The Essentials of Biostatistics for Physicians, Nurses, and Clinicians

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians

Michael R. Chernick
Lankenau Institute for Medical Research
Wynnewood, PA

WILEY
A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Chernick, Michael R.
The essentials of biostatistics for physicians, nurses, and clinicians / Michael R. Chernick.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-64185-9 (pbk.)
1. Biometry. I. Title.
[DNLM: 1. Biostatistics. WA 950]
QH323.5C484 2011
570.1'5195–dc22

2011002198

oBook ISBN: 978-1-118-07195-3
ePDF ISBN: 978-1-118-07193-9
ePub ISBN: 978-1-118-07194-6

Printed in Singapore.

10 9 8 7 6 5 4 3 2 1

Contents

Preface

I have taught biostatistics in the health sciences and published a book in 2003 with Wiley on that topic. That book is a textbook for upper-level undergraduates and graduate students in the health science departments at universities. Since coming to the Lankenau Institute 17 months ago, I was tasked to prepare a course in biostatistics for nurses and physicians (particularly the hospital residents and fellows that do medical research). I quickly learned that although the material in my book was relevant, it contained too much material and was not in a digestible form for them. I prepared a six-lecture course (1 hour each) for physicians, and a two-lecture course for the nurses. To prevent boredom, I introduced some funny but educational cartoon slides. The course currently exists and has been refined as PowerPoint presentations and has been moderately successful. I also am starting a similar course at statistics.com.

The physicians and nurses have a busy schedule, and what they need is a concise and clearly explained set of lectures that cover only the areas of statistics that are essential to know about in medical research. This means topics that are not taught in traditional introductory statistics courses. So Kaplan–Meier curves, repeated measures analysis of variance, hazard ratios, contingency tables, logrank tests, bioequivalence, cross-over designs, noninferiority, selection bias, and group sequential methods are all included, but they are introduced on a conceptual level without the need for theory. It is when and why these methods work that they need to know, and not a detailed account of how they work mathematically. I feel that it would be appropriate to have a textbook for such a course that can be taught in-house at research centers or online courses. The book is intended to be approximately 160 pages along with suitable references.

I am very grateful to Professor Marlene Egger, who carefully reviewed the manuscript and made several wonderful suggestions that helped with the clarity and improved the content of the book.

Michael R. Chernick

CHAPTER 1

The What, Why, and How of Biostatistics in Medical Research

1.1 DEFINITION OF STATISTICS AND BIOSTATISTICS

The *Oxford Dictionary of Statistics* (2002, p. 349) defines statistics as "The science of collecting, displaying, and analyzing data." Statistics is important in any scientific endeavor. It also has a place in the hearts of fans of sports, particularly baseball. Roger Angel in his baseball book, *Late Innings*, says "Statistics are the food of love."

Biostatistics is the branch of statistics that deals with biology, both experiments on plants, animals, and living cells, and controlled experiments on humans, called clinical trials. Statistics is classified by scientific discipline because in addition to many standard methods that are common to statistical problems in many fields, special methods have been developed primarily for certain disciplines. So to illustrate, in biostatistics, we study longitudinal data, missing data models, multiple testing, equivalence and noninferiority testing, relative risk and odds ratios, group sequential and adaptive designs, and survival analysis, because these types of data and methods arise in clinical trials and other medical studies. Engineering statistics considers tolerance intervals and design of experiments. Environmental statistics has a concentration in

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

the analysis of spatial data, and so does geostatistics. Econometrics is the branch of statistics studied by economists, and deals a lot with forecasting and time series.

Statisticians are professionals trained in the collection, display, and analysis of data and the distribution theory that characterizes the variability of data. To become a good applied statistician, one needs to learn probability theory and the methods of statistical inference as developed by Sir Ronald A. Fisher, Jerzy Neyman, Sir Harold Jeffreys, Jimmie Savage, Bruno deFinetti, Harald Cramer, Will Feller, A. N. Kolmogorov, David Blackwell, Erich Lehmann, C. R. Rao, Karl and Egon Pearson, Abraham Wald, George Box, William Cochran, Fred Mosteller, Herman Chernoff, David Cox, and John Tukey in the twentieth century. These are some of the major developers of the foundations of probability and statistics. Of course, when selecting a list of famous contributors like this, many have been unintentionally omitted. In the late twentieth century and early twenty-first century, computer-intensive statistics arose, and a partial list of the leaders of that development are Brad Efron, Leo Brieman, David Freedman, Terry Speed, Jerry Friedman, David Siegmund, and T. L. Lai. In the area of biostatistics, we should mention Thomas Fleming, Stuart Pocock, Nathan Mantel, Peter Armitage, Shein-Chung Chow, Jen-pei Liu, and Gordon Lan. You will be introduced to these and other famous probabilists and statisticians in this book. An applied statistician must also become familiar with at least one scientific discipline in order to effectively consult with scientists in that field.

Statistics is its own discipline because it is much more than just a set of tools to analyze data. Although statistics requires the tools of probability, which are mathematical, it should not be thought of as a branch of mathematics. It is the appropriate way to summarize and analyze data when the data contains an element of uncertainty. This is very common when measurements are taken, since there is a degree of inaccuracy in every measurement. Statisticians develop mathematical models to describe the phenomena being studied. These models may describe such things as the time a bus will arrival at a scheduled stop, how long a person waits in line at a bank, the time until a patient dies or has a recurrence of a disease, or future prices of stocks, bonds, or gasoline.

Based on these models, the statistician develops methods of estimation or tests of hypotheses to solve certain problems related to the data.

Because almost every experiment involves uncertainty, statistics is the scientific method for quantitative data analysis.

Yet in the public eye, statistics and statisticians do not have a great reputation. In the course of a college education, students in the health sciences, business, psychology, and sociology are all required to take an introductory statistics course. The comments most common from these students are "this is the most boring class I ever took" and "it was so difficult, that I couldn't understand any of it." This is the fault of the way the courses are taught and not the fault of the subject. An introductory statistics course can be much easier to understand and more useful to the student than, say, a course in abstract algebra, topology, and maybe even introductory calculus. Yet many people don't view it that way.

Also, those not well trained in statistics may see articles in medicine that are contradictory but still make their case through the use of statistics. This causes many of us to say "You can prove anything with statistics." Also, there is that famous quote attributed to Disraeli. "There are lies, damn lies and statistics." In 1954, Darrell Huff wrote his still popular book, *How to Lie with Statistics*. Although the book shows how graphs and other methods can be used to distort the truth or twist it, the main point of the book is to get a better understanding of these methods so as not to be fooled by those who misuse them. Statisticians applying valid statistical methods will reach consistent conclusions. The data doesn't lie. It is the people that manipulate the data that lie. Four books that provide valuable lessons about misusing statistics are Huff (1954), Campbell (1974), Best (2001), and Hand (2008).

1.2 WHY STUDY STATISTICS?

The question is really why should medical students, physicians, nurses, and clinicians study statistics? Our focus is on biostatistics and the students we want to introduce it to. One good reason to study statistics is to gain knowledge from data and use it appropriately. Another is to make sure that we are not to be fooled by the lies, distortions, and misuses in the media and even some medical journals. The medical journals now commonly require good statistical methods as part of a research paper, and the sophistication of the methods used is greater.

So we learn statistics so that we know what makes sense when reading the medical literature, and in order to publish good research.

We also learn statistics so that we can provide intelligent answers to basic questions of a statistical nature. For many physicians and nurses, there is a fear of statistics. Perhaps this comes from hearing horror stories about statistics classes. It also may be that you have seen applications of statistics but did not understand it because you have no training. So this text is designed to help you conquer your fear of statistics. As you learn and gain confidence, you will see that it is logical and makes sense, and is not as hard as you first thought.

Major employers of statisticians are the pharmaceutical, biotechnology, and medical device companies. This is because the marketing of new drugs, biologics, and most medical devices must be approved by the U.S. Food and Drug Administration (FDA), and the FDA requires the manufacturers to demonstrate through the use of animal studies and controlled clinical trials the safety and effectiveness of their product. These studies must be conducted using valid statistical methods. So any medical investigator involved in clinical trials sponsored by one of these companies really needs to understand the design of the trial and the statistical implications of the design and the sample size requirements (i.e., number of patients need in the clinical trial). This requires at least one basic biostatistics course or good on-the-job training.

Because of uncontrolled variability in any experimental situation, statistics is necessary to organize the data and summarize it in a way so that signals (important phenomena) can be detected when corrupted by noise. Consequently, bench scientists as well as clinical researchers need some acquaintance with statistics. Most medical discoveries need to be demonstrated using statistical hypothesis testing or confidence interval estimation. This has increased in importance in the medical journals. Simple *t*-tests are not always appropriate. Analyses are getting much more sophisticated. Death and other time-to-event data require statistical survival analysis methods for comparison purposes.

Most scientific research requires statistical analysis. When Dr. Riffenburgh (author of the text *Statistics in Medicine*, 1999) is told by a physician "I'm too busy treating patients to do research," he answers, "When you treat a patient, you have treated a patient. When you do research, you have treated ten thousand patients."

In order to amplify these points, I will now provide five examples from my own experience in the medical device and pharmaceutical

industries where a little knowledge of statistics would have made life easier for some of my coworkers.

In the first scenario, suppose you are the coordinator for a clinical trial on an ablation catheter. You are enrolling subjects at five sites. You want to add a new site to help speed up enrollment. The IRB for the new site must review and approve your protocol for the site to enter your study. A member of the IRB asks what stopping rule you use for safety. How do you respond? You don't even know what a stopping rule is or even that the question is related to statistics! By taking this course, you will learn that statisticians construct stopping rules based upon accumulated data. In this case, there may be safety issues, and the stopping rule could be based on reaching a high number of adverse events. You won't know all the details of the rule or why the statistician chose, it but you will at least know that the statistician is the person who should prepare the response for the IRB.

Our second example involves you as a regulatory affairs associate at a medical device company that just completed an ablation trial for a new catheter. You have submitted your premarket approval application (PMA). In the statistical section of the PMA, the statistician has provided statistical analysis regarding the safety and efficacy of your catheter in comparison to other marketed catheters. A reviewer at the FDA sent you a letter asking why Peto's method was not used instead of Greenwood's approximation. You do not know what these two methods are or how they apply.

From this course, you will learn about survival analysis. In studying the effectiveness of an ablation procedure, we not only want to know that the procedure stopped the arrhythmia (possibly atrial fibrillation), but also that the arrhythmia does not recur. Time to recurrence is one measure of efficacy for the treatment. Based on the recurrence data from the trial, your statistician constructs a time-to-event curve called the Kaplan–Meier curve.

If we are interested in the probability of recurrence within 1 year, then the Peto and Greenwood methods are two ways to get approximate confidence intervals for it. Statistical research has shown differences in the properties of these two methods for obtaining approximate confidence intervals for survival probabilities. As an example, Greenwood's estimate of the lower confidence bound can be too high in situations where the number of subjects still at risk at the time point of interest is small.

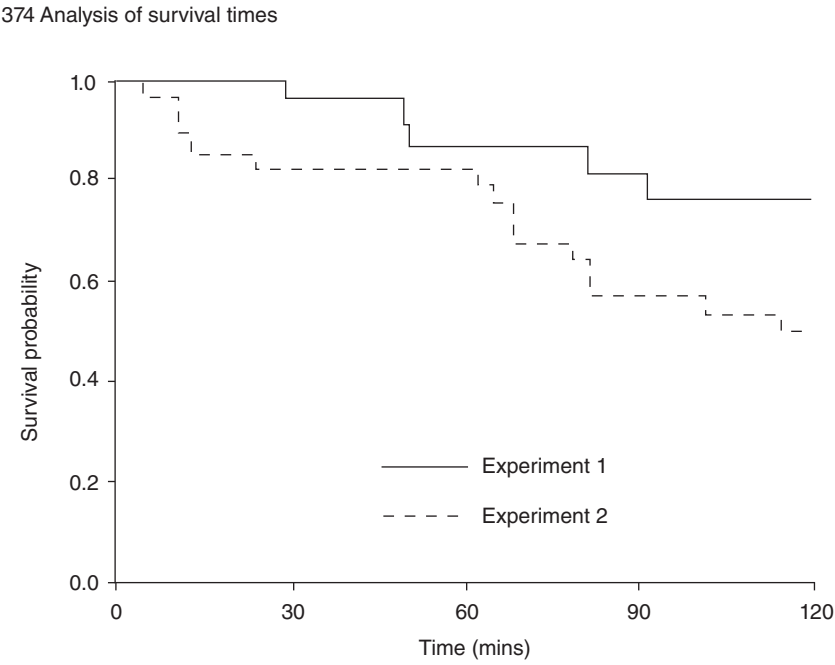


Figure 1.1. Example of a Kaplan–Meier curve. Taken from Altman (1991), *Practical Statistics for Medical Research*. Chapman and Hall/CRC, p. 374.

In these situations, Peto's method gives a better estimate of this lower bound. In general, neither method is always superior to the other. Since the FDA posed this question, the statistician would opt to provide the Peto estimate in addition to Greenwood for the FDA to compare the two lower confidence bounds. Knowing these simple facts would help you deal with the FDA question quickly, effectively, and accurately (Fig. 1.1).

In situation 3, you are in regulatory affairs and are reviewing an FDA letter about a PMA submission. The FDA wants you to report results in terms of confidence intervals, in addition to the *p*-values, before they give final approval to the treatment. You recognize this as a statistical question, but are worried because if it takes significant time to supply the request, the launch date of the new device will be delayed and will upset marketing's plans. You don't even know what a confidence interval is!

In this case, since you have the necessary data to do the binomial test on success probability, you can easily compute an exact confidence

interval. Your statistician can provide this for you in less than 1 day and you are greatly relieved.

In situation 4, you are a clinical research associate in the middle of an important phase III trial. Based upon a data analysis done by the statistics group and an agreement with the FDA prior to the trial, the primary endpoint can be changed from a condition at the 6-month follow-up visit to that same condition at the 3-month follow-up visit. This is great news, because it means that the trial can be finished sooner!

There is a problem though. The protocol only required follow-up visits at 2 weeks and 6 months, and the 3-month follow-up was optional. Unfortunately, some sites opted not to conduct the 3-month follow-up. Your clinical manager now wants you to have all the patients that are past the 3-month time point since the procedure was done and did not have the 3-month follow-up to come in for an unscheduled visit. When you requested that the investigators do this, a nurse and one investigator balked at the idea and demanded to know why this is necessary. You need an answer from your statistician!

To placate the investigator, the statistician tells the investigator that they could not use the 3-month follow-up initially because the FDA had not seen data to indicate that a 3-month follow-up would be enough to determine long-term survival. However, during the early part of the trial, the statistician was able to find relevant survival curves to indicate the survival probability flattens out at 3 months' duration. This was enough to convince the FDA that the 3-month endpoint was sufficient to determine long-term survival. If we now have the unscheduled visits, these could be the subjects' last visit, and many subjects will not need a 6-month follow-up, allowing a shorter accrual time and a chance to get the product to market faster.

This explanation helped, but the problem could have been avoided had the clinician had the foresight to see the importance of making the 3-month follow-up mandatory in the protocol. The investigator was pleased because although it would cost more to add these unscheduled visits, this would be more than compensated by the dropping of the 6-month follow-up, for those getting the unscheduled visit, and possibly some others.

In the last situation (situation 5), imagine you are the VP of the Clinical and Regulatory Affairs Departments at a medical device company. Your company hired a contract research organization (CRO) to run a blinded randomized control phase III clinical trial. You have a

statistics group, but the CRO is tasked to handle the data collection, processing, and analysis, so as to keep your company blinded and thus maintain greater integrity for the data and to avoid any presumption of bias.

The CRO can view the data in an unblinded fashion as they prepare their report. You are very curious to see the results, since a successful trial outcome is of paramount importance. Now, as the report is complete, you are the only representative of the company who can see the report. As you look at the report, you see *p*-values for statistical tests. You recall only a little statistics but remember to look for *p*-values below 0.05 because those were indicative of statistical significance. You are alarmed, when looking at a demographic comparison of treatment and control groups by age and gender, to see high *p*-values. One *p*-value was 0.56. You would like to show this to your statistician, but cannot, because he must remain blinded.

If you had taken a course like this one, you would know that for efficacy variables, the hypotheses are set up to be rejected, and low *p*-values are good. But we want the demographic factors to be nearly the same for both groups. For demographics, we do not want to reject the null hypothesis, and a high *p*-value is actually good news!!

The main reason for similarity between the groups with respect to all these demographic factors is randomization. Fisher originally suggested randomization in experiments because of confounding of effects. Perhaps unknown to the investigators, the treatment is more effective in women than men. Suppose we have 100 patients in each group. In the control group, 30 are women and 70 are men. In the treatment group, 80 are women and 20 are men, and we see a statistically significant effect. Is it due to the treatment or the fact that so many more women are in the treatment group than in the control group? Unfortunately, we do not know! This is what is called confounding.

Randomization overcomes this problem because it tends to balance out factors that we are not interested in. Simple random sampling will proportion the men and women nearly in the proportions that they occur in the patient population. This too avoids bias and confounding. In situation 5, the high *p*-value shows that the randomization is doing its job!

We now summarize what we have learned in this section.

1. Statistics and statisticians played an important role in research. Their role in medical research and particularly randomized

controlled clinical trials continues to increase rapidly, as the demand for finding new and better treatments for severe diseases increases.

2. The regulatory agencies and pharmaceutical companies continue to emphasize controlled clinical trials for the evaluation of efficacy and safety for a new drug.
3. Physicians and nurses cannot ignore statistics. It is everywhere, and is mandated by the FDA to provide proof of safety and efficacy of new drugs, devices, and combination therapies.

1.3 THE MEDICAL LITERATURE

Chapter 6 of Doug Altman's book, Altman (1991), discusses statistical methods in the medical literature. He quotes the famous statistician, Sir David Cox, who in 1983 said: "One does feel that statistical techniques both of design and analysis are sometimes adopted as rituals designed to assuage the last holders of absolute power (editors of journals) and perhaps also regulatory agencies, and not because the techniques are appreciated to be scientifically important." I agree with this statement not only as it applied in 1983, but even to a large extent, still today, 27 years later!

Altman uses very strong language regarding problems with the medical literature. He claims "Examples of substandard design and incorrect analysis can be seen in almost any issue of any medical journal." He goes on to say: "The importance of sound design and analysis cannot be overemphasized. Clearly the conclusions from a study must rely on the methods having been correct. If conclusions are unreliable because of faulty methodology, then the study cannot be clinically worthwhile. Worse, it may be clinically harmful by reason of the conclusions being misleading, and a clinically harmful study is surely unethical."

Evidence of the growth of the use of statistical methods in medical research is given in this table about the journal *Pediatrics*, taken from table 16.1, page 479 of Altman's book. The number of papers is on an increasing trend, the percentage of papers without statistics is decreasing, and the percentage with more sophisticated techniques is increasing over the three decades (Table 1.1).

