.0525	0.0175	0.10
Contol vs. others	1	481.0922	481.0922	2736.59
Error	114	20.0445	0.1758	
Total	179	593.9389		

20.5 Split-Plot Example

Suppose we have a series of experiments over s sites (S) designed as a factorial structure with a levels for the main plot A, and b levels for the sub-plot B, and with r replications (R); then the structure of the ANOVA table is presented in Table 20.13.

Table 20.13 Df under the split-plot model at s sites

Source	Df
S	$s-1$
Reps within S	$s(r-1)$
Main plot A	$(a-1)$
S*A	$(s-1)(a-1)$
Pooled error (a)	$s(r-1)(a-1)$
Subplot factor B	$(b-1)$
A*B	$(a-1)(b-1)$
S*A*B	$(s-1)(a-1)(b-1)$
Error	$sa(r-1)(b-1)$
Total	$srab-1$

The site effects of S will be tested with the Reps within Sites Mean square. The main plot A and interaction $S*A$ will be tested by the pooled error (a) Mean square, while the B, $A*B$ and $S*A*B$ effects are tested with the error mean square.

Example

A study conducted at Samaru and Ibadan was to determine the influence of row spacing and plant density on corn (*Zea mays* L.) yield. The experiment was a 2×3 factorial replicated four times in a randomized complete block design arranged and conducted at two different sites. The factorial was laid out as a split-plot design. In this experiment, factor A is the two row spacings (12 and 25 in) assigned to the main, and factor B is the three target plant densities (12,000, 16,000, and 20,000 plants per acre) assigned to the subplots. The data from the experiments are displayed in Table 20.14.

We present below the analysis of the above data in MINITAB. The data are read into columns C1 to C5. We present the first five and last five observations for these data below. The GLM statement for implementing the model is also presented. The analysis of variance table is presented in Table 20.15.

```
Row  S  A  B  R    Y
  1  1  1  1  1  140
  2  1  1  1  2  138
  3  1  1  1  3  130
  4  1  1  1  4  142
  5  1  1  2  1  145
 ..................
 44  2  2  2  4  132
 45  2  2  3  1  140
```

Table 20.14 Yield of corn at two different sites with factorial design

			Grain yields (bushels/acre)			
	Row spacing	Plant density	Replications			
Sites	(in)	(plants/acre)	I	II	III	IV
Samaru	12	12,000	140	138	130	142
		16,000	145	146	150	147
		20,000	150	149	146	150
			435	433	426	439
	25	12,000	136	132	134	138
		16,000	140	134	136	140
		20,000	145	138	138	142
			421	404	408	420
Badeji	12	12,000	142	132	128	140
		16,000	146	136	140	141
		20,000	148	140	142	140
			436	408	410	421
	25	12,000	132	130	136	134
		16,000	138	132	130	132
		20,000	140	134	130	136
			410	396	396	402

```
 46  2  2  3  2  134
 47  2  2  3  3  130
 48  2  2  3  4  136
```

```
MTB > GLM 'Y' = S R( S) A S* A  A*R(S) B S* B A* B S* A* B;
SUBC>   Random 'R';
SUBC>   Brief 2 .

General Linear Model: Y versus S, A, B, R

Factor  Type    Levels  Values
S       fixed        2  1, 2
R(S)    random       8  1, 2, 3, 4, 1, 2, 3, 4
A       fixed        2  1, 2
B       fixed        3  1, 2, 3
```

The results from the ANOVA table indicate that the main effect A, the subplot factor B and the interaction effects $A * B$ are significant. Since the $A * B$ interaction is significant, we decide to explore this interaction further by partitioning it into both linear and quadratic components each on 1 d.f. This is accomplished in MINITAB by recoding the levels of A and B using orthogonal polynomial coefficients. The codes for implementing this are displayed below and the results are embedded in Table 20.15. Only the linear component is significant, which indicates that for a given row spacing, yield increases linearly with increases in plant density.

```
MTB > code (1) 1 (2) -1 c2 c6
MTB > code (1) 1 (2) 0 (3) -1 c3 c7
MTB >  code (1) 1 (2) -2 (3) 1 c3 c8
MTB > let c9=c6*c7
MTB > let c10=c6*c8
```

I hope readers have by now realized that we have seen the data in Table 20.14 before, in Table 15.15 of Chap. 15. We recognize that the ANOVA table displayed here in Table 20.15 is a replica of that displayed for the split–split plot ANOVA table (see page 773).

Table 20.15 Combined analysis based on the split-plot design

Source	d.f.	SS	MS	F	p value
S	1	238.521	238.521	5.27	0.061
R(S)	6	271.625	45.271	6.83	0.017
A	1	475.021	475.021	71.63	0.000
S*A	1	1.687	1.687	0.25	0.632
A*R(S)	6	39.792	6.632	0.75	0.612
B	2	350.042	175.021	19.92	0.000
Linear	1	338.00	338.00	38.47	0.000
Quadratic	1	12.04	12.04	1.37	0.253
S*B	2	37.042	18.521	2.11	0.143
A*B	2	87.792	43.896	5.00	0.015
$L \times L$	1	55.12	55.12	6.27	0.019
$L \times Q$	1	32.67	32.67	3.72	0.066
S*A*B	2	1.625	0.812	0.09	0.912
Error	24	210.833	8.785		
Total	47	1713.979			

20.6 Experiments Conducted Over Several Years

Because of the unpredictability of environmental factors over the years, the factor years will be assumed to be a random effect. Suppose we have an experiment conducted as a randomized block design with say t treatments (T), replicated (R) r times, and conducted over p years (Y). Thus we have a total of $n = tpr$ plots in the experiment. As in the previous case, we first analyze the data on a yearly basis obtaining the ANOVA table for each year. To combine the data, we need to again do the test of variance homogeneity for the p years. If the hypothesis of homogeneity is affirmed, then we proceed to the combined analysis, with the following structure of the analysis of variance table for the combined data would be as displayed in the table below.

Source	d.f.	MS	F
Y	$p-1$	Y MS	$\frac{\text{Y MS}}{\text{RMS}}$
R(Y)	$p(r-1)$	R MS	-
T	$t-1$	T MS	$\frac{\text{T MS}}{\text{Y*T MS}}$
Y*T	$(p-1)(t-1)$	Y*T MS	$\frac{\text{Y*T MS}}{EMS}$
Error	$p(t-1)(r-1)$	EMS	–
Total	$tpr-1$		

The following data were adapted from the notes of Krishan Lal (Combined Analysis of Data) and relate to experiments conducted over 4 years (Y) having four treatments (T) and replicated five times. Thus we have a total of $n = 4 \times 4 \times 5 = 80$ observations in the experiments.

Table 20.16 Grain yield (kg/plot) with four replications. Adapted from Krishan Lal (2010)

		Replication				
Year	Treatment	1	2	3	4	5
1	1	33.6	33.7	30.9	33.3	15.0
	2	34.0	27.2	46.2	36.7	11.6
	3	30.5	33.2	15.1	33.3	29.7
	4	30.8	14.4	14.2	9.5	12.0
2	1	28.8	28.8	35.2	41.6	43.2
	2	46.4	43.2	38.4	54.4	57.6
	3	35.2	32.0	32.0	25.6	33.6
	4	51.2	40.0	49.6	51.2	49.6
3	1	30.1	38.1	21.4	17.6	14.3
	2	36.1	18.3	38.0	31.0	26.6
	3	27.2	40.7	15.5	18.1	12.3
	4	37.8	54.5	13.2	18.1	7.3
4	1	23.8	48.8	19.5	28.8	34.4
	2	15.2	39.0	39.8	52.0	31.2
	3	40.2	52.0	33.0	41.2	35.0
	4	43.2	46.8	34.5	44.5	38.0

20.6.1 Analysis

As before we analyze the data first on a year by year basis before we do the combined analysis. The estimated error variances for the 4 years are displayed in the following table.

	Year1	Year2	Year3	Year4
s_i^2	78.23	28.31	108.50	67.90
d.f.	12	12	12	12

Because there are very large variations among the estimated variances, we would use Bartlett's test for homogeneity of variances discussed in earlier sections.

$$S_P^2 = \frac{12(78.23 + 28.31 + 108.50 + 67.90)}{48} = 70.735$$

$$\begin{aligned} q &= 48 \log_{10} 70.735 - \big(12[\log_{10} 78.23 + \log_{10} 28.31 + \log_{10} 108.50 \\ &\quad + \log_{10} 67.90]\big) \\ &= 88.7824 - 12(1.8934 + 1.4519 + 2.0354 + 1.8319) \\ &= 88.7824 - 86.5512 \\ &= 2.2312 \end{aligned}$$

$$\begin{aligned} c &= 1 + \frac{1}{3}\left(\frac{4}{12} - \frac{1}{48}\right) = \frac{15 \times 10}{48 \times 9} \\ &= 0.3472 \end{aligned}$$

Hence,

$$\chi_0^2 = 2.3026 \times \frac{2.2312}{0.3472} = 14.7971$$

The p value corresponding to this value is 0.0020. Since $0.0020 < 0.05$, we would therefore strongly reject the null hypothesis that the error variances are homogeneous. Since the error variances are heterogeneous, we would have to employ the method of weighted least squares for our analysis with the weights ω being the reciprocals of the root mean square errors, that is, $\omega_i = \dfrac{1}{\sqrt{s_i^2}}$. The table below gives the computed values of ω_i for each year extracted from the separate analysis of variance Tables in Table 20.16.

	Year1	Year2	Year3	Year4
s_i^2	78.23	28.31	108.50	67.90
ω_i	0.1131	0.1879	0.0960	0.1214

The ANOVA tables reveal that there are only significant differences in the treatment means for the second year only. The other years do not exhibit significant different means.

Table 20.17 Separate ANOVA tables for the three sites

Source	d.f.	SS	MS	F	p value
		Year 1			
R	4	498.29	124.57	1.59	0.239
T	3	695.36	231.79	2.96	0.075
Error	12	938.81	78.23		
Total	19	2132.45			
		Year 2			
R	4	239.87	59.97	2.12	0.141
T	3	1097.09	365.70	12.92	0.000
Error	12	339.71	28.31		
Total	19	1676.67			
		Year 3			
R	4	1379.1	344.8	3.18	0.054
T	3	146.4	48.8	0.45	0.722
Error	12	1301.6	108.5		
Total	19	2827.1			
		Year 4			
R	4	756.30	189.08	2.78	0.076
T	3	339.14	113.05	1.66	0.227
Error	12	814.83	67.90		
Total	19	1910.27			