Table 1.1
Use of Statistical Procedures in the Journal *Pediatrics*

	Year			
	1952	1962	1972	1982
No. of papers	67	98	115	151
% with no statistical procedures	66%	59%	45%	30%
% with procedures other than *t*, chi-square, or *r*	3%	5%	12%	35%

From Altman (1991) with permission.

Table 1.2
Errors Found in *Arthritis and Rheumatism* 1967–1968 Compared With 1982 (Continued)

Year of publication	1967–1968	1982
Number of papers	$n = 47$	$n = 74$
Error type		
Undefined method	14 (30%)	7 (9%)
Inadequate description of measure of location or dispersion	6 (13%)	7 (9%)
Repeated observations treated as independent	1 (2%)	4 (5%)
Two groups compared on more than 10 variables at 5% level	3 (6%)	4 (5%)
Multiple *t*-tests instead of ANOVA	2 (4%)	18 (24%)
Chi-squared tests used when observed frequencies are too small	3 (6%)	4 (5%)
At least one of these errors in the paper	28 (60%)	49 (66%)

From Altman (1991), with permission.

From 1982 to 2010, this trend has continued, and fortunately, the quality of the statistical refereeing has improved as well. Altman also looked at errors in a particular journal, *Arthritis and Rheumatism*, comparing the late 1960s to 1982 (Table 1.2).

We see from the tables that the medical literature was notorious for incorrect use of statistical methods. Trends from the late 1960s to the

early 1980s show an increase in the use of statistical methods and particularly the more sophisticated ones. The frequency of occurrence of elementary-type errors declined over this period. Because statistics is used more frequently and with more sophistication, there is an increase in the percentage of papers that have at least one error, as well as an increase in the percentage of papers that contain the more recent type of errors from multiple testing and the use of multiple *t*-tests instead of the analysis of variance.

1.4 MEDICAL RESEARCH STUDIES

Medical research studies involving human subjects can be put into four categories.

1. Cross-sectional studies
2. Retrospective studies
3. Prospective studies (other than clinical trials)
4. Controlled clinical trials, including pharmacokinetic and pharmacodynamic studies

While the controlled clinical trial falls under the category of prospective studies, we choose to separate it out because of its clear importance in the evaluation of new drugs and medical devices.

1.4.1 Cross-Sectional Studies Including Surveys

Definition: A cross-sectional study is one that is taken at a given point in time.

Surveys including election polls and censuses are both examples of cross-sectional studies. These studies are conducted when only one point in time is relevant to the question at hand (e.g., censuses, public opinion polls, election polls, and marketing surveys). Here, only the current opinion matters. Not interested in looking far into the future. But often in medicine, we are interested in changes over time after a medical intervention. This goes for both efficacy variables and quality of life variables. So we do not see many cross-sectional studies in medical research except in epidemiological studies.

1.4.2 Retrospective Studies

Definition: A retrospective study is one that examines relationships based on past data.

One important example of a retrospective study is the case-control study (these could also be prospective). Such studies are intended to be similar to what prospective clinical trials are intended to do. The cases are the subjects with the outcome of interest (like a treatment group in a clinical trial). The control subjects are similar demographically or otherwise to their matched case subjects, but for which the outcome did not occur.

A particular example might be a situation where subjects who contracted a particular disease such as lung cancer are asked about their past exposure to a risk factor. The same questions are administered to control subjects who did not get lung cancer. In this case, the risk factor is cigarette consumption.

1.4.3 Prospective Studies Other Than Clinical Trials

Definition: A prospective study is one that is planned in the present and takes place in the future.

Examples include cohort studies and clinical trials. Clinical trials are particularly important to us, as we have already mentioned. So we consider them as a category of their own.

An example of a cohort study is a study that follows a group of disease-free subjects who have a certain risk factor for a disease to see if they eventually develop the disease. The subjects could be young college students, the disease could be emphysema, and the risk factor could be smoking. From cohort studies, statisticians and epidemiologists determine relative risks based on exposure levels to risk factors.

1.4.4 Controlled Clinical Trials

In the context of clinical trials, an experiment is a study that assigns subjects to treatment groups in order to assess differences among treatments. A randomized experiment is one in which randomization is used for the selection process.

Definition: An experiment performed to evaluate the effect of intervention(s) or treatment(s) for a group of human subjects against a control group that does not get a treatment(s) (placebo) or gets different treatment(s).

The purpose is to see if the difference in treatment creates differences in outcomes for the treatment group versus the control group. The gold standard for clinical trials is the double-blinded randomized controlled trial. When constructed properly, these trials provide good statistical information about the differences between two groups or several groups (often there can be more than one treatment). The control group could be on a drug that is an active competitor to the study drug or on placebo, or, more generally, a different treatment protocol, a different medical device or surgical procedure, and so on.

The use of randomization and blinding is to protect the study from biases that could invalidate the results. Not all clinical trials are *blinded, randomized, or completely prospective*. Sometimes in device trials, historical controls or objective performance criteria (OPCs) are used for comparison with the treatment. This makes the comparator retrospective while the treatment is done prospectively. Since the trial only has one arm, there is no blinding or randomization in this type of trial.

1.4.5 Conclusions

1. There are several types of studies in medical research.
2. Each study has its advantages and disadvantages.
3. Cross-sectional studies only look at one point in time.
4. Most medical research and particularly clinical trials are concerned with how patients improve or get worse over time as a function of alternative treatments.
5. Because of (4), cross-sectional studies are not common in medical research other than in some epidemiologic studies.
6. Double-blind randomized control clinical trials provide the gold standard for evaluating a new treatment versus current standard care and/or placebo when done properly. But they are also the most costly and difficult to implement studies.

1.5 EXERCISES

1. What is a Kaplan–Meier curve?

2. For what kind of data do you compute Kaplan–Meier curves?

3. Why is randomization important in clinical trials?

4. What does Greenwood's method refer to?

5. Why do we compute *p*-values? When is it good for *p*-values to be small and when is it all right if they are large?

6. What are cross-sectional studies and why are they uncommon in medical research?

7. What are retrospective studies?

8. What are prospective studies?

9. What are controlled clinical trials and why is blinding important?

CHAPTER 2

Sampling from Populations

One of the key aspects of statistics and statistical inference is to draw conclusions about a population based on a sample. Our ability to make good inferences requires an intelligent design and must include some form of random sampling. Random sampling is needed so that the sample can be analyzed based on the probability mechanism that generates the sample. This way, estimates based on the sample data can be obtained, and inference drawn based on the probability distribution associated with the sample.

To illustrate, suppose we select five students at random from a math class of 40 students. We will formally define random sampling later. If we give a math test to these students based on the material they have studied in the class, and we average the five scores, we will have a prediction of what the class average for that test will be. This prediction will be unbiased (meaning that if we repeatedly took samples of and averaged them, the average of the averages will approach the class average).

In practice, we do not repeat the process, but we do draw inference based on the properties of the sampling procedure. On the other hand, suppose we selected the five students to be the ones with the highest class average thus far in the class. In that case, we would not have a

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

random sample, and the average of this group could be expected to be higher than the class average. The amount that it is higher is the bias of the prediction. Bias is something we want to avoid because usually we cannot adjust our estimate to get a good prediction.

In addition to bias (which can be avoided by randomization), an estimate or prediction will have a variance. The variance is a measure of the variability in estimates that would be obtained by repeating the sampling process. While bias cannot be controlled by the sample size, the variance can. The larger the sample size is, the smaller is the variance of the estimate, or in the example, above the prediction of the class average.

Suppose that instead of taking a random sample of size 5, we took a random sample of size 10. Then, for an estimate known to be unbiased (e.g., the sample mean) will still be unbiased, and its variance will be lower, meaning that it will tend to be closer to the value for the entire class. You can imagine that if we chose 39 out of the 40 at random, the prediction would be extremely close to the class average, and if we had taken all 40, it will equal the class average and have zero variance.

An excellent example that illustrates the need for random sampling and the bias in prediction when the sample is not random is the *Literary Digest*'s prediction of the winner of the 1936 U.S. Presidential election. Franklin Roosevelt was the incumbent and the Democratic nominee. Alfred Landon was the Republican nominee. To predict the winner, the *Literary Digest* mailed out 10 million ballots asking registered voters which candidate they preferred. A total of 2.3 million out of the 10 million ballots were returned and on the basis of the results for the 2.3 million the *Literary Digest* predicted Landon to be a big winner.

Although the number of voters in the election would be a lot more than the actual or even the intended sample, that sample size is large enough that if it were a random sample of those who would vote, it would have a very small standard deviation (in political surveys, approximately 2 standard deviations for the estimate is called the margin of error), and the prediction would be highly reliable. The result of the election, however, was that Roosevelt won by a landslide, obtaining 62% of the popular vote. This high visibility poll totally destroyed the credibility of the *Literary Digest*, and soon caused it to cease publication. How could they have gone so wrong?

A subsequent analysis of their sampling method indicated that the original mailing list of 10 million was based primarily on telephone directories and motor vehicle registration lists. In modern times, such a sampling method would be acceptable, since the percentage of eligible voters that have telephones and drivers licenses is nearly 100%.

But, in 1936, the United States was recovering from the great depression, and telephones and automobiles were a luxury. So a large majority of the people with telephones and/or cars were affluent. The affluent Americans tended to be Republicans, and were much more likely to vote for Landon than the Democrats, many of whom were excluded because of this sampling mechanism. As poor and middle-income Americans represented a much larger portion of American society in 1936, and they would be more likely to vote for Roosevelt, this created a large bias that was not recognized by those individuals at the Literary Digest who were conducting the survey. This shows that samples not chosen at random may appear on the surface to be like a random sample, but could have a large enough bias to get the prediction wrong. If a truly random sample of 2.3 million registered voters likely to vote were selected and the true proportion that would vote for Roosevelt were 62%, then it would be nearly impossible for the survey to pick Landon.

2.1 DEFINITIONS OF POPULATIONS AND SAMPLES

At this stage, we have informally discussed populations and samples. Now as we get into the details of random samples and other types of sampling methods, we will be more formal. The term *population* refers to a collection of people, animals, or objects that we are interested in studying. Usually, there is some common characteristic about this population that interests us. For example, the population could be the set of all Americans having type II diabetes. A sample would be a subset of this population that is used to draw inferences about the population. In this example, we might have a drug like metformin that we think will control the sugar levels for these patients. There may be millions of Americans that have type II diabetes.

But we shall draw inference about the population based on a sample of 1000 subjects with type II diabetes that we were able to enroll in a clinical trial. If the trial is properly conducted statistically, we may estimate a treatment effect in the population based on an estimate from the sample of 1000 subjects in the trial. This estimate, if favorable, may lead the FDA to approve the drug for treatment of type II diabetes to any American with type II diabetes.

Without a proper statistical design and analysis, the inference to the population would not be valid and would not lead to an approval even if the results are positive for the sample. The sample estimate could be biased, and the probability that a decision favors the conclusion of effectiveness when the drug is really not effective (called the type I error or significance level) would not be appropriately controlled.

So to summarize, a population is a collection of things or people that have similarities and possibly subgroup differences that you are interested in learning about. A sample is simply a subset of the population that you take measurements on to draw inferences about those measurements for the population the sample was taken from.

2.2 SIMPLE RANDOM SAMPLING

One of the easiest and most convenient ways to take a sample that allows statistical inference is by taking a simple random sample. As mentioned earlier, many methods of sampling can create biases. Simple random sampling assures us that sample estimates like the arithmetic mean are unbiased.

Simple random sampling involves selecting a sample of size n from a population of size N. The number of possible ways to draw a sample of size n out of a population of size N is the binomial coefficient C_n^N (read as "combinations of N choose n"), the number of combinations of N things taken n at a time. This is known in combinatorial mathematics to be $N!/[n!(N-n)!]$. By "$n!$", we mean the product $n(n-1)(n-2)\ldots 3\ 2\ 1$. In simple random sampling, we make the selection probability the same for each possible choice for the sample n. So the probability that any particular set occurs is $1/C_n^N$. In Section 2.3, we will show a method for taking simple random samples based on

using a pseudo-random number generator on the index set for the population.

2.3 SELECTING SIMPLE RANDOM SAMPLES

Simple random sampling can alternatively be defined as sampling at random without replacement from the population. Going by our original definition, a brute force way to generate a random sample would be to enumerate and order all the possible samples from 1 to C_n^N and randomly select an integer k, where $1 \le k \le C_n^N$.

To illustrate this method, we will look at a simple example where $N = 6$ and $n = 4$. Then the number of possible samples is $C_4^6 = 6!/[4!\,2!] = 6 \times 5/2 = 15$. Suppose these six elements represent patients, and we denote them as the set {A, B, C, D, E, F}. Using this notation, we can enumerate the 15 distinct samples any way we want and assign integer indices from 1 to 15. A systematic enumeration might look as follows:

1. {A, B, C, D}
2. {A, B, C, E}
3. {A, B, C, F}
4. {A, B, D, E}
5. {A, B, D, F}
6. {A, B, E, F}
7. {A, C, D, E}
8. {A, C, D, F}
9. {A, C, E, F}
10. {A, D, E, F}
11. {B, C, D, E}
12. {B, C, D, F}
13. {B, C, E, F}
14. {B, D, E, F}
15. {C, D, E, F}

We then use a table of uniform random numbers, or on the computer, generate a uniform pseudorandom number. A computer pseudorandom number generator is an algorithm that will generate a sequence of numbers between 0 and 1 that have properties approximating those of a sequence of independent uniform random numbers. To assign a random index to the random number we generate, we do the following: We first break up the interval [0, 1)* into 15 disjointed (i.e., nonoverlapping) intervals of equal length 1/15. So the intervals are [0, 1/15), [1/15, 2/15) [2/15, 3/15), . . . , [14/15, 1). Let U denote the random number selected by the table or the computer generated value. Then

If $0 \le U < 0.0667$, then the index is 1 (0.0667 is a decimal approximation to 1/15).

If $0.0667 \le U < 0.1333$, then the index is 2.

If $0.1333 \le U < 0.2000$, then the index is 3.

If $0.2000 \le U < 0.2667$, then the index is 4.

If $0.2667 \le U < 0.3333$, then the index is 5.

If $0.3333 \le U < 0.4000$, then the index is 6.

If $0.4000 \le U < 0.4667$, then the index is 7.

If $0.4667 \le U < 0.5333$, then the index is 8.

If $0.5333 \le U < 0.6000$, then the index is 9.

If $0.6000 \le U < 0.6667$, then the index is 10.

If $0.6667 \le U < 0.7333$, then the index is 11

If $0.7333 \le U < 0.8000$, then the index is 12.

If $0.8000 \le U < 0.8667$, then the index is 13.

If $0.8667 \le U < 0.9333$, then the index is 14.

If $0.9333 \le U < 1.0000$, then the index is 15.

For example, suppose the computer generated the number 04017 corresponding to 0.4017. Since $0.4000 \le 0.4017 < 0.4667$, the index is

* "[0, 1)" means all values x such that is greater than or equal to 0 but less than 1, "(0, 1)" means all x greater than 0 but less than 1, "(0, 1]" means all x greater then 0 bur less than or equal to 1, and "[0,1]" means all x greater than or equal to 0 but less than or equal to 1.

7. Then referring to the systematic list, we see that the index 7 corresponds to the sample {A, C, D, E}.

Now this method is feasible when N and n are small like 6 and 4 above, since the number of combinations is only 15. But as N and n get larger, the number of combinations gets out of hand very quickly. So a simpler alternative is to consider the sampling without replacement approach. In this approach, the individual patients get ordered. One ordering that we could have is as follows:

1 is A
2 is B
3 is C
4 is D
5 is E
6 is F

Now we divide [0, 1) into six equal intervals and assign the uniform random number as follows:

If $0.0000 \leq U < 0.1667$, then the index is 1.
If $0.1667 \leq U < 0.3333$, then the index is 2.
If $0.3333 \leq U < 0.5000$, then the index is 3.
If $0.5000 \leq U < 0.6667$, then the index is 4.
If $0.6667 \leq U < 0.8333$, then the index is 5.
If $0.8333 \leq U < 1.0000$, then the index is 6.

Example: From a table of uniform random numbers, suppose the first number to be 00439 for 0.00439, since 0.00439 is in the interval [0, 0.1667], we choose index 1 corresponding to patient A. Now A is taken out so we rearrange the indexing.

1 is B
2 is C
3 is D

4 is E

5 is F

Now we must divide [0, 1) into five equal parts.
So we get:

If $0.0000 \le U < 0.2000$, then the index is 1.

If $0.2000 \le U < 0.4000$, then the index is 2.

If $0.4000 \le U < 0.6000$, then the index is 3.

If $0.6000 \le U < 0.8000$, then the index is 4.

If $0.8000 \le U < 1.0000$, then the index is 5.

The second uniform random number from the table is 29676, corresponding to 0.29676. Now since $0.2000 \le U < 0.4000$, the index is 2 corresponding to C. So now our sample includes A and C. Again, in order to sample without replacement from the remaining four patients B, D, E, and F, we divide [0, 1) into four equal parts and redefine the indices as

1 is B

2 is D

3 is E

4 is F

For the intervals, we get:

If $0.0000 \le U < 0.2500$, then the index is 1.

If $0.2500 \le U < 0.5000$, then the index is 2.

If $0.5000 \le U < 0.7500$, then the index is 3.

If $0.7500 \le U < 1.0000$, then the index is 4.

The third uniform random number in the table is 69386. So $U = 0.69386$. We see that $0.5000 \le U < 0.7500$. So the index is 3, and we choose patient E. Now we have three of the four required patients in our sample. They are A, C, and E. So for the final patient in the sample,

we pick at random between B, D, and F. The indices are chosen as follows:

1 is B

2 is D

3 is F

To divide [0, 1) into three equal parts we get:

If $0.0000 \leq U < 0.3333$ then the index is 1.

If $0.3333 \leq U < 0.6667$ then the index is 2.

If $0.6667 \leq U < 1.0000$, then the index is 3.

The final random number from the table is 68381. So $U = 0.68381$.

We see that $0.6667 \leq U < 1.0000$. So the index for the last patient is 3, corresponding to patient F. The random sample of size 4 that we chose is {A, C, E, F}. This approach seems a little more awkward, but it does generate a simple random sample using only four random numbers. Although it is awkward, it avoids enumerating all 15 combinations and therefore remains a feasible approach as N and n get large.

A simpler approach that also generates a simple random sample is the rejection method. In the rejection method, we do not repartition the interval [0, 1) after choosing each patient. We stay with the original partition. This saves some calculations, but could lead to a longer string of numbers. We simply start with the approach that we previously used in sampling without replacement, but since we do not change the partition or assignment of indices, it is now possible to repeat an index (for bootstrap sampling, this will be perfectly fine). But since a simple random sample cannot repeat an element (a patient in our hypothetical example), we cannot include a repeat. So whenever a patient repeats, we reject the duplicate sample and pick another random sample. This continues until we have a complete sample of size n ($n = 4$ in our example).

Using the same table and running down the first column, the sequence of numbers is 00439, 29676, 69386, 68381, 69158, 00858, and 86972. In the previous examples, we ran across the first row and then the second. In this case, we get a different sequence by going down

the first column. We did this to illustrate repeat values and how they are handled in generating the sample.

Recall that when we subdivide the interval into six equal parts, we get:

If $0.0000 \leq U < 0.1667$ then the index is 1.

If $0.1667 \leq U < 0.3333$ then the index is 2.

If $0.3333 \leq U < 0.5000$ then the index is 3.

If $0.5000 \leq U < 0.6667$ then the index is 4.

If $0.6667 \leq U < 0.8333$ then the index is 5.

If $0.8333 \leq U < 1.0000$ then the index is 6.

Also recall the correspondence of patients to indices:

1 is A

2 is B

3 is C

4 is D

5 is E

6 is F

So the random sequence generates A, B, E, E, E, A, F. Since we didn't get a repeat among the first three patients, A, B, and E are accepted. But the fourth random number repeats E, so we reject it and take the fifth random number. The fifth number repeats E again so we reject it and look at the sixth random number in the sequence. The sixth random number chooses A, which is also a repeated patient, so we reject it and go to the seventh random number. This number leads to the choice of F, which is not a repeat so we accept it. We now have four different patients in our sample so we stop.

The rejection method can also be shown mathematically to generate a simple random sample. So we had the advantage of only doing one partitioning, but with it came the repeats and the need to sometimes have to generate more than four random numbers. In theory, we could get many repeats, but a long series of repeats is not likely. In this case, we needed seven random numbers instead of just four. The rejection

method is preferred on the computer because generating new random numbers is faster than calculating new partitions. So although it looks to be wasteful in practice, it is usually computationally faster.

Now for each patient, we are interested in a particular characteristic that we can measure. For this example, we choose age in years at their last birthday. Let us assume the ages for the patients are as follows:

A is 26

B is 17

C is 45

D is 70

E is 32

F is 9

The parameter of interest is the average age of the population. We will estimate it using the sample estimate. Since the population is 6, and we know the six values, the population mean, denoted as μ, is (26 + 17 + 45 + 70 + 32 + 9)/6 = 33.1667. In our example, we will not know the parameter value because we will only see the sample of size 4 and will not know the ages of the two patients that were not selected. Now, if we generated the random sample using the exact random numbers that we got from the reject technique, we would have {B, C, E, F} as our sample, and the sample mean will be (17 + 45 + 32 + 9)/4 = 19.5. This is our estimate. It is a lot smaller than the true population mean of 33.1667.

This is because patient D is not in the sample. D is the oldest patient and is 70. So his addition in the average would increase the mean and his absence decreases it. So if we added D to the sample, the average would be (17 + 45 + 70 + 32 + 9)/5 = 34.6. So adding D to the sample increases the mean from 19.5 to 34.6. On the other hand, if we think of the sample as being {B, C, D, E, F}, the removal of D drops the mean from 34.6 to 19.5. So the influence of D is 15.1 years! This is how much D influences the mean. This shows that the mean is a parameter that is heavily influenced by outliers. We will address this again later.

The sample mean as an estimator is unbiased. That means that if we averaged the estimate for the 15 possible samples of size 4, we

would get exactly the population mean, which is 31.1667. This result can be proven mathematically. In this case, since we know what the finite population is, we can calculate the 15 possible sample means and take their average to verify that it equals the population mean. Recall the 15 possible samples are:

{A, B, C, D} with sample mean = (26 + 17 + 45 + 70)/4 = 39.50
{A, B, C, E} with sample mean = (26 + 17 + 45 + 32)/4 = 30.00
{A, B, C, F} with sample mean = (26 + 17 + 45 + 9)/4 = 24.25
{A, B, D, E} with sample mean = (26 + 17 + 70 + 32)/4 = 36.25
{A, B, D, F} with sample mean = (26 + 17 + 70 + 9)/4 = 30.50
{A, B, E, F} with sample mean = (26 + 17 + 32 + 9)/4 = 21.00
{A, C, D, E} with sample mean = (26 + 45 + 70 + 32)/4 = 43.25
{A, C, D, F} with sample mean = (26 + 45 + 70 + 9)/4 = 43.25
{A, C, E, F} with sample mean = (26 + 45 + 32 + 9)/4 = 28.00
{A, D, E, F} with sample mean = (26 + 70 + 32 + 9)/4 = 34.25
{B, C, D, E} with sample mean = (17 + 45 + 70 + 32)/4 = 41.00
{B, C, D, F} with sample mean = (17 + 45 + 70 + 9)/4 = 35.25
{B, C, E, F} with sample mean = (17 + 45 + 32 + 9)/4 = 25.75
{B, D, E, F} with sample mean = (17 + 70 + 32 + 9)/4 = 32.00
{C, D, E, F} with sample mean = (45 + 70 + 32 + 9)/4 = 39.00

In this case, the largest mean is 43.25, and the smallest is 21.00, and the value closest to the population mean is 34.25. This shows that the estimate has a lot of variability. To verify the property of unbiasedness, we need to average these 15 estimates and verify that the average is 33.1667. This goes as follows:

The expected value of the averages is (39.5 + 30.0 + 24.25 + 36.25 + 30.5 + 21.0 + 43.25 + 37.5 + 28.0 + 34.25 + 41.0 + 35.25 + 25.75 + 32.0 + 39.0)/15 = 497.50/15 = 33.1667 rounded to four decimal places.

In Section 2.5, we will generate bootstrap samples. In bootstrapping, we do simple random sampling with replacement. So this can be

accomplished easily by using the approach of the rejection method without the need to reject, since repeats are acceptable.

The basic bootstrap idea is to let the original sample data serve as the population, and then you sample with replacement from the sample data. The original sample is size n, and bootstrapping is usually done by taking m samples with replacement with $m = n$.

However, in recent years, it has been discovered that although $m = n$ usually works best, there are situations where the choice of $m = n$ leads to a particular type of incorrect solution which statisticians call inconsistency. In several of these cases, a consistent bootstrap approach can be obtained by making $m << n$.* This is called the m-out-of-n bootstrap. For the statistical theory of consistency to hold, both n and m tend to infinity, but with m going at a slower rate. So we will see that bootstrap sampling is conceptually very similar to simple random sampling. The only difference is that replacement is used for the bootstrap. For the bootstrap estimates, we will also look at the ages of the six patients.

2.4 OTHER SAMPLING METHODS

Stratified random sampling is just a little more complicated; the simple random sampling in a set of m strata are defined indexed by k where $k = 1, 2, \ldots, m$, and each strata gets a simple random sample of size n_k. An example of stratification might be age group, with $k = 1$ for ages 1–12, $k = 2$ for ages 13–20, $k = 3$ for ages 21–35, $k = 4$ for ages 36–55, $k = 5$ for ages 56–75, and $k = 6$ for anyone over 75. A stratified random sample can work better than a simple random sample if each stratum has a relatively homogeneous group, but there are marked differences between strata.

Other forms of sampling are convenience sampling, cluster sampling, and systematic sampling. Cluster sampling is a random sampling approach that is used when it is easier to randomly select a group of elements for a sample rather than the individual elements themselves. Examples could be lists of districts or counties within a state. In cluster sampling, the item being sampled is a cluster. A cluster is a group of objects generally found in the same location. For example, in sampling

* By "$m << n$," we mean that m is much less than n.

households in a particular city, the city could be divided up into blocks. A subset of the city blocks is selected at random, and each household on the block is included in the sample. Cluster sampling is a convenient and economic way for organizations such as the U. S. Census Bureau to conduct surveys. So, for example, we may look at residents of Manhattan, New York as the population. Every city block in Manhattan is eligible for selection, and a random sample of city blocks is taken, and every household on the chosen blocks are included.

Convenience sampling and systematic sampling are both nonrandom methods and are not recommended in general. In special cases, these methods may work, but often they don't. Systematic sampling can be used (but not necessarily recommended) when an ordered list of the population members is available. Samples are chosen by a systematic algorithm. For example, if the population size $n = 500$, and we want a sample of size 100, we can choose every fifth case on the list, such as those with indices 1, 6, 11, 16, 21, 26, 31, . . . , 491, and 496. This is not the only way if we skip 1; we can accomplish the sample choosing 2, 7, 12, 17 . . . , 492, and 497. We could also start with the third, fourth, or fifth index in the sequence. If we start with 5, the sequence is 5, 10, 15, 20, 25, . . . , 495, and 500.

Systematic sampling can work if the ordering has no relationship to the value of the outcome variable. A case where systematic sampling can fail is when the outcomes are cyclical in time. For instance, if the pattern is sinusoidal and the period is 5 units, then we could be sampling at the peaks of the cycle when we pick every fifth case in sequence and the first case is a peak. This would lead to a positive bias in the estimate for the outcome variable's mean. On the other hand, starting at a trough would create a negative bias on the estimate of the outcome variable's mean.

Convenience sampling only means that you find a sample of size n out of the population of size N in a simple and convenient way. There is no way to draw inference from such a sample. Convenience sampling should never be recommended.

2.5 GENERATING BOOTSTRAP SAMPLES

Bootstrap sampling is simple random sampling from the observed data (also called the empirical distribution). It amounts to sampling with

replacement m times from the data where each data point has probability $1/n$ for each of the m draws, where n is the number of data points. As mentioned in Section 2.3, m is usually equal to n, but sometimes it is advantageous to take $m << n$.

In this section, we will show how bootstrap samples can be generated, much as we did for simple random samples in Section 2.3. We will further discuss the bootstrap when we get to hypothesis testing and confidence intervals, where it is commonly applied. Without going into detail now, let us say that bootstrap estimation is based on using the sampling distribution of estimates obtained from bootstrap samples.

In theory, that sampling distribution can be derived directly from the data. However, this is not often easy to do (especially as n gets large), so the distribution is approximated by Monte Carlo methods. That means that we get a collection of B bootstrap samples by sampling with replacement from the original data B times, each time taking a sample of size m. In our example, we will take $m = n$, where n is the size of the original sample.

The bootstrap samples, typically, differ from the original sample because some observations get repeated in the bootstrap sample and others are left out. This will become apparent in the example. To generate a bootstrap sample, we again partition the interval [0, 1). In this case, since we have n samples indexed 1, 2, 3, . . . , n, we divide the interval into n equal disjoint parts. Again taking U to be a uniform random number from a table of random numbers, we get:

If $0 \leq U < 1/n$, the index is 1.

If $1/n \leq U < 2/n$, the index is 2.

If $2/n \leq U < 3/n$, the index is 3.

.

.

.

If $(n-2)/n \leq U < (n-1)/n$, the index is $n-1$.

If $(n-1)/n \leq U < n/n = 1$, the index is n

Let us take the same population of six patients {A, B, C, D, E, F} that we used in Section 2.3, but it now represents the sample of patients. Again, the correspondence of patients to indices:

1 is A
2 is B
3 is C
4 is D
5 is E
6 is F

The variable of interest is the patient's age, so again:

A is 26
B is 17
C is 45
D is 70
E is 32
F is 9

A bootstrap sample will sample six times with replacement from the six patients, and mean age will be computed for each bootstrap sample. There are $6^6 = 46{,}656$ possible bootstrap samples when order is counted. This is a little too much for a human to handle, but not so large to cause difficulty for today's computers. To get the bootstrap distribution for the mean, we would enumerate all 46,656 possible bootstrap samples get the age distribution for each of these bootstrap samples. For each bootstrap sample we compute, its mean and the set of all 46,656 means provides the bootstrap sampling distribution for the mean. This is very tedious and unnecessary.

We can get a good approximation of the distribution from just 100 to 1000 randomly selected bootstrap samples. The number of randomly selected bootstrap samples is often denoted as B. That approach is what we call the Monte Carlo approximation to the bootstrap distribution. For illustrative purposes, we will take B = 10 even though in practice the number B needs to be much larger to get a good approximation to the bootstrap distribution. The random numbers and the corresponding patients and ages for the ten bootstrap samples are as follows:

Bootstrap sample 1: 69386, 71708, 88608, 67251, and 00169 corresponding to patients E, E, F, E, B, and A and ages 32, 32, 9, 32, 17, 26, with bootstrap mean estimate 24.67.

Bootstrap sample 2: 68381, 61725, 49122, 75836, 15368, and 52551 corresponding to patients E, D, C, E, A, and D, and ages 32, 70,45, 32, 26, and 70, with bootstrap mean estimate 45.83.

Bootstrap sample 3: 69158, 38683, 41374, 17028, 09304, and 10834 corresponding to patients E, C, C, B, A, and A, and ages 32, 45, 45, 17, 26, and 26, with bootstrap mean estimate 31.83.

Bootstrap sample 4: 00858, 04352, 17833, 41105, 46569, and 90109 corresponding to patients A, A, B, C, C, and F, and ages 26, 26, 17, 45, 45, and 9, with bootstrap mean estimate 28.00.

Bootstrap sample 5: 86972, 51707, 58242, 16035, 94887, and 83510 corresponding to patients F, D, D, A, F, and F, and ages 9, 70, 70, 26, 9, and 9, with bootstrap mean estimate 32.17.

Bootstrap sample 6: 30606, 45225, 30161, 07973, 03034, and 82983 corresponding to patients B, C, B, A, A, and E, and ages 17, 45, 17, 26, 26, and 32, with bootstrap mean estimate 27.17.

Bootstrap sample 7: 93864, 49044, 57169, 43125, 11703, and 87009 corresponding to patients F, C, D, C, A, and F, and ages 9, 45, 70, 45, 26 and 9, with bootstrap mean estimate 34.0.

Bootstrap sample 8: 61937, 90217, 56708, 35351, 60820, and 90729 corresponding to patients D, F, D, C, D, and F, and ages 70, 9, 70, 45, 70 and 9, with bootstrap mean estimate 45.5.

Bootstrap sample 9: 94551, 69538, 52924, 08530, 79302, and 34981 corresponding to patients F, E, D, A, D, and C, and ages 9, 32, 70, 26, 70 and 45, with bootstrap mean estimate 42.0.

Bootstrap sample 10: 68381, 61725, 49122, 75836, 15368, and 52551 corresponding to patients E, C, C, D, E, and B, and ages 32, 45, 45, 70, and 17, with bootstrap mean estimate 33.83.

The mean of the bootstrap distribution is (24.67 + 45.83 + 31.83 + 28.0 + 32.17 + 27.17 + 34.0 + 45.5 + 42.0 + 34.83)/10 = 31.88. The bootstrap mean will converge to the true mean for the six patients as the number of bootstrap samples B gets large. As we already

suggested, 10 is not a large number, and you can see that since the original sample mean is 33.17, the estimate is off by 1.29 years. A value of B = 100 or 500 should make the estimate much closer.

Properties of the bootstrap samples to note are the repetitions. In bootstrap sample 1, E occurs three times and C and D are both left out. In bootstrap sample 2, E and D each repeat once, and B and F are left out. Bootstrap sample 9 has only one repetition, and only B is left out. I that sense it is closest to the original sample, but its mean is 42.0 compared with the mean of 33.17 for the original sample. The large difference is due to the fact that the oldest patient E is repeated and the second youngest is the one left out.

2.6 EXERCISES

1. Why do we need to collect samples when we want to determine population characteristics?

2. Provide a definition in your own words for the following terms:

(a) Sample

(b) Census

(c) Parameter

(d) Statistic

3. Describe and contrast the following types of sampling designs. Also, state when if ever it is appropriate to use the particular designs.

(a) Simple random sample

(b) Stratified random sample

(c) Convenience sample

(d) Systematic sample

(e) Cluster sample

(f) Bootstrap sample

4. What is meant by parameter estimation?

5. For sample designs (a), (b), (c), and (d) in exercise 3, explain under what circumstances bias can enter?

6. How does bootstrap sampling differ from simple random sampling?

7. What is the rejection sampling method and when is it used?

8. Why would a convenience sample of the elderly on vacation in Hawaii probably not be representative of the elderly in retirement homes?

9. What role does the sample size play in the accuracy of a statistical inference?

10. Why is the choice of the design for the sample more critical than the size of the sample?

11. Why is bias more important than variance in research?

CHAPTER 3

Graphics and Summary Statistics

3.1 CONTINUOUS AND DISCRETE DATA

Numerical or quantitative data can be continuous or discrete. Discrete data are data that consist of a finite or a countably infinite (mathematically equivalent to the integers) set of numbers. The binomial distribution that counts the number of successes is a discrete distribution with a finite number of outcomes 0 to n successes out of n. In contrast, the Poisson distribution counts the number of events occurring in a unit time interval. It can take on any integer value that is nonnegative. So it has a countably infinite set of values for the probability distribution. A property of discrete data is that between any two values, there are real numbers that are not possible data points.

On the other hand, continuous data have the property that there exist two real numbers that are possible values, and any real number between those numbers is a possible data point. Data that are continuous include such things as weight, volume, area, and density. Although height and weight are considered continuous, they are usually measured on a discrete scale, such as inches and pounds respectively.

We call these data continuous because although we can only measure height to the nearest inch, say, in theory, a person could have a height between two units of measurement. Practically speaking, if the

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

units are fine and the possible values are large, it makes more sense to treat the data as though it were continuous even though technically it may be discrete.

3.2 CATEGORICAL DATA

Categorical data is data that is not numerical. Often there is no natural order to categorical data, although there may be a qualitative ordering, such as degree of severity. However, if we use a scale such as a Likert scale to order the categories, they do not have the typical numerical meaning. For example, a 2 on the scale may not be twice as severe as 1. So ratios of the scaled data have no real meaning. Categorical data can be dichotomous, such as true or false, male or female, yes or no, alive or dead. It also can consist of three or more categories. So race, religion, ethnicity, and education level are all examples of categorical data with more than two categories. Among these four examples, only education level has a natural ordering in terms of the hierarchy of grade levels: graduate school > college > high school > elementary school, for example.

3.3 FREQUENCY HISTOGRAMS

Fo r continuous data, frequency histograms offer us a nice visual summary of the data and the shape of its distribution. The range of possible values for the data is divided into disjoint intervals usually of equal length, and the number of data points in each interval is shown as a bar. The art of generating frequency histograms is in the decision as to how many intervals to choose. If you choose too many intervals, some intervals could be sparse or empty, and the bars could look spikey. If you take too few intervals, the bars may flatten and you lose some of the shape of the distribution.

We shall produce a histogram for a set of body mass index (BMI) measurements for 120 U.S. adults. The data looks as follows in Table 3.1.