Row	Y	T	R	Yd	YY
1	2	1	1	28.8	5.4115
2	2	2	1	46.4	8.7186
3	2	3	1	35.2	6.6141
4	2	4	1	51.2	9.6205
5	2	1	2	28.8	5.4115
6	2	2	2	43.2	8.1173
7	2	3	2	32.0	6.0128
8	2	4	2	40.0	7.5160
9	2	1	3	35.2	6.6141
10	2	2	3	38.4	7.2154
11	2	3	3	32.0	6.0128
12	2	4	3	49.6	9.3198
13	2	1	4	41.6	7.8166
14	2	2	4	54.4	10.2218
15	2	3	4	25.6	4.8102
16	2	4	4	51.2	9.6205
17	2	1	5	43.2	8.1173
18	2	2	5	57.6	10.8230
19	2	3	5	33.6	6.3134
20	2	4	5	49.6	9.3198
.....					

```
 61  4  1  1  28.8   3.4963
 62  4  1  2  52.0   6.3128
 63  4  1  3  41.2   5.0017
 64  4  1  4  44.5   5.4023
 65  4  1  5  15.0   1.8210
 66  4  2  1  11.6   1.4082
 67  4  2  2  29.7   3.6056
 68  4  2  3  12.0   1.4568
 69  4  2  4  43.2   5.2445
 70  4  2  5  57.6   6.9926
 71  4  3  1  33.6   4.0790
 72  4  3  2  49.6   6.0214
 73  4  3  3  14.3   1.7360
 74  4  3  4  26.6   3.2292
 75  4  3  5  12.3   1.4932
 76  4  4  1   7.3   0.8862
 77  4  4  2  34.4   4.1762
 78  4  4  3  31.2   3.7877
 79  4  4  4  35.0   4.2490
 80  4  4  5  38.0   4.6132


MTB > GLM 'YY' = Y R( Y) T Y* T;
SUBC>   Random 'Y';
SUBC>   Brief 2 ;
SUBC>   Means Y* T;
SUBC>   GNormalplot;
SUBC>   GFits;
SUBC>   NoDGraphs;
SUBC>   RType 1 .


General Linear Model: YY versus Y, T, R

Factor  Type    Levels  Values
Y       random       4  1, 2, 3, 4
MTB > GLM 'YY' = Y R( Y) T Y* T;
SUBC>   Random 'Y';
SUBC>   Brief 2 ;
SUBC>   Means Y* T;
SUBC>   GNormalplot;
SUBC>   GFits;
SUBC>   NoDGraphs;
SUBC>   RType 1 .
```

```
General Linear Model: YY versus Y, T, R

Factor  Type    Levels  Values
Y       random       4  1, 2, 3, 4
R(Y)    random      20  1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1. 2, 3, 4, 5, 1, 2, 3,
                        4, 5
T       fixed        4  1, 2, 3, 4


Analysis of Variance for YY, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS   Adj MS      F      P
Y        3  336.067  336.067  112.022  19.90  0.000 x
R(Y)    16   45.396   45.396    2.837   1.77  0.065
T        3   12.729   12.729    4.243   0.96  0.451
Y*T      9   39.582   39.582    4.398   2.74  0.011
Error   48   77.051   77.051    1.605
Total   79  510.825

x Not an exact F-test.


S = 1.26697   R-Sq = 84.92%   R-Sq(adj) = 75.17%

Error Terms for Tests, using Adjusted SS

                                 Synthesis of
   Source  Error DF  Error MS   Error MS
1  Y          11.71     5.630  (2) + (4) - (5)
2  R(Y)       48.00     1.605  (5)
3  T           9.00     4.398  (4)
4  Y*T        48.00     1.605  (5)


Variance Components, using Adjusted SS

       Estimated
Source     Value
Y         5.3196
R(Y)      0.3080
Y*T       0.5586
Error     1.6052


Least Squares Means for YY

Y*T   Mean
1 1  3.314
1 2  3.522
1 3  3.208
1 4  1.830
2 1  6.674
2 2  9.019
2 3  5.953
2 4  9.079
3 1  2.333
3 2  2.880
3 3  2.185
3 4  2.513
4 1  4.407
4 2  3.742
4 3  3.312
4 4  3.542
```

The ANOVA table for the combined analysis indicates that there is very strong significant interactions between years and treatment. The hypothesis that $\sigma^2_{YT} = 0$ is therefore rejected and the estimated variance components is $\hat{\sigma}^2_{YT} = 0.5586$.

20.7 Exercises

1. The data in this exercise relate to the yield in tons/ha for a randomized complete block design on six varieties of sorghum planted over 3 years in three randomized blocks.

Variety	Year 1 replications			Year 2 replications			Year 3 replications		
no.	1	2	3	1	2	3	1	2	3
1	4.72	3.89	4.28	2.66	3.76	4.21	4.84	2.80	3.03
2	3.79	2.85	3.76	3.44	2.12	3.68	3.54	3.38	3.19
3	3.34	2.89	4.26	3.01	2.32	3.70	3.81	3.61	2.68
4	3.72	4.25	3.47	2.53	3.19	3.58	3.94	3.91	3.53
5	3.54	3.86	3.79	3.59	3.64	3.42	3.72	2.77	3.22
6	4.86	3.81	4.25	3.98	2.27	2.92	3.70	3.74	2.10

Yields in tons/ha for the experiment

   (a) Perform a combined analysis over the years.
   (b) Perform a test of homogeneity of variance and draw your conclusions.
   (c) Is there a significant interaction effect between the varieties and years?

2. In an attempt to determine the effect of harvest management on forage yield from 1 Nitro' a cultivar of alfalfa (*Medicago satva* L.), a randomized complete block experiment with four replications was carried out. The treatments consist of four management systems:

   - H_1 : No harvest during the growing season
   - H_2 :Two harvests at bud and herbage regrowth harvested in the fall
   - H_3 : Three harvests at bud and herbage regrowth harvested in the fall
   - H_4 : Two harvests at first flower and herbage regrowth harvested in the fall.

   The data from the experiment are presented in the following table (problem adapted from (Hosmond 2005)).

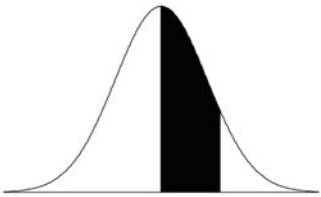
Harvest management	Rep I	Rep II	Rep III	Rep IV
		Summer		
H_1	–	–	–	–
H_2	1.9	2.0	2.1	2.0
H_3	3.0	3.1	3.0	2.8
H_4	2.5	2.4	2.8	2.3
		Fall		
H_1	0.5	2.3	1.4	2.1
H_2	0.7	0.9	0.8	0.7
H_3	0.4	0.6	0.2	0.1
H_4	0.3	0.5	0.3	0.4

Effect of management on forage yield (ton/acre)

(a) Perform a combined analysis over the seasons.
(b) Perform a test of homogeneity of variance. Are the variances homogeneous?
(c) Is there a significant interaction effect between the seasons and harvest management practices?

Appendix: Statistical Tables

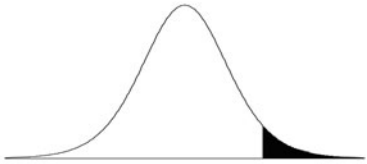
Table 1 Standard normal probabilities (area between 0 and z)



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

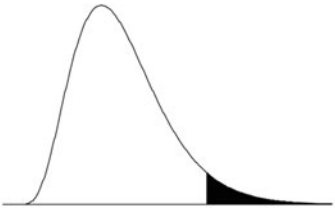
B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8,

Table 2 Values of t_α in a t distribution with df degrees of freedom. (*shaded area* $P(t > t_\alpha) = \alpha$)



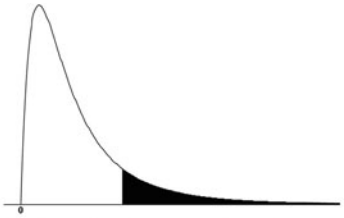
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	df
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
30	1.310	1.697	2.042	2.457	2.750	30
z	1.282	1.645	1.960	2.326	2.576	z