To better discern patterns in this data, it is convenient to order the data from lowest in the top left corner to highest in the bottom right corner in ascending order down the columns, for example (could alternatively have chosen to go ascending across the rows). The result is show in Table 3.2.

Table 3.1
Body Mass Index: Sample of 120 U.S. Adults

27.4	31.0	34.2	28.9	25.7	37.1	24.8	34.9	27.5	25.9
23.5	30.9	27.4	25.9	22.3	21.3	37.8	28.8	28.8	23.4
21.9	30.2	24.7	36.6	25.4	21.3	22.9	24.2	27.1	23.1
28.6	27.3	22.7	22.7	27.3	23.1	22.3	32.6	29.5	38.8
21.9	24.3	26.5	30.1	27.4	24.5	22.8	24.3	30.9	28.7
22.4	35.9	30.0	26.2	27.4	24.1	19.8	26.9	23.3	28.4
20.8	26.5	28.2	18.3	30.8	27.6	21.5	33.6	24.8	28.3
25.0	35.8	25.4	27.3	23.0	25.7	22.3	35.5	29.8	27.4
31.3	24.0	25.8	21.1	21.1	29.3	24.0	22.5	32.8	38.2
7.3	19.2	26.6	30.3	31.6	25.4	34.8	24.7	25.6	28.3
26.5	28.3	35.0	20.2	37.5	25.8	27.5	28.8	31.1	28.7
24.1	24.0	20.7	24.6	21.1	21.9	30.8	24.6	33.2	31.6

Table 3.2
Body Mass Index Data: Sample of 120 U. S. Adults (Ascending Order Going Down the Columns)

18.3	21.9	23.0	24.3	25.4	26.6	27.5	28.8	30.9	34.8
19.2	21.9	23.1	24.3	25.6	26.9	27.5	28.8	30.9	34.9
19.8	21.9	23.1	24.5	25.7	27.1	27.6	28.9	31.0	35.0
20.2	22.3	23.3	24.6	25.7	27.3	28.2	29.3	31.1	35.5
20.7	22.3	23.4	24.6	25.8	27.3	28.3	29.5	31.3	35.8
22.4	22.3	23.5	24.7	25.8	27.3	28.3	29.8	31.6	35.9
21.1	22.4	24.0	24.7	25.9	27.3	28.3	30.0	31.6	36.6
21.1	22.5	24.0	24.8	25.9	27.4	28.4	30.1	32.6	37.1
21.1	22.7	24.0	24.8	26.2	27.4	28.6	30.2	32.8	37.5
21.3	22.7	24.1	25.0	26.5	27.4	28.7	30.3	33.2	37.8
21.3	22.8	24.1	25.4	26.5	27.4	28.7	30.8	33.6	38.2
21.5	22.9	24.2	25.4	26.5	27.4	28.8	30.8	34.2	38.8

From this table, it is now easy to see at a glance that 18.3 is the lowest BMI value and 38.8 is the highest. It is also easy to see the values that repeat by scanning down the columns, and we quickly see that 27.4 occurs the most times (five) and 27.3 next (four times). The values: 21.1, 21.9, 22.3, 24.0, 25.4, 26.5, 28.3, and 28.8 occur three

times, and 21.3, 22.7, 23.1, 24.1. 24.3, 24.6, 24.7, 24.8, 25.8, 25.9, 27.5, 28.7, 30.8, 30.9, and 31.6 occur two times. We decide for the histogram to break the data into 7 equally spaced intervals. The histogram for this data is displayed here in Table 3.3.

Expressing this as a bar chart, we get Figure 3.1.

Table 3.3
Frequency and Cumulative Frequency Histogram for BMI Data

Class interval for BMI levels	Frequency (*f*)	Cumulative frequency (*cf*)	Relative frequency (%)	Cumulative relative frequency (%)
18.0–20.9	6	6	5.00	5.00
21.0–23.9	24	30	20.00	25.00
24.0–26.9	32	62	26.67	51.67
27.0–29.9	28	90	23.33	75.00
30.0–32.9	15	105	12.50	87.50
33.0–35.9	9	114	7.50	95.00
36.0–38.9	6	120	5.00	100.00
Total	120	120	100.00	100.00

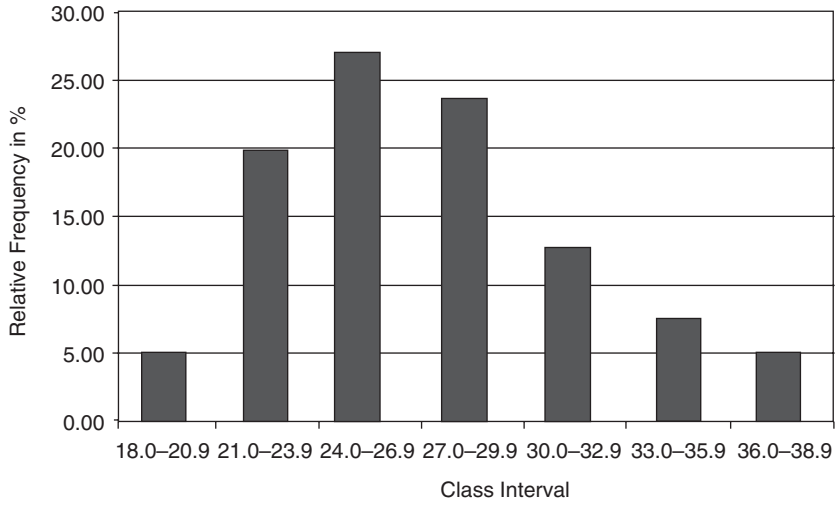


Figure 3.1. Relative frequency histogram for BMI data.

3.4 STEM-AND-LEAF DIAGRAMS

John Tukey devised a quick way to summarize the data in a similar way to a histogram but still preserving all the individual data points. He called it a stem-and-leaf diagram because it looks like a stem with the individual points protruding out like leaves on a tree branch.

We see from Table 3.4 that the leaves produce the shape of the histogram on its side. This is a little different, because we chose 21 equally spaced intervals instead of 7 so that the stem could be the first

Table 3.4
Stem-and-Leaf Diagram for BMI Data

Stems (intervals)	Leaves (observations)	Frequency
18.0–18.9	3	1
19.0–19.9	28	2
20.0–20.9	278	3
21.0–21.9	111335999	9
22.0–22.9	333457789	9
23.0–23.9	011345	6
24.0–24.9	000112335667788	15
25.0–25.9	04446778899	11
26.0–26.9	255569	6
27.0–27.9	1333344444556	13
28.0–28.9	233346778889	12
29.0–29.9	358	3
30.0–30.9	01238899	8
31.0–31.9	01366	5
32.0–32.9	68	2
33.0–33.9	26	2
34.0–34.9	289	3
35.0–35.9	0589	4
36.0–36.9	6	1
37.0–37.9	158	3
38.0–38.9	28	2
Total	—	120

two digits. The way to reconstruct the data from the diagram is as follows: The interval with the stem 18 only has one value, and the leaf is a 3. So the value of the data point is 18.3. The fourth interval from the top has a stem of 21, and there are nine observations in the intervals. The leaves are 1, 1, 1, 3, 3, 5, 9, 9, and 9. So the nine values are 21.1, 21.1, 21.1, 21.3, 21.3, 21.5, 21.9, 21.9, and 21.9. The data in the other intervals are reconstructed in exactly the same way.

3.5 BOX PLOTS

In order to appreciate box plots, we need to explain the interquartile range. The interquartile range is the middle 50% of the data. The lower end is the 25th percentile of the data, and the upper end is the 75th percentile. The width of the interquartile range is equal to:

75th Percentile–25th Percentile.

As we will see later in the text, the interquartile range is a robust measure of variability. The box plot, or, more formally, the box-and-whisker plot, is given as follows: The midline of a box-and-whisker plot is the median or 50th percentile. The body or box portion of the plot is the interquartile range going from the 25th percentile to the 75th percentile. The ends of the whiskers are given different definitions by several authors. Often, it runs in the lower end from the 1st percentile to the 25th percentile, and in the upper end from the 75th percentile to the 99th percentile. Sometimes, the lower end is the 5th percentile, and the upper end, the 95th percentile. Points beyond the ends of the whiskers are potential outliers, and are highlighted as individual dots.

The following cartoon shows an example of what the box-and-whisker plot looks like (Fig. 3.2).

3.6 BAR AND PIE CHARTS

We shall use a particular data set that we call the Pugh data to exhibit both bar charts and pie charts. Both types of charts can be applied to categorical data. Bar charts are preferred when the categories have a natural ordering. The bars are displayed across as the order of the categories increases. The pie chart is preferred to give a good idea of the proportion of the data in each category. By using a pie or circular shape,

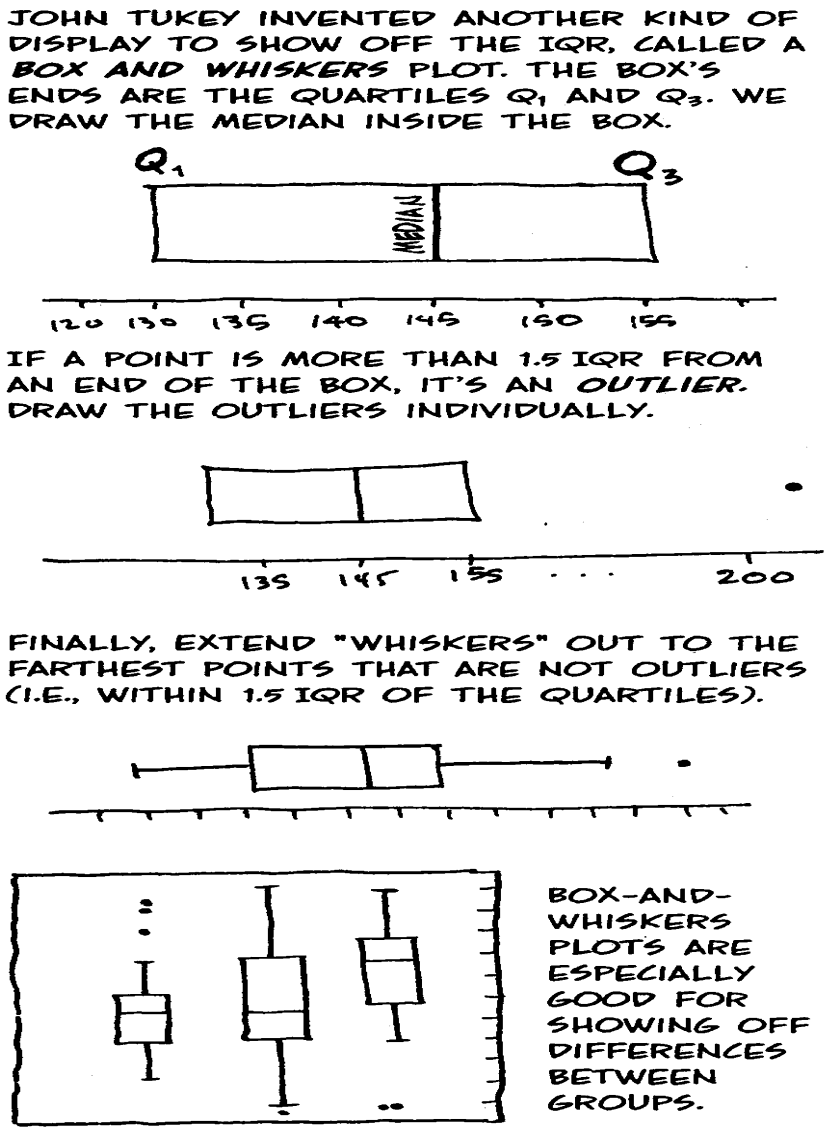


Figure 3.2. Explanation of box-and-whisker plots (taken from the *Cartoon Guide to Statistics* with permission).

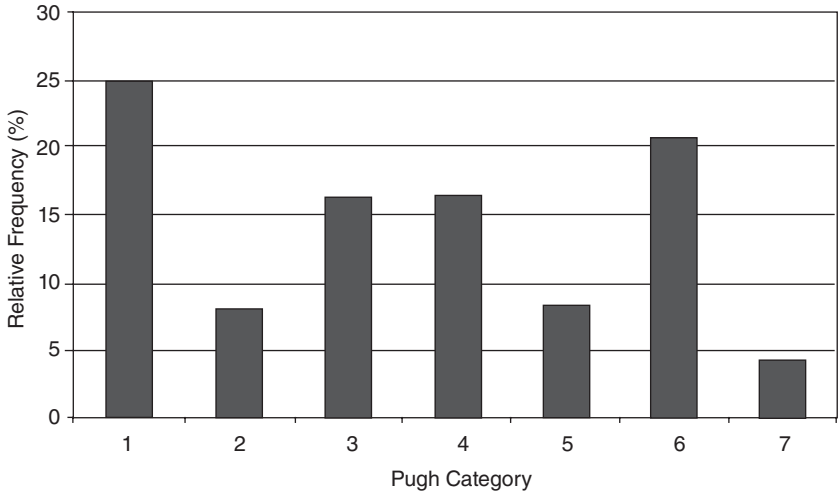


Figure 3.3. Relative frequency bar graph for Pugh categories of 24 pediatric patients with liver disease.

there is no natural order exhibited. The bar chart can also be used when there is no natural order, but it is easy for the viewer to think that, since the bars go from left to right, the bar chart, like the histogram, is displaying the categories in increasing order.

In Figure 3.3, the Pugh data provides a measure of severity of liver disease. The Pugh categories run from 1 to 7 in increasing level of severity. Here is a bar chart for the Pugh data.

Note that the bar chart looks just like a relative frequency histogram, but remember that the numbers represent categories and not intervals of real numbers. So a 2 is not twice as severe as a 1, for example. But as we move from left to right, the severity increases. So, for the Pugh, data the bar chart is appropriate

Next, we shall look at the same data viewed as a pie chart (Fig. 3.4).

It is much easier to identify the differences in proportions visually from the pie chart. The order is lost unless you recognize that order of severity starts with 1 in the upper right quadrant and increases as you move clockwise from there. For data like this, it may be useful to present both types of graphs so that the viewer will recognize both features clearly. But had there not been a natural ordering to the data, only the pie chart should be used.

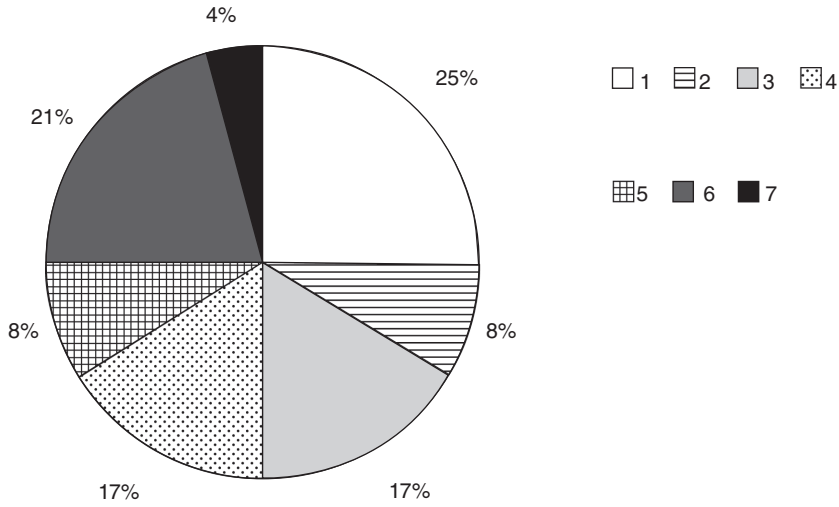


Figure 3.4. Pie chart for Pugh level for 24 children with liver disease.

Bar charts are sometimes used to compare two groups with respect to a measure. To show the variability in the data that went into the value presented by the bar, an error bar like the whisker portion of a box plot is used to show the variability. However, the box plot is a much better choice for the comparison because it shows the difference between the medians of the distribution in the proper way and provides more detail about the variability and skewness of the data.

3.7 MEASURES OF THE CENTER OF A DISTRIBUTION

There are several measures of central tendency for a data set. They include

1. arithmetic mean;
2. geometric mean;
3. harmonic mean;
4. mode; and
5. median.

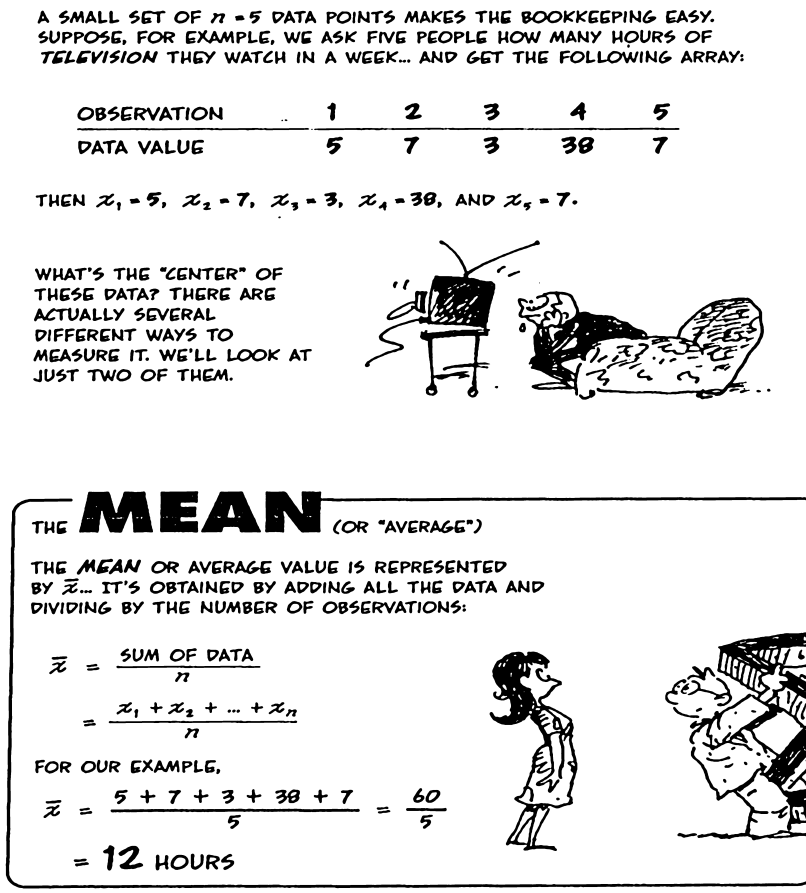


Figure 3.5. Explanation of the sample mean (taken from the *Cartoon Guide to Statistics* with permission).

Of these five, we will only discuss the most common: 1, 4, and 5.

The next cartoon (Fig. 3.5) describes the sample mean of a data set (this is the arithmetic mean, but when "arithmetic" is left off, it is understood).