Table 3 Values of $\chi^2_{\alpha,\mathrm{df}}$ in a chi-square distribution with *df* degrees of freedom (*shaded area* $P(\chi^2 > \chi^2_{\alpha,\mathrm{df}}) = \alpha$)



df	$\alpha = .995$	$\alpha = .990$	$\alpha = .975$	$\alpha = .950$	$\alpha = .050$	$\alpha = .025$	$\alpha = .010$	$\alpha = .005$	df
1	0.0000393	0.000157	0.000982	0.00393	3.841	5.024	6.635	7.879	1
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597	2
3	0.0717	0.115	0.216	0.352	7.815	9.348	11.345	12.838	3
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860	4
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750	5
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548	6
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	12
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	13
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319	14
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801	15
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267	16
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718	17
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156	18
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582	19
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997	20
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401	21
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796	22
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181	23
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559	24
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928	25
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290	26
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645	27
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993	28
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336	29
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672	30

Table 4 Values of f_{α,ν_1,ν_2} in an F distribution (*shaded area* $P(F > f_{\alpha,\nu_1,\nu_2}) = \alpha$). Numerator degrees of freedom is ν_1 and denominator degrees of freedom is ν_2.



		ν_1																
ν_2	α	1	2	3	4	5	6	7	8	9	10	11	12	15	20	25	30	40	1000
1	0.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.47	60.71	61.22	61.74	62.05	62.26	62.53	63.30
	0.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98	243.91	245.95	248.01	249.26	250.10	251.14	254.19
	0.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	973.03	976.71	984.87	993.10	998.08	1001.41	1005.60	1017.75
	0.010	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6083.32	6106.32	6157.28	6208.73	6239.83	6260.65	6286.78	6362.68
2	0.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41	9.42	9.44	9.45	9.46	9.47	9.49
	0.050	18.51	19.00	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.49
	0.025	38.51	39.00	39.17	39.25	39.3	39.33	39.36	39.37	39.39	39.40	39.41	39.41	39.43	39.45	39.46	39.46	39.47	39.50
	0.010	98.5	99.00	99.17	99.25	99.3	99.33	99.36	99.37	99.39	99.40	99.41	99.42	99.43	99.45	99.46	99.47	99.47	99.50
3	0.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22	5.20	5.18	5.17	5.17	5.16	5.13
	0.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.70	8.66	8.63	8.62	8.59	8.53
	0.025	17.44	16.04	15.44	15.1	14.88	14.73	14.62	14.54	14.47	14.42	14.37	14.34	14.25	14.17	14.12	14.08	14.04	13.91
	0.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.13	27.05	26.87	26.69	26.58	26.50	26.41	26.14
4	0.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90	3.87	3.84	3.83	3.82	3.80	3.76
	0.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.86	5.80	5.77	5.75	5.72	5.63
	0.025	12.22	10.65	9.98	9.6	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.66	8.56	8.50	8.46	8.41	8.26
	0.010	21.2	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.20	14.02	13.91	13.84	13.75	13.47
5	0.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27	3.24	3.21	3.19	3.17	3.16	3.11
	0.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.62	4.56	4.52	4.50	4.46	4.37
	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.43	6.33	6.27	6.23	6.18	6.02
	0.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.72	9.55	9.45	9.38	9.29	9.03
6	0.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90	2.87	2.84	2.81	2.80	2.78	2.72
	0.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.94	3.87	3.83	3.81	3.77	3.67
	0.025	8.81	7.26	6.6	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.27	5.17	5.11	5.07	5.01	4.86
	0.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.56	7.40	7.30	7.23	7.14	6.89
7	0.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.68	2.67	2.63	2.59	2.57	2.56	2.54	2.47
	0.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.51	3.44	3.40	3.38	3.34	3.23
	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.57	4.47	4.40	4.36	4.31	4.15
	0.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.31	6.16	6.06	5.99	5.91	5.66
8	0.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.52	2.50	2.46	2.42	2.40	2.38	2.36	2.30
	0.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.22	3.15	3.11	3.08	3.04	2.93
	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24	4.20	4.10	4.00	3.94	3.89	3.84	3.68
	0.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.52	5.36	5.26	5.20	5.12	4.87
9	0.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.40	2.38	2.34	2.30	2.27	2.25	2.23	2.16
	0.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.01	2.94	2.89	2.86	2.83	2.71
	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.77	3.67	3.60	3.56	3.51	3.34
	0.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	4.96	4.81	4.71	4.65	4.57	4.32
10	0.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.30	2.28	2.24	2.20	2.17	2.16	2.13	2.06
	0.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.85	2.77	2.73	2.70	2.66	2.54
	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.52	3.42	3.35	3.31	3.26	3.09
	0.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.56	4.41	4.31	4.25	4.17	3.92
11	0.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.23	2.21	2.17	2.12	2.10	2.08	2.05	1.98
	0.050	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.72	2.65	2.60	2.57	2.53	2.41
	0.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47	3.43	3.33	3.23	3.16	3.12	3.06	2.89
	0.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.25	4.10	4.01	3.94	3.86	3.61
12	0.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.17	2.15	2.10	2.06	2.03	2.01	1.99	1.91
	0.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.62	2.54	2.50	2.47	2.43	2.30
	0.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.18	3.07	3.01	2.96	2.91	2.73
	0.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.01	3.86	3.76	3.70	3.62	3.37

Table 4 Values of f_{α,ν_1,ν_2} in an F distribution (continued)

		ν_1																	
ν_2	α	1	2	3	4	5	6	7	8	9	10	11	12	15	20	25	30	40	1000
13	0.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.12	2.10	2.05	2.01	1.98	1.96	1.93	1.85
	0.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.53	2.46	2.41	2.38	2.34	2.21
	0.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.05	2.95	2.88	2.84	2.78	2.60
	0.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.82	3.66	3.57	3.51	3.43	3.18
14	0.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.07	2.05	2.01	1.96	1.93	1.91	1.89	1.80
	0.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.46	2.39	2.34	2.31	2.27	2.14
	0.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09	3.05	2.95	2.84	2.78	2.73	2.67	2.50
	0.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.66	3.51	3.41	3.35	3.27	3.02
16	0.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	2.01	1.99	1.94	1.89	1.86	1.84	1.81	1.72
	0.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.35	2.28	2.23	2.19	2.15	2.02
	0.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.79	2.68	2.61	2.57	2.51	2.32
	0.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.41	3.26	3.16	3.10	3.02	2.76
18	0.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.95	1.93	1.89	1.84	1.80	1.78	1.75	1.66
	0.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.27	2.19	2.14	2.11	2.06	1.92
	0.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81	2.77	2.67	2.56	2.49	2.44	2.38	2.20
	0.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.23	3.08	2.98	2.92	2.84	2.58
20	0.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.91	1.89	1.84	1.79	1.76	1.74	1.71	1.61
	0.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.20	2.12	2.07	2.04	1.99	1.85
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.57	2.46	2.40	2.35	2.29	2.09
	0.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.09	2.94	2.84	2.78	2.69	2.43
22	0.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.88	1.86	1.81	1.76	1.73	1.70	1.67	1.57
	0.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.15	2.07	2.02	1.98	1.94	1.79
	0.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65	2.60	2.50	2.39	2.32	2.27	2.21	2.01
	0.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	2.98	2.83	2.73	2.67	2.58	2.32
24	0.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.85	1.83	1.78	1.73	1.70	1.67	1.64	1.54
	0.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.11	2.03	1.97	1.94	1.89	1.74
	0.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59	2.54	2.44	2.33	2.26	2.21	2.15	1.94
	0.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.89	2.74	2.64	2.58	2.49	2.22
26	0.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.83	1.81	1.76	1.71	1.67	1.65	1.61	1.51
	0.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.07	1.99	1.94	1.90	1.85	1.70
	0.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.54	2.49	2.39	2.28	2.21	2.16	2.09	1.89
	0.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.81	2.66	2.57	2.50	2.42	2.14
28	0.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.81	1.79	1.74	1.69	1.65	1.63	1.59	1.48
	0.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.04	1.96	1.91	1.87	1.82	1.66
	0.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.49	2.45	2.34	2.23	2.16	2.11	2.05	1.84
	0.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.75	2.60	2.51	2.44	2.35	2.08
30	0.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.79	1.77	1.72	1.67	1.63	1.61	1.57	1.46
	0.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.01	1.93	1.88	1.84	1.79	1.63
	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.31	2.20	2.12	2.07	2.01	1.80
	0.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.70	2.55	2.45	2.39	2.30	2.02
40	0.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.74	1.71	1.66	1.61	1.57	1.54	1.51	1.38
	0.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.92	1.84	1.78	1.74	1.69	1.52
	0.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33	2.29	2.18	2.07	1.99	1.94	1.88	1.65
	0.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.52	2.37	2.27	2.20	2.11	1.82
1000	0.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64	1.61	1.58	1.55	1.49	1.43	1.38	1.35	1.30	1.08
	0.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.80	1.76	1.68	1.58	1.52	1.47	1.41	1.11
	0.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13	2.06	2.01	1.96	1.85	1.72	1.64	1.58	1.50	1.13
	0.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.27	2.20	2.06	1.90	1.79	1.72	1.61	1.16

Table 5 Orthogonal polynomial coefficients

		Coefficients								
k	Polynomial	1	2	3	4	5	6	7	8	$\sum c_i^2$
3	Linear	−1	0	1						2
	Quadratic	1	−2	1						6
4	Linear	−3	−1	1	3					20
	Quadratic	1	−1	−1	1					4
	Cubic	−1	3	−3	1					20
5	Linear	−2	−1	0	1	2				10
	Quadratic	2	−1	−2	−1	2				14
	Cubic	−1	2	0	−2	1				10
	Quartic	1	−4	6	−4	1				70
6	Linear	−5	−3	−1	1	3	5			70
	Quadratic	5	−1	−4	−4	−1	5			84
	Cubic	−5	7	4	−4	−7	5			180
	Quartic	1	−3	2	2	−3	1			28
	Quintic	−1	5	−10	10	−5	1			252
7	Linear	−3	−2	−1	0	1	2	3		28
	Quadratic	5	0	−3	−4	−3	0	5		84
	Cubic	−1	1	1	0	−1	−1	1		6
	Quartic	3	−7	1	6	1	−7	3		154
	Quintic	−1	4	−5	0	5	−4	1		84
	Sextic	1	−6	15	−20	15	−6	1		924
8	Linear	−7	−5	−3	−1	1	3	5	7	168
	Quadratic	7	1	−3	−5	−5	−3	1	7	168
	Cubic	−7	5	7	3	−3	−7	−5	7	264
	Quartic	7	−13	−3	9	9	−3	−13	7	616
	Quintic	−7	23	−17	−15	15	17	−23	7	2184
	Sextic	1	−5	9	−5	−5	9	−5	1	264
	Septic	−1	7	−21	35	−35	21	−7	1	3432

Table 6 Upper α point of Studentized range, $q_\alpha(k, \nu)$, where $k = r$ = number of treatments to be compared and ν = the number of degrees of freedom. (Source: The Analysis of Variance by Scheffé, H. (1959). Reproduced with permission from John Wiley & Sons, Inc.)

$1 - \alpha = .95$

	r																		
ν	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	18.0	27.0	32.8	37.1	40.4	43.1	45.4	47.4	49.1	50.6	52.0	53.2	54.3	55.4	56.3	57.2	58.0	58.8	59.6
2	6.08	8.33	9.80	10.9	11.7	12.4	13.0	13.5	14.0	14.4	14.7	15.1	15.4	15.7	15.9	16.1	16.4	16.6	16.8
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	9.72	9.95	10.2	10.3	10.5	10.7	10.8	11.0	11.1	11.2
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.03	8.21	8.37	8.52	8.66	8.79	8.91	9.03	9.13	9.23
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65	6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.10	7.17
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.11	6.19	6.27	6.34	6.40	6.47
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.98	6.06	6.13	6.20	6.27	6.33
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51	5.61	5.71	5.80	5.88	5.95	6.02	6.09	6.15	6.21
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79	5.86	5.93	5.99	6.05	6.11
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.55	5.64	5.71	5.79	5.85	5.91	5.97	6.03
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.57	5.65	5.72	5.78	5.85	5.90	5.96
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.90
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.54	5.61	5.67	5.73	5.79	5.84
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23	5.31	5.39	5.46	5.53	5.59	5.65	5.70	5.75
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32	5.38	5.44	5.49	5.55	5.59
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.47
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82	4.90	4.98	5.04	5.11	5.16	5.22	5.27	5.31	5.36
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00	5.06	5.11	5.15	5.20	5.24
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71	4.78	4.84	4.90	4.95	5.00	5.04	5.09	5.13
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.68	4.74	4.80	4.85	4.89	4.93	4.97	5.01