Note that the value, 38, is much larger than all the other values, and may be considered an outlier. We have already seen that outliers can have a large influence on the sample mean. If we removed 38, the sample mean would be $(5 + 7 + 3 + 7)/4 = 22/4 = 5.5$.

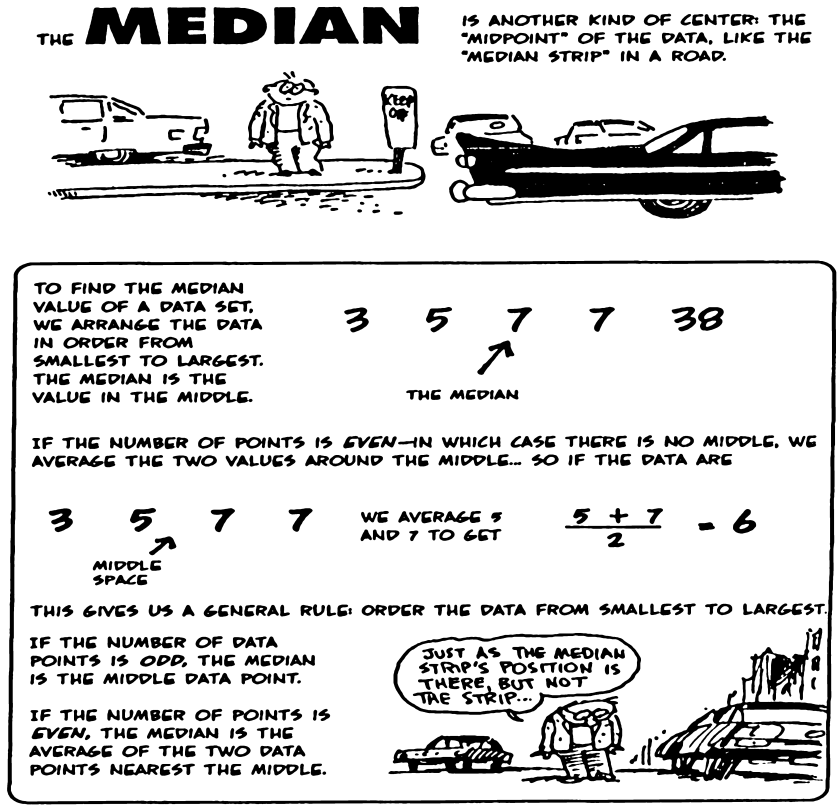


Figure 3.6. Definitions for the sample median: even and odd cases (taken from the *Cartoon Guide to Statistics* with permission).

This is a lot smaller than the average of 12, when the number 38 is included. We shall next look at the median, and for this example, see how the median is affected by the outlier.

The following cartoon (Fig. 3.6) defines median and illustrates how it is affected by the outlier in the TV viewing example.

In this case, with the value 38 included, the median is 7 (compared with 12 for the mean), and by taking the outlier out of the data set, the median drops only to 6 (compared with 5.5 for the mean). So the removal of the outlier has a big effect on the mean, dropping it by 6.5 hours, but not so large for the median, dropping it by only 1 hour. In statistics, we say that the median is "robust" with respect to outliers, and the mean is not robust. Note that when 38 is removed, the data has a distribution that is far less skewed to the right. When the data are

symmetric or close to symmetric, the mean and median are nearly equal, and the mean has statistical properties that favor it over the median. But for skewed distributions or data with one or more gross outliers, the median is usually the better choice.

For discrete data, the mode is the most frequently occurring value. Sometimes, there can be more than one mode. For continuous data, the mode of the distribution is the highest peak of the probability density function. If the density has two or more peaks of equal height that is the highest, then these peaks are all modes. Sometimes, authors will be less strict and refer to all the peaks of the probability density function to be modes. Such distributions are called multimodal, and in the case of two peaks, bimodal.

The normal distribution and other symmetric distributions (such as Student's t distribution) have one mode (called unimodal distributions), and in that case, the mode = median = mean. So the choice of the measure to use depends on a statistical property called efficiency. There are also symmetric distributions that do not have a finite mean.* The Cauchy distribution is an example of a unimodal symmetric distribution that does not have a finite mean. For the Cauchy, the median and mode both exist and are equal.

Now let's give a formal definition for the mode. The mode of a sample is the most frequently occurring value. It will not be unique if two or more values tie for the highest frequency of occurrence. Probability distributions with one mode are called unimodal. Distributions with two or more peaks are called multimodal. Strictly speaking, a distribution only has two or more distinct modes if the peaks have equal maximum height in the density (probability distribution for a continuous distribution) or probability mass function (name for the frequency distribution for a discrete distribution). However, when not strict, Figure 3.7b is called bimodal even though the peaks do not have the same height.

Figure 3.7 shows the distinction between a unimodal and a bimodal density function.

Had the two peaks had the same height, then the bimodal distribution would have two distinct modes. As it is, it only has one mode. But we still call it bimodal to distinguish it from the unimodal distribution.

* A continuous distribution has an infinite mean if $=xf(x)dx = \infty$, where $f(x)$ is the probability density function, and the integral is taken over all x where $f(x) > 0$.

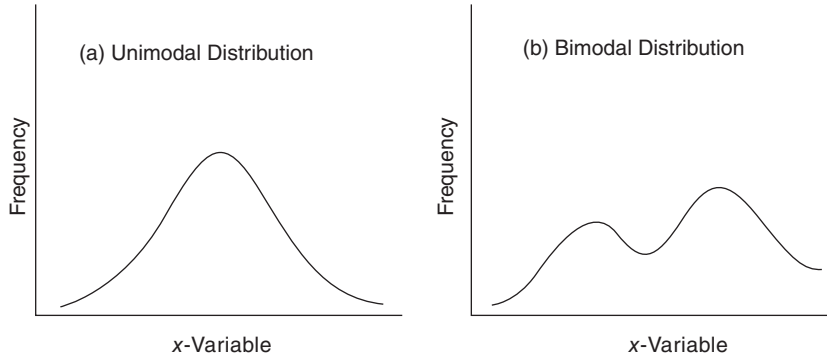


Figure 3.7. Example of a unimodal and a bimodal distribution.* The bimodal distribution in the picture has two peaks but the peak to the right is the mode because it is the highest peak.

The mean, median, and mode may all be different in the bimodal case. In the symmetric bimodal case, the mean and median may be the same, but neither of the two modes would equal the median (one will be below and the other above both the mean and the median).

For symmetric unimodal distributions: mean = median = mode. For unimodal distributions that are right skewed: mean < median < mode. For unimodal distributions that are left skewed: mean > median > mode. Although the mode can sometimes be a good measure of central tendency, at least in the case of the symmetric bimodal distribution, the natural center is in the "middle" between the two modes at where there is a trough. That middle of the valley between the peaks is where the median and mean are located.

3.8 MEASURES OF DISPERSION

Measures of dispersion or spread (also called variability) that we discuss in this section are:

* In the example above, we chose a symmetric unimodal distribution and an asymmetric bimodal distribution. Unimodal distributions can also be skewed and bimodal distributions symmetric.

1. interquartile range;

2. mean absolute deviation; and

3. standard deviation.

We have already encountered and defined the interquartile range because of its importance in illustrating the spread of the data in a box-and-whisker plot. So we will now move on to the definitions of (2) and (3) above.

The mean absolute deviation of a sample is defined as follows:

The *mean absolute deviation* (MAD) for a sample is the average absolute difference between the sample mean and the observed value. Its formula is:

$$MAD = \sum_{i=1}^{n} \left| E_i - \bar{E} \right| / n, ,$$

where E_i = the ith observation, $\bar{E}$ is the sample mean, and n is the sample size. The MAD like the median is less sensitive to outliers than the next measure we discuss, the standard deviation.

The standard deviation of a sample is the square root of the estimate of the variance. The variance measures the average of the squared deviations from the mean. The sample estimate of variance would then be given by the formula $V = \sum_{i=1}^{n} \left(E_i - \bar{E} \right)^2 / n$, where $\bar{E}$ is the sample mean, and E_i is the ith observation. Usually, we would divide by n as in the formula above, but for normally distributed observations from a random sample, this estimate is proportional to a chi-square random variable with $n - 1$ degrees of freedom.* The expected value† of a chi-square random variable is equal to its degrees of freedom. So if instead of n in the denominator we used $n - 1$ and label the estimate as S^2, then $(n - 1)\ S^2/\sigma^2$ is known to have exactly a chi-square distribution with $n - 1$ degrees of freedom, where σ^2 is the true population variance.

Therefore, as previously discussed, its expected value $E[(n - 1)\ S^2/\sigma^2] = n - 1$, or $E[S^2/\sigma^2] = 1$. But this means $E[S^2] = \sigma^2$. So S^2 is an

* Think of the degrees of freedom as a parameter that determines the shape of the chi-square distribution.

† Expected value is a statistical term for the mean of a distribution.

unbiased estimate of the population variance. So this means that $E(V) = (n-1)\sigma^2/n$, which means that V is biased on the side of underestimating σ^2. Hence, it is more common to use S^2 for the estimate. But sometimes, I think too much is made of this, because for large n, the bias is small (i.e., $(n-1)/n$ is close to 1). Furthermore, since we estimate the standard deviation by taking the square root of an unbiased estimate of the variance, that will give us a slightly biased estimate of the standard deviation anyway.

Although we have only discussed the unbiasedness property for the variance estimate S^2 when the observations come from a random sample

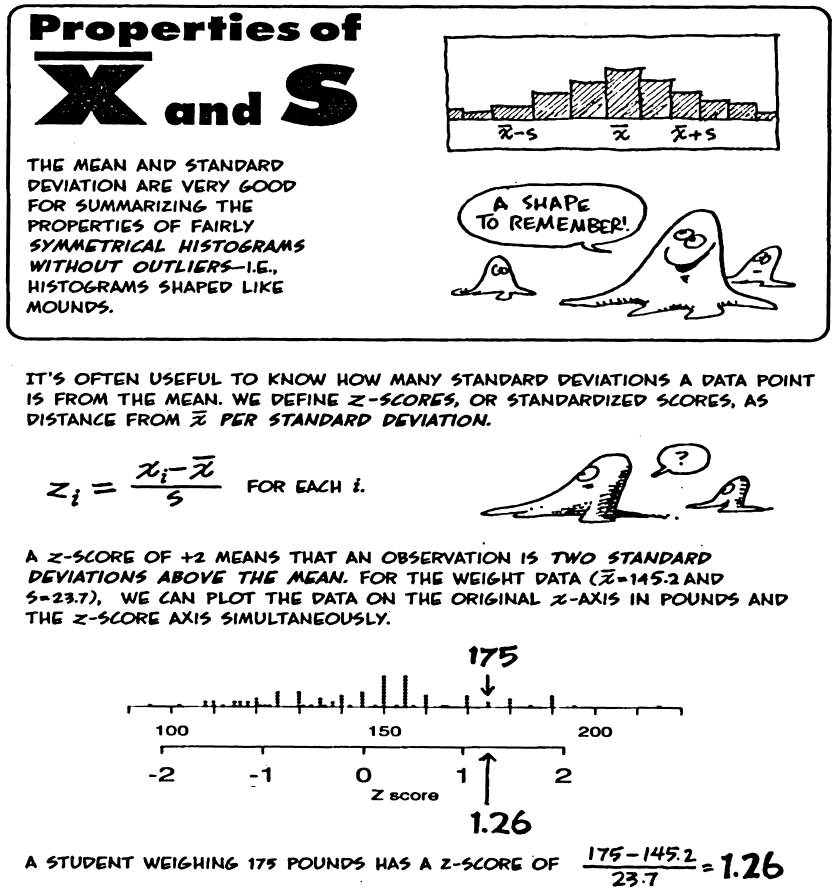


Figure 3.8. Properties of the sample mean and sample standard deviation (taken from the *Cartoon Guide to Statistics* with permission).

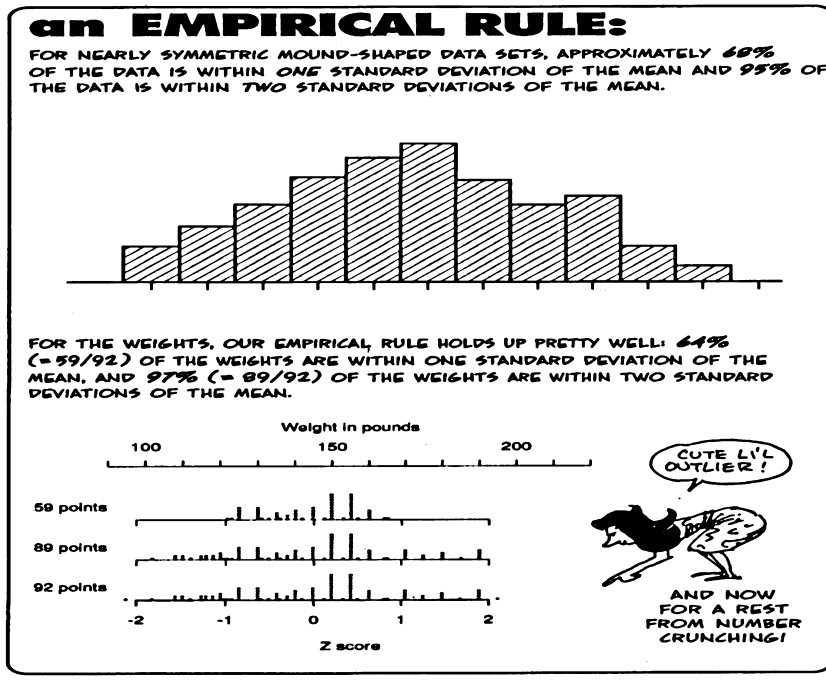


Figure 3.9. The empirical rule for mound-shaped distributions (taken from the *Cartoon Guide to Statistics* with permission).

(i.e., are independent and identically distributed normal random variables), the unbiasedness of S^2 actually holds more generally. We shall now discuss some properties of the mean and standard deviation to explain why these two measures are sometimes preferred. This is again illustrated nicely by a few cartoons (Figs. 3.8 and 3.9).

We call this an empirical rule because it was discovered by looking at mound-shaped data. It works because mound-shaped data look approximately like samples from the normal distribution, and the normal distribution has exactly those percentages given in the rule. If a distribution has a variance,* the Chebyshev inequality gives a lower bound on the percentage of cases within k standard deviations of the mean.

* A variance is defined for any finite population or finite sample. However, if a distribution has an infinite range the distribution (or infinite population) does not necessarily have a finite variance. We require $\mu = \int x f(x)\, dx < \infty$ and $\sigma^2 = \int (x - \mu)^2 f(x)\, dx < \infty$ for the distribution with density f to have a finite variance.

Chebyshev's inequality: The interval $[\mu - k\sigma, \mu + k\sigma]$ contains at least $100(1 - 1/k^2)\%$ of the distribution or data, where μ is the mean and σ is the standard deviation. Compare this with the empirical rule. Chebyshev's inequality guarantees at least 0% within 1 standard deviation of the mean (essentially guarantees nothing), while the empirical rule gives 68%. Chebyshev's inequality guarantees at least 75% with 2 standard deviations of the mean, while the empirical rule gives 95%. Chebyshev's inequality always guarantees lower percentages than the empirical rule. This is because Chebyshev's rule must apply to all distributions that have variances while the empirical rule applies only to distributions that are approximately normally distributed.

3.9 EXERCISES

1. What does a stem-and-leaf diagram show?

2. What does a relative frequency histogram show?

3. What is the difference between a histogram and a relative frequency histogram?

4. How is a relative frequency histogram different from a cumulative relative frequency histogram?

5. What portion of the data is contained in the box portion or body of a box-and-whiskers plot?

6. When are pie charts better than bar charts?

7. What relationship can you make to the three measures of location (mean, median, and mode) for right-skewed distributions?

8. What is the relationship between these measures for left-skewed distributions?

9. What is the definition of mean absolute error (deviation)?

10. What is the definition of mean square error?

11. Under what conditions does a probability distribution contain approximately 95% of its mass within 2 standard deviations of the mean?

CHAPTER 4

Normal Distribution and Related Properties

4.1 AVERAGES AND THE CENTRAL LIMIT THEOREM

How does the sample mean behave? If the sample comes from a normal distribution with mean μ and standard deviation σ, then the sample average of n observations is also normal with the mean μ, but with standard deviation $\sigma/\sqrt{n}$. So the nice thing here is that the standard deviation gets smaller as n increases. This means that our estimate (the sample mean) is an unbiased estimator of μ, and so it tends to get closer to μ as n gets large.

However, even knowing that we cannot make exact inference because what we actually know is that $Z = (\hat{X} - \mu)/(\sigma/\sqrt{n}) = \sqrt{n}(\hat{X} - \mu)/\sigma$, where $\hat{X}$ is the sample mean, has a normal distribution with mean 0 and variance 1. To draw inference about μ we need to know σ. Because σ causes difficulties, we call it a nuisance parameter. In the late nineteenth century and in the first decade of the twentieth century, researchers would replace σ with a consistent estimate of it, the sample standard deviation S. They would then do the inference assuming that $\sqrt{n}(\hat{X} - \mu)/\mathrm{S}$ has a standard normal distribution.

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

This, however, is not exactly right, because S is a random quantity and not the constant σ. However, for large n, the resulting distribution is close to the standard normal. But this is not so when n is small. Gosset, whose pen name was Student, had experiments involving small n. In this case, Gosset was able to discover the exact distribution, and a formal mathematical proof that he was correct was later derived by R. A. Fisher. We will discuss Gosset's t-distribution later in this chapter.

It is also true for any distribution with a finite variance that the sample mean is an unbiased estimator of the population mean, and if σ is the standard deviation for these observation, which we assume are independent and come from the same distribution, then the standard deviation of the sample mean is $\sigma/\sqrt{n}$. However, inference cannot be exact unless we know the distribution of the sample mean, except for the parameter μ. Again, σ is a nuisance parameter, and we will use $\sqrt{n}(\hat{X}-\mu)/\mathrm{S}$ to draw inference.

However, we no longer can assume that each observation has a normal distribution. In Gosset's case, as long as the observations were independent and normally and identically distributed with mean μ and standard deviation σ, $\sqrt{n}(\hat{X}-\mu)/\mathrm{S}$ would have the t-distribution with $n-1$ degrees of freedom. The "degrees of freedom" is the parameter for the t-distribution, and as the degrees of freedom get larger, the t-distribution comes closer to a standard normal distribution. But in our current situation where the distribution for the observations may not be normal, $\sqrt{n}(\hat{X}-\mu)/\mathrm{S}$ may not have a t-distribution either. Its exact distribution depends on the distribution of the observations. So how do we do the statistical inference?

The saving grace that allows approximate inference is the central limit theorem, which states that under the conditions assumed in the previous paragraph, as long as the distribution of the observations has a moment slightly higher than 2 (sometimes called the $2+\delta$ moment),* $\sqrt{n}(\hat{X}-\mu)/\mathrm{S}$ will approach the standard normal distribution as n gets large. Figure 4.1 illustrates this.