Table 7 Upper α point of Studentized range (continued)

$1 - \alpha = .99$

	r																		
ν	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	90.0	135	164	186	202	216	227	237	246	253	260	266	272	277	282	286	290	294	298
2	14.0	19.0	22.3	24.7	26.6	28.2	29.5	30.7	31.7	32.6	33.4	34.1	34.8	35.4	36.0	36.5	37.0	37.5	37.9
3	8.26	10.6	12.2	13.3	14.2	15.0	15.6	16.2	16.7	17.1	17.5	17.9	18.2	18.5	18.8	19.1	19.3	19.5	19.8
4	6.51	8.12	9.17	9.96	10.6	11.1	11.5	11.9	12.3	12.6	12.8	13.1	13.3	13.5	13.7	13.9	14.1	14.2	14.4
5	5.70	6.97	7.80	8.42	8.91	9.32	9.67	9.97	10.2	10.5	10.7	10.9	11.1	11.2	11.4	11.6	11.7	11.8	11.9
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.49	9.65	9.81	9.95	10.1	10.2	10.3	10.4	10.5
7	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55	8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65
8	4.74	5.63	6.20	6.63	6.96	7.24	7.47	7.68	7.87	8.03	8.18	8.31	8.44	8.55	8.66	8.76	8.85	8.94	9.03
9	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.32	7.49	7.65	7.78	7.91	8.03	8.13	8.23	8.32	8.41	8.49	8.57
10	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36	7.48	7.60	7.71	7.81	7.91	7.99	8.07	8.15	8.22
11	4.39	5.14	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13	7.25	7.36	7.46	7.56	7.65	7.73	7.81	7.88	7.95
12	4.32	5.04	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73
13	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90	7.01	7.10	7.19	7.27	7.34	7.42	7.48	7.55
14	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66	6.77	6.87	6.96	7.05	7.12	7.20	7.27	7.33	7.39
15	4.17	4.83	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55	6.66	6.76	6.84	6.93	7.00	7.07	7.14	7.20	7.26
16	4.13	4.78	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46	6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15
17	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38	6.48	6.57	6.66	6.73	6.80	6.87	6.94	7.00	7.05
18	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41	6.50	6.58	6.65	6.72	6.79	6.85	6.91	6.96
19	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25	6.34	6.43	6.51	6.58	6.65	6.72	6.78	6.84	6.89
20	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.29	6.37	6.45	6.52	6.59	6.65	6.71	6.76	6.82
24	3.96	4.54	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02	6.11	6.19	6.26	6.33	6.39	6.45	6.51	6.56	6.61
30	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85	5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41
40	3.82	4.37	4.70	4.93	5.11	5.27	5.39	5.50	5.60	5.69	5.77	5.84	5.90	5.96	6.02	6.07	6.12	6.17	6.21
60	3.76	4.28	4.60	4.82	4.99	5.13	5.25	5.36	5.45	5.53	5.60	5.67	5.73	5.79	5.84	5.89	5.93	5.98	6.02
120	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.38	5.44	5.51	5.56	5.61	5.66	5.71	5.75	5.79	5.83
∞	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23	5.29	5.35	5.40	5.45	5.49	5.54	5.57	5.61	5.65

Table 8 Upper $\alpha = 0.05$ points of Duncan's multiple range tests

	k								
df	2	3	4	5	6	7	8	9	10
1	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969	17.969
2	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085
3	4.501	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516
4	3.926	4.013	4.033	4.033	4.033	4.033	4.033	4.033	4.033
5	3.635	3.749	3.796	3.814	3.814	3.814	3.814	3.814	3.814
6	3.460	3.586	3.649	3.680	3.694	3.697	3.697	3.697	3.697
7	3.344	3.477	3.548	3.588	3.611	3.622	3.625	3.625	3.625
8	3.261	3.398	3.475	3.521	3.549	3.566	3.575	3.579	3.579
9	3.199	3.339	3.420	3.470	3.502	3.523	3.536	3.544	3.547
10	3.151	3.293	3.376	3.430	3.465	3.489	3.505	3.516	3.522
11	3.113	3.256	3.341	3.397	3.435	3.462	3.480	3.493	3.501
12	3.081	3.225	3.312	3.370	3.410	3.439	3.459	3.474	3.484
13	3.055	3.200	3.288	3.348	3.389	3.419	3.441	3.458	3.470
14	3.033	3.178	3.268	3.328	3.371	3.403	3.426	3.444	3.457
15	3.014	3.160	3.250	3.312	3.356	3.389	3.413	3.432	3.446
16	2.998	3.144	3.235	3.297	3.343	3.376	3.402	3.422	3.437
17	2.984	3.130	3.222	3.285	3.331	3.365	3.392	3.412	3.429
18	2.971	3.117	3.210	3.274	3.320	3.356	3.383	3.404	3.421
19	2.960	3.106	3.199	3.264	3.311	3.347	3.375	3.397	3.415
20	2.950	3.097	3.190	3.255	3.303	3.339	3.368	3.390	3.409
21	2.941	3.088	3.181	3.247	3.295	3.332	3.361	3.385	3.403
22	2.933	3.080	3.173	3.239	3.288	3.326	3.355	3.379	3.398
23	2.926	3.072	3.166	3.233	3.282	3.320	3.350	3.374	3.394
24	2.919	3.066	3.160	3.226	3.276	3.315	3.345	3.370	3.390
25	2.913	3.059	3.154	3.221	3.271	3.310	3.341	3.366	3.386
26	2.907	3.054	3.149	3.216	3.266	3.305	3.336	3.362	3.382
27	2.902	3.049	3.144	3.211	3.262	3.301	3.332	3.358	3.379
28	2.897	3.044	3.139	3.206	3.257	3.297	3.329	3.355	3.376
29	2.892	3.039	3.135	3.202	3.253	3.293	3.326	3.352	3.373
30	2.888	3.035	3.131	3.199	3.250	3.290	3.322	3.349	3.371
31	2.884	3.031	3.127	3.195	3.246	3.287	3.319	3.346	3.368
32	2.881	3.028	3.123	3.192	3.243	3.284	3.317	3.344	3.366
33	2.877	3.024	3.120	3.188	3.240	3.281	3.314	3.341	3.364
34	2.874	3.021	3.117	3.185	3.238	3.279	3.312	3.339	3.362
35	2.871	3.018	3.114	3.183	3.235	3.276	3.309	3.337	3.360
36	2.868	3.015	3.111	3.180	3.232	3.274	3.307	3.335	3.358
37	2.865	3.013	3.109	3.178	3.230	3.272	3.305	3.333	3.356
38	2.863	3.010	3.106	3.175	3.228	3.270	3.303	3.331	3.355
39	2.861	3.008	3.104	3.173	3.226	3.268	3.301	3.330	3.353
40	2.858	3.005	3.102	3.171	3.224	3.266	3.300	3.328	3.352
48	2.843	2.991	3.087	3.157	3.211	3.253	3.288	3.318	3.342
60	2.829	2.976	3.073	3.143	3.198	3.241	3.277	3.307	3.333
80	2.814	2.961	3.059	3.130	3.185	3.229	3.266	3.297	3.323
120	2.800	2.947	3.045	3.116	3.172	3.217	3.254	3.286	3.313
240	2.786	2.933	3.031	3.103	3.159	3.205	3.243	3.276	3.304
Inf	2.772	2.918	3.017	3.089	3.146	3.193	3.232	3.265	3.294

Table 9 Upper $\alpha = 0.01$ point of Duncan's multiple range tests

	k									
df	2	3	4	5	6	7	8	9	10	11
1	90.024	90.024	90.024	90.024	90.024	90.024	90.024	90.024	90.024	90.024
2	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036	14.036
3	8.260	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321
4	6.511	6.677	6.740	6.755	6.755	6.755	6.755	6.755	6.755	6.755
5	5.702	5.893	5.989	6.040	6.065	6.074	6.074	6.074	6.074	6.074
6	5.243	5.439	5.549	5.614	5.655	5.680	5.694	5.701	5.703	5.703
7	4.949	5.145	5.260	5.333	5.383	5.416	5.439	5.454	5.464	5.470
8	4.745	4.939	5.056	5.134	5.189	5.227	5.256	5.276	5.291	5.302
9	4.596	4.787	4.906	4.986	5.043	5.086	5.117	5.142	5.160	5.174
10	4.482	4.671	4.789	4.871	4.931	4.975	5.010	5.036	5.058	5.074
11	4.392	4.579	4.697	4.780	4.841	4.887	4.923	4.952	4.975	4.994
12	4.320	4.504	4.622	4.705	4.767	4.815	4.852	4.882	4.907	4.927
13	4.260	4.442	4.560	4.643	4.706	4.754	4.793	4.824	4.850	4.871
14	4.210	4.391	4.508	4.591	4.654	4.703	4.743	4.775	4.802	4.824
15	4.167	4.346	4.463	4.547	4.610	4.660	4.700	4.733	4.760	4.783
16	4.131	4.308	4.425	4.508	4.572	4.622	4.662	4.696	4.724	4.748
17	4.099	4.275	4.391	4.474	4.538	4.589	4.630	4.664	4.692	4.717
18	4.071	4.246	4.361	4.445	4.509	4.559	4.601	4.635	4.664	4.689
19	4.046	4.220	4.335	4.418	4.483	4.533	4.575	4.610	4.639	4.664
20	4.024	4.197	4.312	4.395	4.459	4.510	4.552	4.587	4.617	4.642
21	4.004	4.177	4.291	4.374	4.438	4.489	4.531	4.567	4.597	4.622
22	3.986	4.158	4.272	4.355	4.419	4.470	4.513	4.548	4.578	4.604
23	3.970	4.141	4.254	4.337	4.402	4.453	4.496	4.531	4.562	4.588
24	3.955	4.126	4.239	4.322	4.386	4.437	4.480	4.516	4.546	4.573
25	3.942	4.112	4.224	4.307	4.371	4.423	4.466	4.502	4.532	4.559
26	3.930	4.099	4.211	4.294	4.358	4.410	4.452	4.489	4.520	4.546
27	3.918	4.087	4.199	4.282	4.346	4.397	4.440	4.477	4.508	4.535
28	3.908	4.076	4.188	4.270	4.334	4.386	4.429	4.465	4.497	4.524
29	3.898	4.065	4.177	4.260	4.324	4.376	4.419	4.455	4.486	4.514
30	3.889	4.056	4.168	4.250	4.314	4.366	4.409	4.445	4.477	4.504
31	3.881	4.047	4.159	4.241	4.305	4.357	4.400	4.436	4.468	4.495
32	3.873	4.039	4.150	4.232	4.296	4.348	4.391	4.428	4.459	4.487
33	3.865	4.031	4.142	4.224	4.288	4.340	4.383	4.420	4.452	4.479
34	3.859	4.024	4.135	4.217	4.281	4.333	4.376	4.413	4.444	4.472
35	3.852	4.017	4.128	4.210	4.273	4.325	4.369	4.406	4.437	4.465
36	3.846	4.011	4.121	4.203	4.267	4.319	4.362	4.399	4.431	4.459
37	3.840	4.005	4.115	4.197	4.260	4.312	4.356	4.393	4.425	4.452
38	3.835	3.999	4.109	4.191	4.254	4.306	4.350	4.387	4.419	4.447
39	3.830	3.993	4.103	4.185	4.249	4.301	4.344	4.381	4.413	4.441
40	3.825	3.988	4.098	4.180	4.243	4.295	4.339	4.376	4.408	4.436
48	3.793	3.955	4.064	4.145	4.209	4.261	4.304	4.341	4.374	4.402
60	3.762	3.922	4.030	4.111	4.174	4.226	4.270	4.307	4.340	4.368
80	3.732	3.890	3.997	4.077	4.140	4.192	4.236	4.273	4.306	4.335
120	3.702	3.858	3.964	4.044	4.107	4.158	4.202	4.239	4.272	4.301
240	3.672	3.827	3.932	4.011	4.073	4.125	4.168	4.206	4.239	4.268
Inf	3.643	3.796	3.900	3.978	4.040	4.091	4.135	4.172	4.205	4.235

Bibliography

Angela M. Dean and Daniel Voss (1999) *Design and Analysis of Experiments*. Springer, New York.

Armitage, P. and Berry, G. (1985). *Statistical Methods in Medical Research*. 2nd Edition. Blackwell Scientific Publications. London.