So we see distributions with a variety of shapes, and all have very different distributions when $n = 2$, and less different when $n = 5$, but all very close to the shape of a normal distribution when $n = 30$.

* Recall that the population mean is $E(X)$. This is called the first moment. $E(X^2)$ is called the second moment. The variance is $E[X - E(X)]^2 = E(X^2) - [E(X)]^2$, and is called the second central moment. The $2+\delta$ moment is then $\mathrm{E}(X^{2+\delta})$ with $\delta > 0$.

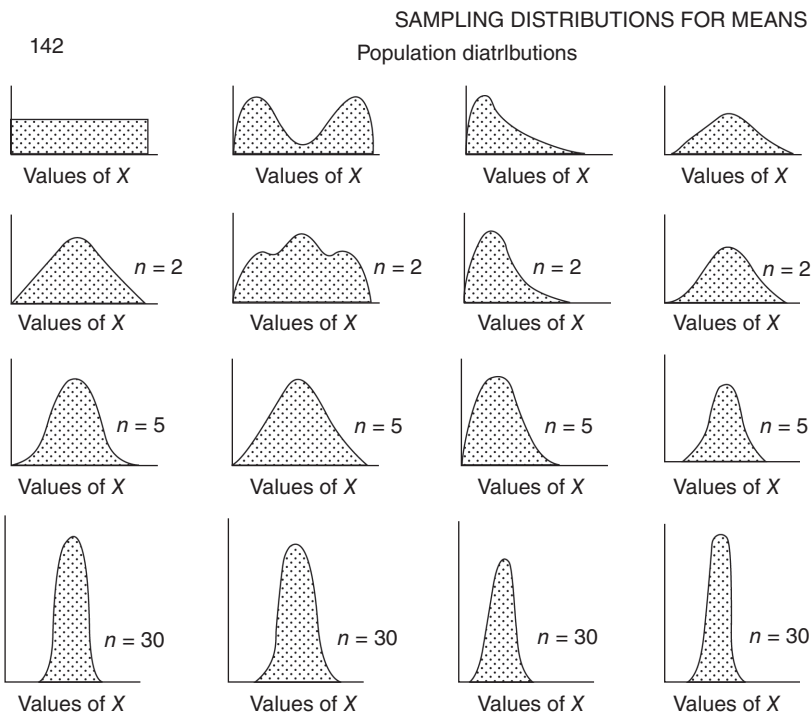


Figure 4.1. The effect of shape of population distribution and sample size on the distribution of means of random samples.
Source: Kuzma, J. W. (1984). *Basic Statistics for the Health Sciences*. Mountain View, CA: Mayfield Publishing Company, figure 7.2, p. 82.

4.2 STANDARD ERROR OF THE MEAN

The standard deviation of the sample mean is sometimes called the standard error of the mean. We have seen in the previous section that the standard error of the mean is $\sigma / \sqrt{n}$. This is very important, because it indicates that the variance approaches 0 as n gets large.

4.3 STUDENT'S *T*-DISTRIBUTION

We have already explained some of the history regarding the *t*-distribution. Now let's look at it in more detail. The *t* is a symmetric

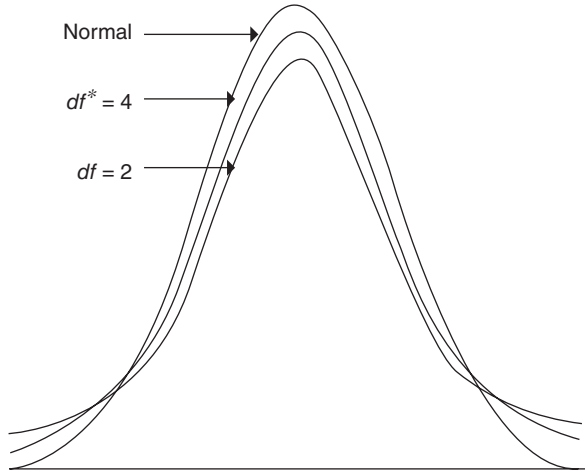


Figure 4.2. Picture of Student's *t*-distributions (2 and 4 degrees of freedom) and the standard normal distribution ^.
* *df* is an abbreviation for degrees of freedom.

unimodal distribution with fatter tails (density drops slower than the normal), and especially fatter when the degrees of freedom is 5 or less. See Figure 4.2.

Here is why the *t*-distribution is important. Our test statistic will be standard normal when we know the standard deviation and the observations are normal. But to know what the standard deviation is equal to is not common in practice. So in place of our test statistic

$$Z = (m - \mu)/(\sigma/\sqrt{n}), \text{ where } m = \sum X_i / n,$$

the sample mean, and μ is the population mean, we use

$$T = (m - \mu)/(S/\sqrt{n}), \text{ where } S = \sqrt{\left[\sum_{i=1}^{n} (X_i - m)^2/(n-1)\right]}.$$

This pivotal quantity for testing has a Student's *t*-distribution with $n - 1$ degrees of freedom, and T approaches Z as n gets large. These statements hold exactly when the X_is are independent and identically distributed normal random variables. But it also works for large n for other distributions thanks to the central limit theorem.

4.4 EXERCISES

1. What is a continuous distribution?
2. What is important about the normal distribution that makes it different from other continuous distributions?
3. How is the standard normal distribution defined?
4. For a normal distribution, what percentage of the distribution is within one standard deviation of the mean?
5. What percentage of the normal distribution falls with two standard deviations of the mean?
6. How are the median, mean, and mode related for the normal distribution
7. What two parameters determine a normal distribution?
8. In a laboratory in a hospital where you are testing for subjects with low-density lipoprotein and the distribution for healthy individuals is a particular known normal distribution, how would use this information to define abnormal amounts of lipoprotein?
9. What are degrees of freedom for a *t*-statistic for the sample mean?
10. How is the t distribution related to the normal distribution? What is different about the t statistic particularly when the sample size is small?
11. Assume that the weight of women in the United States who are between the ages of 20 and 35 years has a normal distribution (approximately), with a mean of 120 lbs and a standard deviation of 18 lbs. Suppose you could select a simple random sample of 100 of these women. How many of these women would you expect to have their weight between 84 and 156 lbs? If the number is not an integer, round off to the nearest integer.
12. Given the sample population of women as in 11, suppose you could choose a simple random sample of size 250. How many women would you expect to have weight between 102 and 138 lbs? Again round off to the nearest integer if necessary.
13. The following table shows patients with rheumatoid arthritis treated with sodium aurothiomalate (SA). The patients are divided into those that had adverse reactions (AE) and those that didn't. In addition to SA, their age is given. Of the 68 patients, 25 did not have AEs and 43 did (Table 4.1).
    (a) Construct a stem-and-leaf diagram for the ages for each group.
    (b) Construct a stem-and-leaf diagram for the total dose for each group.
    (c) Do a side-by-side comparison of a box-and-whisker plot for age for each group.

Table 4.1
Table of Rheumatoid Arthritis on Sodium Aurothiomalate

Patients without adverse reactions			Patients with adverse reactions		
Patient ID	Age	Total SA dose (mg)	Patient ID	Age	Total SA dose (mg)
001	42	1510	003	59	1450
002	67	1280	005	71	960
004	60	890	006	53	1040
007	55	1240	009	53	370
008	52	900	012	74	2000
010	60	860	014	29	1390
011	32	1200	015	54	650
013	61	1400	018	68	1150
016	48	1480	019	66	500
017	69	3300	023	52	400
020	39	2750	024	57	350
021	49	850	026	63	1270
022	36	1800	027	51	540
025	31	1340	028	68	1100
031	37	1220	029	51	1420
032	45	1220	030	39	1120
035	39	1480	033	60	990
036	55	2310	034	59	1340
037	44	1330	041	44	1200
038	41	1960	042	57	2800
039	72	960	043	48	370
040	60	1430	044	49	1920
050	48	2500	045	63	1680
055	60	1350	046	28	450
062	73	800	047	53	300
			048	55	330
			049	49	400
			051	41	680
			052	44	930
			053	59	1240

(*Continued*)

Table 4.1
(Continued)

Patients without adverse reactions			Patients with adverse reactions		
Patient ID	Age	Total SA dose (mg)	Patient ID	Age	Total SA dose (mg)
			054	51	1280
			056	46	1320
			057	46	1340
			058	40	1400
			059	37	1460
			060	62	1500
			061	49	1550
			063	55	2050
			064	52	820
			065	45	1310
			066	33	750
			067	29	990
			068	65	1100
Averages	51	1494.4	Averages	51.79	1097.91

(d) Do a side-by-side comparison of a box-and-whisker plot for total dose.

(e) Do the box-and-whisker plots for age look the same or different? What do you infer from this?

(f) Do the box-and-whisker plots for total dose look the same or different? If they are different, what are some possible explanations?

CHAPTER 5

Estimating Means and Proportions

5.1 THE BINOMIAL AND POISSON DISTRIBUTIONS

Consider a discrete variable that has two possible values, such as success or failure (e.g., success could be complete remission, while failure would be incomplete or no remission). Let 1 denote success and 0 denote failure. Suppose that we want to determine the proportion of successes in a population that for practical purposes we can consider to be infinite. We take a simple random sample of size n. We can consider this sample to represent a set of observations of n independent identically distributed random variables that each have probability p to be a success and $1 - p$ to be a failure Then the number of successes is a discrete random variable with parameters n and p, and is called the binomial distribution. As n gets large, the central limit theorem applies, and even though the binomial distribution is discrete and the normal distribution is continuous, the binomial is well approximated by the normal distribution. Sometimes, to improve the approximation due to the discrete nature of the binomial, a continuity correction is applied. However, with the increased speed of the modern computer, it is now very feasible to do exact inference using the Clopper—Pearson approach.

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

Now we will describe another important discrete distribution the Poisson distribution. In clinical trials, we often consider the time from entrance in the study to the occurrence of a particular event as an endpoint. We will cover this in more depth when we reach the survival analysis topic. One of the simplest parametric models of time to an event is the exponential distribution. This distribution involves a single parameter λ called the rate parameter. It is a good model for some time to failure data, such as light bulbs. For the exponential distribution, the probability that the time to the first event is less than t is $1 - \exp(-\lambda t)$ for $0 \le t < \infty$. The Poisson distribution is related to the exponential distribution in the following way: It counts the number of events that occur in an interval of time of a specified length t (say $t = 1$).

We have the following relationship: Let N be the number of events in the interval [0, 1] when events occur according to an exponential distribution with parameter λ. Let the exponential random variable be T. Then $P[N \ge k] = P[T \le 1]^k = [1 - \exp(-\lambda)]^k$. This relates the Poisson to the exponential mathematically. This says that there will be at least k events in [0, 1] as long as the first k events are all less than 1. So:

$$P[N < k] = 1 - [1 - \exp(-\lambda)]^k = P[N \le k - 1].$$

For $k \ge 1$, using the binomial expansion for $[1 - \exp(-\lambda)]^k$, we can derive the cumulative Poisson distribution.

The following figure shows an example of a binomial distribution with $n = 12$ and $p = 1/3$. In the figure, π is used to represent the parameter p. The Poisson distribution is also given for $\lambda = 0.87$ and $t = 1$. Note that the binomial random variable can take on integer values from 0 to 12 in this case, but the Poisson can be any integer greater or equal to zero (though the probability that $N > 5$ is very small. Also note that the probability that the number of successes is 11 is very small, and for 12, it is even smaller, while the probability of 0 or 1 success is much larger than for 11 or 12. This shows that this binomial is skewed to the right. This Poisson is also skewed right to an even larger extent (Figs. 5.1 and 5.2).

5.2 POINT ESTIMATES

In Chapter 3, we learned about summary statistics. We have discussed population parameters and their sample analogs for measures of central

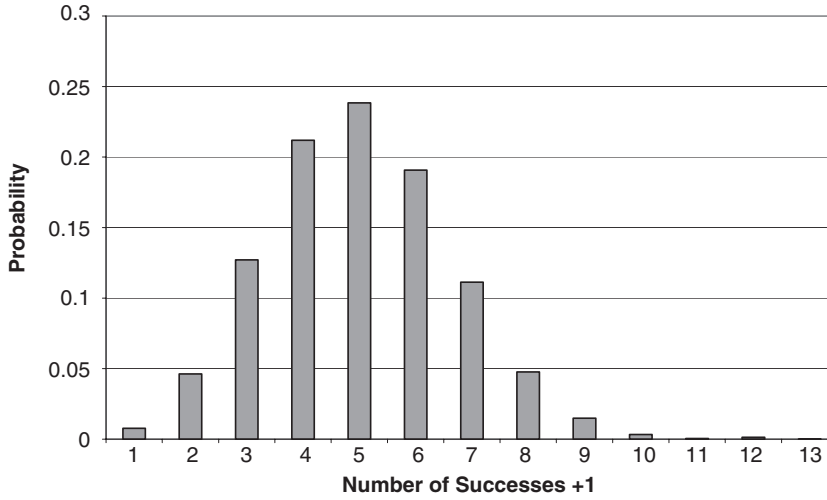


Figure 5.1. Binomial distribution $n = 12$, $p = 1/3$. Note that the number of successes is 1less than the number displayed on the x-axis. So 1 corresponds to 0 successes, 2 corresponds to 1 success, 3 corresponds to 2 successes, . . . , 13 corresponds to 12 successes.

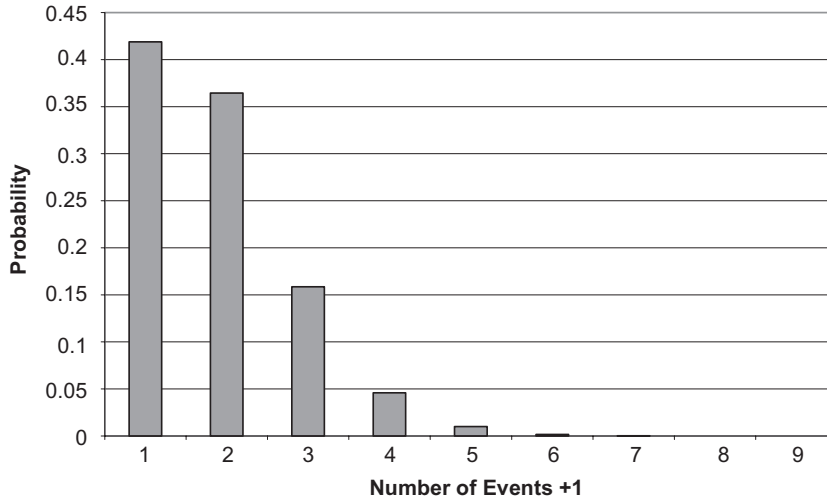


Figure 5.2. Poisson distribution with lambda = 0.87. Note that the number of events is 1 less than the number displayed on the *x*-axis. So 1 corresponds to 0 events, 2 corresponds to 1 event and so on.

tendency and dispersion. These sample analogs are often used as point estimates for the parameters. Sometimes, for a given population parameter from an assumed parametric family of distributions (e.g., the normal distribution), there are two or more possible choices for a point estimate.

For example, with continuous parametric families like the Gamma and Beta distributions, we can find maximum likelihood estimates or method of moment estimates for the parameters. How then can we choose an optimal estimate? Statistical theory has been developed to define properties that estimators should have. Among the nice properties, we have consistency, unbiasedness, minimum variance, minimum mean square error, and efficiency. Consistency is an important property. It tells us that even though the sample is random and subject to variability, as the sample size gets larger, the estimate gets close to the true parameter and will become arbitrarily close as n goes to infinity.

The sample mean is consistent because if the population distribution has mean μ and standard deviation σ, then the sample mean has for its sampling distribution mean μ and standard deviation $\sigma/\sqrt{n}$. So as n gets larger, the standard deviation goes to zero. This is enough to show consistency in probability.

The sample mean is also unbiased. To be unbiased, we must have for every n that the sampling distribution for the estimator has its mean equal to the true value of the parameter. We know this is the case for the sample mean. If we consider the class of all unbiased estimators for a parameter, we might consider the best estimate from this class to be the one with the lowest variance.

We call these minimum variance unbiased estimates. However, even a minimum variance unbiased estimator may not always be the best. Accuracy is a measure of how close the estimate tends to be to the parameter. An estimate with a small bias and small variance can be better or more accurate than an estimate with no bias but a large variance.

To see this, let us consider mean square error. The mean square error is the average of the squared distance between the estimator and the parameter. It is natural to want the mean square error to be small. Denote the mean square error by MSE, and B the bias, and σ^2 the variance of the estimator. It then happens that $MSE = B^2 + \sigma^2$. So mathematically, what we have just said in word simply means that if one estimator has $MSE_1 = B_1^2 + \sigma_1^2$, and another estimator is unbiased with mean square error $MSE_2 = \sigma_2^2$, then $MSE_2 > MSE_1$, if $\sigma_2^2 > B_1^2 + \sigma_1^2$.

This can easily happen if B_1 is small, and σ_2^2 is much larger than σ_1^2. An estimator is called efficient if as n gets large, it approaches the lowest possible mean square error. So if there is an unbiased estimator that has the smallest possible variance among all consistent estimates, then it is the best. If a biased estimator is consistent and has its variance approaching the lowest possible value, then it is efficient because the bias approaches zero under these same conditions. This is important when considering maximum likelihood estimation.