Bates, D. and Watts, D. (1988) *Nonlinear Regression Analysis and Its Applications*. John Wiley & Sons. Inc. New York, NY.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press.

Blaisdell, E. A. (1993). *Statistics in Practice*. Saunders College Publishing. Fort Worth.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of Royal Statistical Society, Series B*, 26: 211–246.

Breslow, N.E., & Day, N.E. (1980). *Statistical Methods in Cancer Research*. Lyon: International Agency for Research on Cancer.

Brown S., Selvin, S. and Winkelstein, W. J. (1975). The association of economic status with the occurrence of lung cancer. *Cancer*, 36(5):1903–11.

Christensen, R. (1990). *Log-Linear Models*. Springer-Verlag, New York.

Cochran, W.G & Cox, G.M. (1957). *Experimental Designs*. John Wiley and Sons. New York.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc.*, **B74**:187–220.

Daniel, W. W. (1999). *Biostatistics: A foundation for Analysis in the Health Sciences*. Seventh ed. Wiley. New York.

Donald Weber and John H. Skillings (2000). A first course in the Design of Experiments. A linear Models Approach. CRC Press. Boca Raton.

Festing, MWT and Altman, D.G. (2002). Guidelines for the design and Statistical analysis of experiments using laboratory animals. ILAR J 43 (supp) 000–000

Finney, D. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika*, 34: 320–334.

Goldstein, A. (1965). *Biostatistics: An Introductory Text*. New York: Macmillan.

Gomez, K.A. & Gomez, A.A. (1984) *Statistical Procedures for Agricultural Research*. 2nd Edition. Wiley & Sons. Canada

Graybill, F. and Iyer, H. (1994). *Regression Analysis: concepts and Applications*. Duxbury Press. Belmont, CA.

Hastie, T.J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & hall/CRC monographs. London

Hoshmand A. R. (1994). *Experimental Research Design and Analysis: a practical approach for agricultural and Natural Sciences*. CRC Press, Inc., Boca Raton, Florida.

Keppel, G. (1991). *Design and Analysis: A Researcher's Handbook*. 3rd ed. Englewood Cliffs: Prentice Hill

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8,

Kuehl, R.O. (1994). *Statistical Principles of Research Design and Analysis*. Duxbury Press. Belmont, CA.

Kutner, M., Nachtsheim, C., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill & Irwin, New York, NY. Fifth edition.

Lawal, H.B. (1980). Tables of percentage points of Pearson's goodness-of-fit statistic for use with small expectations. *Appl. Statist.*, 29, 292–298.

Lawal, H.B. (1989). On the X^2 statistic for testing independence in two-way contingency tables *AMSE Review*, 12: 37–51.

Lawal, B. (2003). *Categorical Data Analysis with SAS and SPSS Applications*. Lawrence Erbaum Assoc., New Jersey.

Lawal, H.B., & Upton, G.J.G. (1984). On the use of X^2 as a test of independence in contingency tables with small cell expectations. *Australian J. Statist.*, 26, 75–85.

Litchfield, J.T. & Wilcoxon, F. A. (1949). A simplified method of evaluating dose-response Experiments. *J. Pharmacol. Exp. Ther.*, 96(2): 99–113.

Lombard H. L. and Doering C.R. (1947). Treatment of the four-fold table by partial correlation as it relates to public health problems. *Biometrics*, **3**: 123–128

Lunneborg, C. E. (1994). *Modeling Experimental and Observational Data*. Duxbury. Belmont, CA.

Mead, R. & Curnow, R.N. (1983). *Statistical Methods in Agriculture and Experimental Biology*. Chapman and Hall. London.

Michaelis, L. and Menten, M. (1913). Die kinetik der invertinwirkung. *Biochemische Zeitschrift*, 49:333–369.

Montgomery, Douglas (2005) *Design and Analysis of Experiments*. 5th ed. John Wiley & Sons, New York.

Pagano, M. & Gauvreau, K. (2000). *Principles of Biostatistics*. 2nd Edition. Duxbury, CA.

Pearce, S. E. (1983). *The Agricultural Field Experiment: A Statistical Examination of Theory and Practice*. Chichester, England: John Wiley & Sons.

Rosner, B. (2000). *Fundamentals of Biostatistics*. 5th Edition. Duxbury. Belmont, CA.

Samuels, M. L. and Witmer, J. A. (1999). *Statistics for the Life Sciences*. 2nd edition. Prentice Hall. New Jersey.

Schork, M. A. and and Remington, R. D. (2000). *Statistics with Applications to the Biological and Health Sciences*. Prentice Hall, Englewood Cliffs, NJ.

Sedmak, D., Meineke, T., Knechtges, D., and Anderson, J. (1989). Prognostic significance of cytokeratin-positive breast cancer metases. *Modern Pathology*, 2:516–520.

Source: Experimentation in Biology by Ridgman, V.J., pg. 55

Steel, R. G.D. & Torrie, J.H. (1960). *Principles and Procedures of Statistics*. McGraw-Hill Book Company, New York.

Winer, B.J., Brown, D.R., & Michaels, K.M. (1991). *Statistical Analysis in Experimental Design*. McGraw-Hill, New York.

Woodward, G., Lange, S.W., Nelson, K.W., & Calvert, H.O. (1941). The acute oral toxicity of acetic, chloracetic, dichloracetic and trichloracetic acids. *J. of Industrial Hygience and Toxicology*, 23, 78–81.

Wu, C. F. J. and Hamada, Michael, S. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. Wiley, New York.

Yarnold, J.K. (1970). The minimum expectation in χ^2 goodness-of-fit tests and the accuracy of approximations for the null distribution. *J. Amer. Statist. Assoc.*, 65, 865–886.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *J. R. Statist. Soc., Suppl., 1,* 217–235.

Zar, J. H. (1999). *Biostatistical Analysis*. Prentice Hall, New Jersey. Fourth Edition.
The two books at home

Credits

(This page contains an extension of the copyright page.)

We have made every effort to trace the ownership of all copyrighted material and to secure permission from copyright holders. In the event of any question arising as to the use of any material, we will be happy to make the necessary correction in future printings including e-printing.

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8,

Index

B. Lawal, *Applied Statistical Methods in Agriculture, Health and Life Sciences*, DOI 10.1007/978-3-319-05555-8,