5.3 CONFIDENCE INTERVALS

Point estimates are useful but do not describe the uncertainty associated with them. Confidence intervals include the point estimate (often at the center of the interval), and they express the uncertainty by being an interval whose width depends on the uncertainty of the estimate. Formally, confidence intervals are defined as being one-sided or two-sided, and they have a confidence level associated with them. For example, a 95% two-sided confidence interval for the mean would have the interpretation that if samples of size n are repeatedly taken, and for each such sample, a 95% confidence interval for the mean is calculated, then approximately 95% of those intervals would include the population mean and approximately 5% of the intervals would not.*

As an example, we will show you how to determine a two-sided 95% confidence interval for the mean, μ, of a normal distribution when the standard deviation, σ, is assumed known. In that case, the sample mean $\hat{X}$ has a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. So let $Z = (\hat{X} - \mu)/(\sigma/\sqrt{n})$. Z has a normal distribution with mean 0 and standard deviation 1. From the table of the standard normal distribution, we have $P[-1.96 \leq Z \leq 1.96] = 0.95$. We use this fact to

* In contrast for another form of inference called the Bayesian approach, the analogue to the confidence interval is the credible interval. Because it treats parameters like random variables, a 95% credible interval is an interval that has probability 0.95 of including the parameter. This is not so for confidence intervals. The Bayesian method takes what is called a prior distribution, and based on Bayes' rule creates a posterior distribution combining the prior with the likelihood function for the data. A credible region is determined by integrating the probability density of the posterior distribution until the area under the curve between a and b, with $a < b$, integrates to 0.95.

construct the interval. Since $Z = (\hat{X} - \mu)/(\sigma/\sqrt{n}) = \sqrt{n}(\hat{X} - \mu)/\sigma$, we have $P[-1.96 \le \sqrt{n}(\hat{X} - \mu)/\sigma \le 1.96] = 0.95$. We invert this probability statement about Z into a probability statement about μ falling inside an interval as follows:

$$P[-1.96 \le \sqrt{n}(\hat{X} - \mu)/\sigma \le 1.96] = P[-1.96\sigma/\sqrt{n} \le (\hat{X} - \mu) \le 1.96\sigma/\sqrt{n}] = P[-\hat{X} - 1.96\sigma/\sqrt{n} \le -\mu \le -\hat{X} + 1.96\sigma/\sqrt{n}].$$

Then multiplying all three sides of the inequality in the probability statement by −1, we have $P[\hat{X} + 1.96\sigma/\sqrt{n} \ge \mu \ge \hat{X} - 1.96\sigma/\sqrt{n}] = 0.95$. This probability statement can be interpreted as the interval $[\hat{X} - 1.96\sigma/\sqrt{n}, \hat{X} + 1.96\sigma/\sqrt{n}]$ is a two-sided 95% confidence interval for the unknown parameter μ. We can calculate the endpoints of this interval since σ is known. However, in most practical problems σ is an unknown nuisance parameter. For n very large, we can use the sample estimate S for the standard deviation in place of σ and calculate the endpoints of the interval in the same way.

If the sample size is small, then Z is replaced by $T = \sqrt{n}(\hat{X} - \mu)/S$. This statistic T has a Student t-distribution with $n - 1$ degrees of freedom. But then to make the same statement with 95% confidence the normal percentile value of 1.96 must be replaced by the corresponding value from the t distribution with $n - 1$ degrees of freedom. From a table for the central t-distribution that can be found in many text books (Chernick and Friis (2003, p. 371), we see for $n - 1 = 4$, 9, 14, 19, 29, 40, 60, 120, we have the comparable t-percentile $C = 2.776$, 2.262, 2.145, 2.093, 2.045, 2.021, 2.000, 1.980. As the degrees of freedom get larger, C approaches the normal percentile of 1.960. So between 40 and 60, the approximation by the normal is pretty good.

The cartoon in Figure 5.3 illustrates the concept visually. In an experiment like the one shown there, since the confidence interval is a 95% two-sided interval, and the true parameter value is 0.5, we would expect 19 intervals to include 0.5 and 1 to miss. But this too is subject to variability. In the example above, all 20 intervals included 0.5, although one almost missed. If we repeated this experiment independently, we could get 19, 18, 17, or all 20 intervals containing 0.5. It is theoretically possible for a number smaller than 17 to include 0.5 but that would be highly unlikely.

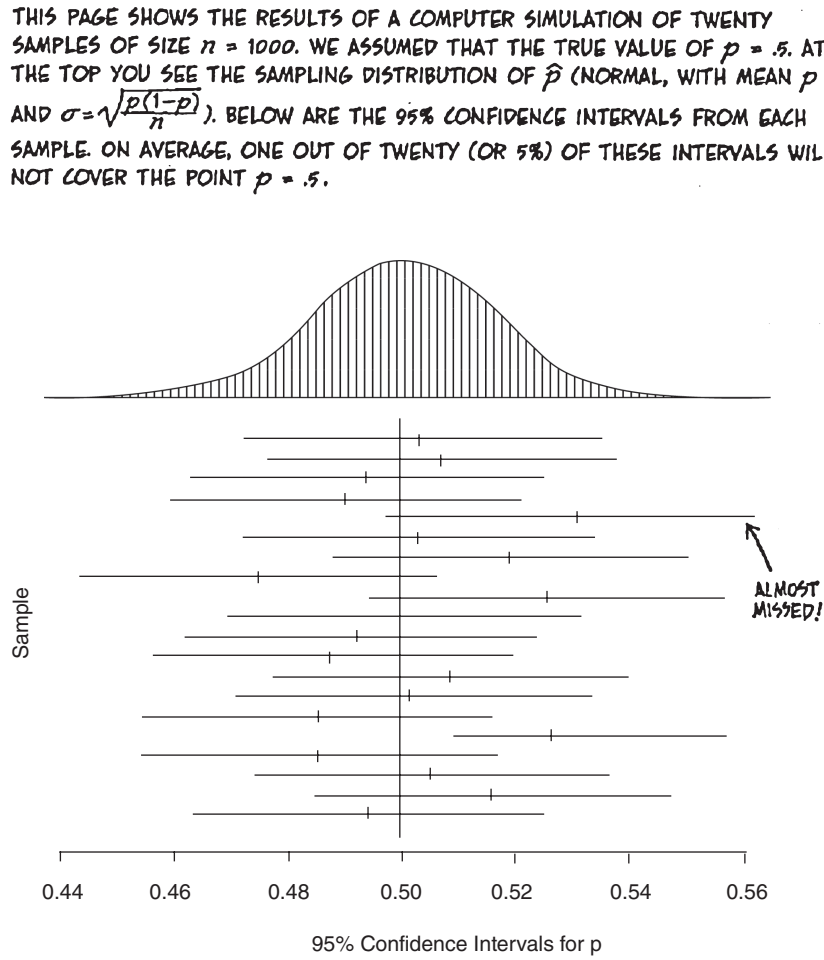


Figure 5.3. Explanation of 95% confidence interval, taken from Chernick and Friis (2003), figure 8.2, p. 156 with permission.

A one-sided confidence interval will either be an interval of the form $[a, \infty)$ or $(-\infty, b]$. These come about most often when looking at the difference of two parameters, such as arithmetic means for one group versus another. Suppose group 2 has mean greater than group 1, that is, $\mu_1 - \mu_2 < 0$. Let X be the sample mean for group 1, and let Y be the sample mean for group 2. Then we construct a confidence interval for $\mu_1 - \mu_2$ of the form $(-\infty, b]$ with say a 95% confidence level. Then, in repeating this process many times, 95% of the time, the true

mean difference will be less than the b that is determined from the sample, and 5% of the time, it will be larger.

We conclude with 95% confidence that the difference $\mu_1 - \mu_2$ is less than b. If $b < 0$, we reject the notion that the group 1 mean is larger than the group 2 mean, and conclude that group 2 has the larger mean. We do not set a finite lower limit because we are not concerned about how much larger it is. On the other hand, if we are only interested if group 1 has a larger mean than group 2, we would take an interval of the form $[a, \infty)$ and reject the notion that group 1 has a larger mean than group 2 if $a > 0$. Here we do not worry about the upper bound because we do not care how much larger it is.

In the next chapter, we will discuss hypothesis tests and will see the relationship between hypothesis testing and confidence intervals presented there. The two-tailed and one-tailed hypothesis tests correspond exactly to the two-sided and one-sided confidence intervals.

We illustrated confidence intervals for a one sample problem for simplicity. This easily extends to the two sample situation for mean differences and for other parameters in one-sample and two-sample problems for parametric families of distributions. In our examples, Z and T play the role of what we call pivotal quantities. A pivotal quantity is a random variable whose distribution is known and the resulting probability statement can be converted into a confidence interval.

Because of the 1–1 correspondence between hypothesis testing and confidence intervals, nonparametric confidence intervals can be obtained through nonparametric tests. So too can bootstrap confidence interval be defined.

5.4 SAMPLE SIZE DETERMINATION

We will demonstrate fixed sample size estimation criteria for confidence intervals using parametric assumptions. The approach is to specify a width or half-width for the interval and a confidence level. Then, the width can be expressed in terms of the sample size n. We will demonstrate that for the estimation of a population mean and for the difference between two population means.

Why is sample size determination important in medical research? When conducting an experiment or a clinical trial, cost is an important consideration. The number of tests in an experiment has an effect on the

cost of the experiment. In a clinical trial, the sample size is usually determined by the number of patients that are recruited. Each patient must get a regimen of drugs, have tests taken on each of a series of hospital visits, and be examined by the investigating doctors and nurses. The patients volunteer and do not incur much of the cost. Sometimes, the pharmaceutical company will even pay the transportation cost. So the sample size is one of the main cost drivers for the sponsor. Therefore, meeting objectives with the smallest defensible sample size is important.

To illustrate the idea, let us consider a normal distribution with a known variance, and we are simply interested in accurate estimation of the population mean. Recall that for a sample size n, a two-sided 95% confidence interval for the mean is $[\hat{X} - 1.96\sigma/\sqrt{n}, \hat{X} + 1.96\sigma/\sqrt{n}]$. The width of this interval is $2(1.96)\sigma/\sqrt{n}$, and since the interval is symmetric, we can specify the requirement equally well by the half-width, which is $1.96\sigma/\sqrt{n}$. We require the half-width to be no larger than d. Then we have $1.96\sigma/\sqrt{n} \leq d$. Since n is in the denominator of this inequality, the minimum would occur when equality holds. But that value need not always be an integer. To meet the requirement, we take the next integer above the estimated value. So we solve the equation $1.96\sigma/\sqrt{n} = d$ for n. Then $\sqrt{n} = 1.96\sigma/d$ or $n = (1.96)^2\sigma^2/d^2$.

Chernick and Friis (2003, p. 177) also derive the required equal or unequal sample sizes when considering confidence intervals for the difference of two normal means with a known common variance. Without losing generality, we take n to be the smaller sample size, and kn to be the larger sample size, with $k \geq 1$ to be the ratio of the larger to the smaller sample size. The resulting sample size n is the next integer larger than $(1.96)^2(k + 1)\sigma^2/(kd^2)$. The total sample size is then $(k + 1)n$. This is minimized at $k = 1$ but for practical reasons, we may want a larger number of patients in the treatment group in a clinical trial.

5.5 BOOTSTRAP PRINCIPLE AND BOOTSTRAP CONFIDENCE INTERVALS

The bootstrap is a nonparametric method for making statistical inferences without making parametric assumptions about the population distribution. All that we infer about the population is the distribution we obtain from the sample (the empirical distribution). The bootstrap does it in a very different way than the parametric approach. It is also

quite different from most nonparametric approaches that differ from the bootstrap because these nonparametric tests are based solely on rankings, while the bootstrap uses the actual values. Similar to parametric procedures, which require pivotal quantities, the bootstrap appears to function best when an asymptotically pivotal quantity can be used.

Recall that the difference between bootstrap sampling and simple random sampling is that

1. Instead of sampling from a population the bootstrap samples from the original sample.
2. Sampling is done with replacement instead of without replacement.

Bootstrap sampling behaves in a similar way to random sampling in that each sample is a random sample of size n taken from the empirical distribution function F_n, which gives each observation an equal chance each draw, while simple random sampling is sampling from a population distribution F (finite in population size N), but for which, unconditionally on each draw, each observation has the same chance of selection, and for the overall sample of size n, every distinct sample has the same chance $1/C_n^N$, where C_n^N is the number of ways n objects can be selected out of N as defined in Chapter 1.

The bootstrap principle is very simple. We want to draw inference about a population based on the sample without make extraneous unverifiable assumptions. So we consider sampling with replacement from the empirical distribution F_n. It is a way to mimic the sampling process. Like actors in a play, the empirical distribution acts the part of the population distribution. Sampling with replacement produces a bootstrap sample that plays the role of the original sample. Repeating the process (like performing a play over again) acts like what repeated sampling of size n from the population would be. Generating bootstrap samples is like simulating the sampling process.

We now illustrate the simplest bootstrap confidence interval, called Efron's percentile method, which is obtained by generating a histogram of bootstrap estimates of the parameter and using the appropriate percentiles to form the confidence interval. We consider an example taken from Chernick and Friis (2003).

In this experiment, a pharmaceutical company wants to market a new blood-clotting agent that will minimize blood loss during surgery

or from injury such as liver trauma. In the early stages, a small experiment to be part of the proof of concept is conducted on pigs. In the experiment, 10 pigs are in the treatment group, and 10 in the control group. All 20 pigs have the same type of liver injury induced. The control group gets a low dose of the drug, and the treatment group gets a high dose. Blood loss is measured for each pig, and we are interested in seeing if the high dose is significantly more effective at reducing the loss of blood. We will do the inference by generating 95% confidence intervals for the difference in blood loss. The sample size results are given in Table 5.1.

Perusing the data, we see the appearance that there is significantly less blood loss in the treatment pigs. If we generate a two-sided 95% confidence interval for the mean difference, assuming normal distributions with the same variance, for simplicity, the pivotal quantity involves a pooled estimate of variance, and it has a *t*-distribution with 18 degrees of freedom. The 95% confidence interval for the treatment mean—the control mean is [−2082.07, −120.93]. Since this does not contain 0, we would conclude that the treatment mean is lower than the control mean.

Table 5.1
Pig Blood Loss Data (mL)

Control pig ID number	Control group blood loss	Treatment pig ID number	Treatment group blood loss
C1	786	T1	543
C2	375	T2	666
C3	4446	T3	455
C4	2886	T4	823
C5	478	T5	1716
C6	587	T6	797
C7	434	T7	2828
C8	4764	T8	1251
C9	3281	T9	702
C10	3837	T10	1078
Sample mean	2187.40	Sample mean	1085.90
Sample standard deviation	1824.27	Sample standard deviation	717.12

However, the data suggest that the distributions are not normal, and the sample sizes are too small for the central limit theorem to take effect. Also, the equal variance assumption is highly suspect. If we use a well-known *t*-distribution approximation, the approximate 95% confidence interval is [–2361.99, 158.99]. This interval includes 0. Based on these assumptions, we cannot conclude that the means are different. For detailed calculations see Chernick and Friis (2003, pp. 163–166).

Because of the small sample size and apparent nonnormality, a nonparametric or bootstrap confidence interval would be more appropriate. Chernick and Friis (2003) also generate a bootstrap confidence interval using the percentile method.

Now, as in the case of Chernick and Friis (2003), let us compare a 95% confidence interval for the treatment mean is [572.89, 1598.91] based on the parametric method that uses the *t*-distribution. Using Resampling Stats software and generating 10,000 bootstrap samples, the bootstrap percentile method 95% confidence interval is [727.1, 1558.9]. This is quite different from the parametric interval, and is a tighter interval. The difference is another indication that the treatment data is not normal, and neither is the sample mean. The same type of result could be shown for the control group and for the mean difference.

Other bootstrap confidence intervals can be generated and are called bootstrap *t*, double bootstrap, BCa, and tilted bootstrap. See Chernick (2007) for details on these methods.

5.6 EXERCISES

1. Define the following
   (a) Inferential statistics
   (b) Point estimate of a population parameter
   (c) Confidence interval for a population parameter
   (d) Bias of an estimate
   (e) Mean square error

2. What are the two most important properties for an estimator?

3. What is the disadvantage of just providing a point estimate?

4. What is the standard error of the mean?

5. If a random sample of size n is taken from a population with a distribution with mean μ and standard deviation σ, what is the standard deviation (or standard error) of the sample mean equal to?
6. Suppose you want to construct a confidence interval for the mean of a single population based on a random sample of size n from a normal distribution. How does a 95% confidence interval differ if the variance is known versus when the variance is unknown?
7. Describe the bootstrap principle.
8. Explain how the percentile method bootstrap confidence interval for a parameter is obtained.
9. Suppose we randomly select 25 students who are enrolled in a biostatistics course and their heart rates are measured at rest. The sample mean is 66.9 and the sample standard deviation is $S = 9.02$. Assume the sample comes from a normal distribution and the standard deviation is unknown. Calculate a 95% two-sided confidence interval for the mean.
10. How would you compute a one-sided 95% confidence interval of the form $(-\infty, a]$ based on the data in exercise 9? Why would you use a one-sided confidence interval?
11. The mean weight of 100 men in a particular heart study is 61 kg, with a standard deviation of 7.9 kg. Construct a 95% confidence interval for the mean.

Table 5.2
Plasma Glucose Levels for Ten Diabetic Patients

	Plasma glucose (mmol/L)		
Patient	Before	After	Difference
01	4.64	5.44	0.80
02	4.95	10.01	5.06
03	5.11	8.43	3.22
04	5.21	6.65	1.44
05	5.30	10.77	5.47
06	6.24	5.69	–0.55
07	6.50	5.88	–0.62
08	7.15	9.98	2.83
09	6.01	8.55	2.54
10	4.90	5.10	0.20

12. Ten diabetic patients had their plasma glucose levels (mmol/L) before and after 1 hour of oral administration of 100 g of glucose. The results are shown in Table 5.2.

(a) Calculate the mean difference in plasma glucose levels.

(b) Calculate the standard error of the mean.

(c) Assuming a normal distribution for the change in plasma glucose. Based on the results in the table how many diabetic patients would you need to sample to get a 95% two-sided confidence interval for the mean difference to have width 0.5 mmol/L? Treat the estimated standard error as if it were a known constant for this calculation.

CHAPTER 6

Hypothesis Testing

The classic approach to hypothesis testing is the approach of Neyman and Pearson, initially developed in the 1930s. It differed from the approach of significance testing that was proposed by R. A. Fisher but was clear and methodical, whereas some of Fisher's ideas were obtuse and poorly explained. The differing opinions of the giants in the field of statistics led to many controversial exchanges. However, although Fisher was probably the greatest contributor to the rigorous development of mathematical statistics, his fiducial theory was not convincing and was largely discredited.

The Neyman and Pearson approach starts out with the notion of a null and alternative hypothesis. The null hypothesis represents an uninteresting result that the experimenter wants to refute on the basis of the data from an experiment. It is called the null hypothesis because it usually represents no difference, as, for instance, there is no difference in the primary endpoint of a clinical trial when comparing a new treatment with a control treatment (or placebo).

The approach fixes the probability of falsely rejecting the null hypothesis and then determines a fixed sample size that will likely result in correctly rejecting the null hypothesis, when the difference is at least a specified amount, say δ. When we reject the null

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

hypothesis, we are accepting the alternative. Whether we reject or do not reject the null hypothesis, we are making a decision, and associated with that decision is a probability that we made the wrong decision. These ideas will be discussed more thoroughly in the next section.

6.1 TYPE I AND TYPE II ERRORS

The type I error or significance level (denoted as α) for a test is the probability that our test statistic is in the rejection region for the null hypothesis, but in fact the null hypothesis is true. The choice of a cutoff that defines the rejection region determines the type I error, and can be chosen for any sample size $n \geq 1$.

The type II error (denoted as β) depends on the cutoff value and the true difference $\delta \neq 0$, when the null hypothesis is false. It is the probability of not rejecting the null hypothesis when the null hypothesis is false, and the true difference is actually δ. The larger delta is, the lower the type II error becomes. The probability of correctly rejecting at a given δ is called the power of the test. The power of the test is $1 - \beta$. We can define a power function f $(\delta) = 1 - \beta\ (\delta)$. We use the notation $\beta(\delta)$ to indicate the dependency of β on δ. When $\delta = 0$, $f(\delta) = \alpha$.

We can relate to these two types of errors by considering a real problem. Suppose we are trying to show the effectiveness of a drug by showing that it works better than placebo. The type I and type II errors correspond to false claims. The type I error is the claim that the drug is effective when it is not (i.e., is not better than placebo by more than δ). The type II error is the claim that the drug is not effective when it really is effective (i.e., better than placebo by at least δ).

However, it increases as $|\delta|$ increases (often in a symmetric fashion about 0, i.e., $f(\delta) = f[-\delta]$). Figure 6.1 shows the power function for a test that a normal population has a mean zero versus the alternative that the mean is not zero for sample sizes $n = 25$ and 100 and a significance level of 0.05. The solid curve is for $n = 25$, and the dashed for $n = 100$. We see that these power functions are both symmetric about 0, and meet with a value of 0.05 at $\delta = 0$. Since 100 is four times larger than 25, the power function increases more steeply for $n = 100$ compared to $n = 25$.

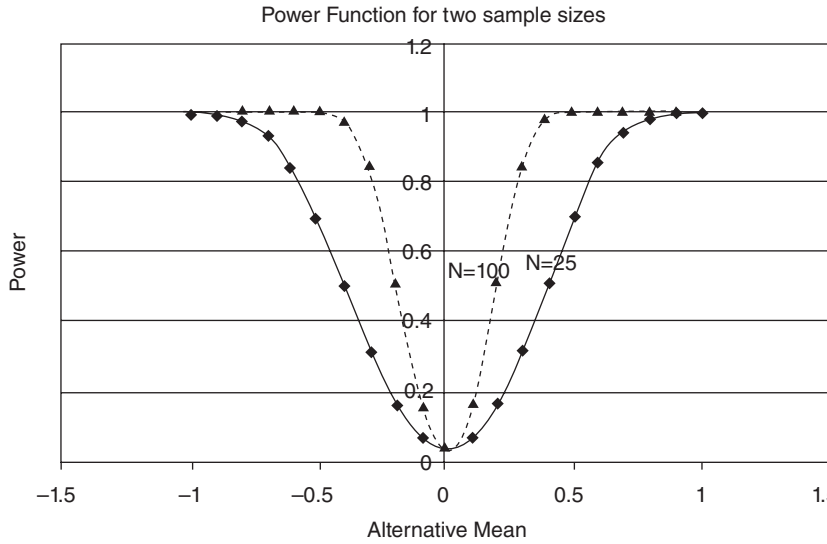


Figure 6.1. Power functions for a normal distribution with mean δ and sample sizes 25 and 100. Null hypothesis $\delta = 0$.

6.2 ONE-TAILED AND TWO-TAILED TESTS

The test described with the power function in Figure 6.1 is an example of a two-tailed test. Two-tailed tests are test where we consider both $\delta > 0$ and $\delta < 0$ as part of the alternative. A one-tailed test is a test where only one side is of interest for the alternative. So, for example, if you want to show drug A is better than drug B at lowering cholesterol, we would only be interested to see if drug A had a larger drop from baseline in cholesterol than drug B. Then, if we take $\delta =$ change from baseline for A $-$ change from baseline for B, we are interested if $\delta < 0$. But $\delta > 0$ is no more interesting than $\delta = 0$. So in this case, $\delta > 0$ is as much a part of the null hypothesis as $\delta = 0$. There are also cases where $\delta \leq 0$ is not interesting, and is included in the null hypothesis because we are only interested if we believe $\delta > 0$.

6.3 *P*-VALUES

The *p*-value is simply the probability of getting a value as extreme or more extreme than the actual value of the observed statistic when the

null hypothesis is true. There is a relationship between *p*-values and the level of the test. The *p*-value is the lowest level at which the test will reject the null hypothesis. Therefore, it is more informative about the evidence against the null hypothesis. A *p*-value can be one sided or two sided, depending whether or not the test is one or two tailed.

6.4 COMPARING MEANS FROM TWO INDEPENDENT SAMPLES: TWO-SAMPLE *T*-TEST

We start out by considering comparison of capture thresholds from two treatment groups as a way to introduce the *t*-test for two independent samples. In a clinical trial where pacing leads are implanted along with a pacemaker, we want to show that the treatment, a steroid-eluting lead attached in the heart, provides a 1 V lower capture threshold than a nonsteroid lead, the control treatment. The test hypothesis is that the difference in mean capture threshold at 6 months postimplant is zero. This is the uninteresting result that we call the null hypothesis. For the trial to be successful, we need to reject the null hypothesis in favor of the alternative hypothesis that the difference: Treatment Group Average—Control Group Average is negative.

We then choose the sample size to be large enough that we are very likely to reject the null hypothesis when the mean threshold for the treatment group is at least 1 V lower than for the control group. This we call a clinically significant difference.

If we reject the null hypothesis, we say the difference is statistically significant. We use the Neyman–Pearson approach discussed earlier. In the clinical trial, we can determine a value for the test statistic called the critical value such that we reject the null hypothesis if the test statistic is as negative, or even more negative than the critical value.

We set $\alpha = 0.05$ and do a one-sided test (i.e., only reject for large negative values, since we are only interested in showing statistically significantly lower thresholds and not significantly higher ones). This determines, based on the chosen significance level and the sample size, a critical value for the test statistic: The mean threshold difference normalized by dividing by an estimate of the standard deviation of the difference. This test statistic may be assumed to have a Student's *t*-distribution with $2n - 2$ degrees of freedom (*df*) when the null

hypothesis of zero difference is true, and n is the number of patients in the treatment group and also the number in the control group.

So then based on the t-distribution, we find a critical value, call it $-C\alpha$. If the test statistic $T \leq -C\alpha$, we reject the null hypothesis. If $T > -C\alpha$, we cannot reject the null hypothesis. As we know, the power function depends on the distribution of the test statistic under the alternative hypothesis and the chosen critical value $-C\alpha$. This distribution is a noncentral t-distribution. Just trust that statisticians can use such distributions to compute power and required sample size. It is not something that you need to learn.

In this test, we assume both samples come from normal populations with the same variance and hence the same standard deviation. This is a more realistic assumption for the pacing leads trial. Also, because steroid-eluting leads had already been approved by the FDA for a competitor, it is accepted that the steroid lead is preferred. Consequently, the patients and the sponsor would both like to see more steroid leads implanted during the trial, but still enough control leads so that the test for difference in means will have the required statistical power (generally taken to be 0.80).

The test statistic $t = (m_1 - m_2)/SD$, where m_1 is the sample mean for the first population with sample size n_1 and m_2 is the sample mean for the second population with sample size n_2 and the pooled standard deviation given by the following equation:

$$SD = \sqrt{\{(1/n_1) + (1/n_2)\}\left[(n_1 - 1) + (n_2 - 1)s_2^2 / n_1 + n_2 - 2\right]}.$$

Under the null hypothesis and the above conditions, t has Student's t-distribution with $n_1 + n_2 - 2$ df. We have seen the power function for this test in Figure 6.1.

6.5 PAIRED *T*-TEST

The tests we have studied so far that involve two populations considered independent samples. With the paired t-test, we are deliberately making the samples dependent, since we have matched pairs. The pairing is used to create positive correlation that will reduce the variability of the estimate (say the difference of two sample means). One common way to do this is to have the patient as the pairing variable.

This will usually lead to a smaller variance for the difference of the two means, such as in crossover trials or the change in test scores in a particular subject before and after an intervention.

Steps for the paired *t*-test.

1. Form the paired differences $d_i = X_{Ti} - X_{Ci}$ for $i = 1, 2, \ldots, n$, where i is the index for the paired variables (e.g., patients)
2. State the null hypothesis H_0: $\mu_T = \mu_C$ versus the alternative H_1: $\mu_T \neq \mu_C$ (or equivalently H_0: $\mu_T - \mu_C = 0$ versus the alternative H_1: $\mu_T - \mu_C \neq 0$)
3. Choose a significance level α (often $\alpha = 0.01$, 0.05, or 0.10).
4. Determine the critical region: the region that has values of the test statistic t in the upper $\alpha/2$ or lower $\alpha/2$ tails of the sampling distribution (in this case for a central t with $n - 1$ df where $n = n_T = n_C$.
5. Calculate $t = \{\hat{d} - (\mu_T - \mu_C)\}/[S_d/\sqrt{n}]$, where S_d is the standard deviation of the paired differences, and $\hat{d}$ is the mean of the paired differences.
6. Reject the null hypothesis if the test statistic t (computed in step 5 above) falls in the rejection region for the test; otherwise, do not reject the null hypothesis.

The example of daily temperatures in Washington, DC, compared with New York is next illustrate to dramatically depict the situations where pairing works best. Although this is a weather example, similar improvements can occur in clinical trials or epidemiology studies where pairing is done by patient as in a cross-over trial or propensity score matching in a case-control study. The paired data is given in Table 6.1.

The data are paired by date. We see that over the course of the year, the average temperature varies periodically with the lowest temperatures in the winter months and the highest in the summer. The date the temperatures were taking is the 15th of the month for all months in the year. Because New York and Washington are relatively close, they often share the same weather system on a particular date. Note that from this data, the highest mean temperature in Washington is 93°F occurring on July 15 and the lowest mean temperature is 26°F on December 15. For New York, the highest mean temperature is 89°F, also occurring on July 15, and the lowest mean temperature is 24°F on December 15. Over

Table 6.1
Daily Average Temperature in Washington, DC, and New York City

Date	Washington, DC (°F)	New York City (°F)
1. January 15	31	28
2. February 15	35	33
3. March 15	40	37
4. April 15	52	45
5. May 15	70	68
6. June 15	76	74
7. July 15	93	89
8. August 15	90	85
9. September 15	74	69
10. October 15	55	51
11. November 15	32	27
12. December 15	26	24

the course of the year, temperatures in DC range from 93°F to 26°F, a difference of 67°F, and in New York, from 89°F to 24°F, a difference of 65°F. But the difference in mean temperature between New York and Washington ranges only from 2 to 5°F. However, New York is always lower than DC each date of matching.

This is a clear case where this small difference would not be detectable with a two-sample (independent samples) *t*-test. But it would be easily detected by a paired *t*-test or a nonparametric approach (sign test).

6.6 TESTING A SINGLE BINOMIAL PROPORTION

The binomial distribution depends on two parameters n and p. It represents the sum of n independent Bernoulli trials. A Bernoulli trial is a test with two possible outcomes that are often labeled as success and failures. The binomial random variable is the total number of successes out of the n trials. So the binomial random variable can take on any value between 0 and n. The binomial distribution has mean equal to np and variance $np(1 - p)$. These results are to construct the pivotal

quantity for construction confidence interval and testing hypotheses about the unknown parameter p.

For a confidence interval, the central limit theorem can be applied for large n. So let $Z = (X - np - 1/2)/\left[\sqrt{n\hat{p}(1-\hat{p})}\right]$, where $\hat{p}$ is the sample estimate of p, and X is the number of successes. The estimate we use is $\hat{p} = X/n$. Z has an approximate normal distribution with mean 0 and variance 1. This is the continuity-corrected version. Removing the term $-1/2$ from the numerator gives an approximation without the continuity correction. Here Z can be used to invert to make a confidence interval statement about p using the standard normal distribution. However, for hypothesis testing, we can take advantage of the fact that $p = p_0$ under the null hypothesis to construct a more powerful test. p_0 is used in place of $\hat{p}$ and p in the definition of Z. So we have $Z = (X - np_0 - 1/2)/\left[\sqrt{np_0(1-p_0)}\right]$. Under the null hypothesis, this continuity-corrected version has an approximate standard normal distribution.

6.7 RELATIONSHIP BETWEEN CONFIDENCE INTERVALS AND HYPOTHESIS TESTS

Suppose we want to test the null hypothesis that $\mu_1 - \mu_2 = 0$ versus the two-sided alternative that $\mu_1 - \mu_2 \neq 0$. We wish to test at the 0.05 significance level. Construct a 95% confidence interval for the mean difference. For the hypothesis test, we reject the null hypothesis if and only if the confidence interval does not contain 0. The resulting hypothesis test has significance level 0.05. Conversely, suppose we have a hypothesis test with the null hypothesis $\mu_1 - \mu_2 = 0$ versus the alternative that $\mu_1 - \mu_2 \neq 0$. Look at the region of values for the test statistic where the null hypothesis is rejected. This region determines a set of values for $\mu_1 - \mu_2$ that defines a 95% confidence region for $\mu_1 - \mu_2$.

The same type of argument can be used to equate one-sided confidence intervals with one-sided tests. So what we have shown is that for every hypothesis test about a parameter with a given test statistic, there corresponds a confidence interval whose confidence level = 1 − significance level of the test. On the other hand, if we can construct a confidence interval (one or two-sided) for a parameter θ, we can define a test of hypothesis about θ based on the confidence interval, and the hypothesis test will have significance level α if the confidence level is

$1 - \alpha$. Confidence levels are often expressed as a percentage, so the confidence level for the interval is $100(1 - \alpha)\%$.

It should be noted that hypothesis tests are often constructed in a way that the test statistic assumes the value of nuisance parameters (e.g., the standard deviation of a normal distribution when testing that the mean is different from 0) under the null hypothesis. This is done because the test is designed to reject the null hypothesis, and such a formulation generally leads to a more powerful test than the one you would get by simply inverting the null hypothesis. Remember that confidence intervals have a different goal, namely to identify the most plausible values for parameter based on the data, and the null hypothesis has no relevance. For example, in hypothesis testing for a proportion, when using a normal approximation, the unknown standard deviation (which statisticians call a nuisance parameter) is replaced by the value under the null hypothesis. Under the null hypothesis, let us assume $p = 1/2$. Then if n is the sample size, the standard deviation for the sample proportion is $\sqrt{p(1-p)/n}$, or, substituting $p = 1/2$, it is $1/(2\sqrt{n})$. But for the confidence interval we would use p in place of the unknown p making it $\sqrt{\hat{p}(1-\hat{p})/n}$, which will be different in general.

6.8 SAMPLE SIZE DETERMINATION

We will again look at the pacing leads example to demonstrate sample size determination. Here we are only considering fixed sample sizes. Group sequential and adaptive designs allow the final sample size to depend on the data, and hence the sample size is unconditionally a random integer N.

What is the required sample size for a test? It depends on how big the treatment effect has to be. It also depends on the standard deviation of the test statistic. Averaging sample values reduces the standard deviation. If a random variable X has a standard deviation σ, then if you average n, such variables that have the same mean and standard deviation and are independent of each other the sample mean has standard deviation $\sigma/\sqrt{n}$. This explains why we get increasing power as we increase n.

The sample standard deviation gets smaller and approaches 0 as $n \to \infty$. So the idea is to specify a power that you want to achieve, say

0.80, at an alternative mean difference. Then pick the first value of n that achieves that desired power. Just as with the confidence intervals, we can sometimes construct a formula for the sample size. However, when things get more complicated, the sample size can still be determined numerically by computer and software packages, such as SAS, STATA, Minitab, Power and Precision, PASS 2000, and nQuery Advisor, all have capabilities to do sample size determination.

In the Tendril DX trial, one of the steroid-eluting pacing lead trials that I was involved in, an unpaired t-test (as well as a bootstrap test) were carried out. For the t-test, I assumed a common standard deviation for the capture thresholds for the steroid and the control leads. I used $\delta = 0.5\,\text{V}$ and consider the equal sample size case and the case where the treatment group gets three times the number of patients that the control group gets. The test was done at the 0.10 significance level for a two-sided test (even though a one-sided test is appropriate). The result was that 99 patients were need for the treatment group and 33 for the control, with a total sample size of 132.

On the other hand, if we were able to recruit equal numbers in both groups, we would only have need 49 in each group for a total of 98, saving 34 patients. Choosing equal sample sizes is the optimal choice if there were no practical constraints and both groups had distributions with the same variance. It would not, however, be optimal if the variances were known to be very different. Detailed output from nQuery Advisor 4.0 can be found on page 202 of Chernick and Friis (2003).

6.9 BOOTSTRAP TESTS

We shall demonstrate the use of Efron's percentile method bootstrap for testing. We will illustrate the approach with a numerical example, the pig blood loss data. Recall that previously, we listed the 10 blood loss values for the treatment group. They were 543, 666, 455, 823, 1716, 797, 2828, 1251, 702, and 1078. This gives a sample mean of 1085.9.

We found, using Resampling Stats software, that a two-sided approximate percentile method 95% confidence interval for the population mean μ (based on 10,000 bootstrap samples would be [727.1, 1558.9]. Now, consider the test where we have a null hypothesis that $\mu = \mu_0$ versus the alternative that $\mu \neq \mu_0$. Then recalling the relationship

in Section 6.7 relating confidence intervals and to hypothesis tests, we reject H_0 if $\mu_0 < 727.1$, or if $\mu_0 > 1558.9$, and do not reject if $727.1 \leq \mu_0 \leq 1558.9$. There are many other bootstrap confidence intervals, and each can be used to construct a test. See Efron and Tibshirani (1993) or Chernick (2007) for details.

6.10 MEDICAL DIAGNOSIS: SENSITIVITY AND SPECIFICITY

Screening tests are used to identify patients who should be referred for further diagnostic evaluation. To determine the quality of a screening test, it is best to have a gold standard to compare it with. The gold standard provides a definitive diagnosis of the disease. For healthy individuals, the tests, if they are numerical, there is a range called the normal range.

Formulating the screening test as a statistical hypothesis testing problem, we would see that these two types of error could be the type I and type II errors for the hypothesis test. In medical diagnosis, we have special terminology. Table 6.2 shows the possible results.

In this case, we apply a screening test to n patients with the following outcomes. Based on the gold standard, m of the patients had the disease, and $n - m$ did not. Of the m diseased patients, "a" were found positive based on the test, and c were found negative. So $m = a + c$. Of the $n - m$ patients that were not diseased based on the gold standard b tested positive, and d were found negative. So $b + d = n - m$. The off-diagonal terms represent the two types of error. The number of false positives is b, and the number of false negatives is c.

Table 6.2
Sensitivity and Specificity for a Diagnostic Test Compared to a Gold Standard

Test results	True condition of the patient based on gold standard		Total
	Diseased	Not diseased	
Positive for disease	A	b	$s = a + b$
Negative for disease	C	d	$n - s = c + d$
Total	$m = a + c$	$n - m = b + d$	$n = a + b + c + d$

The estimate of the unconditional probability of a false positive is estimated to be b/n based on this sample. The estimate of the unconditional false negative probability is c/n. Perhaps of greater interest are the conditional probabilities of error. These rates are estimates as c/m for false negative probability, given the patient has the disease (by the gold standard), and the conditional false positive probability $b/(b + d) = b/(n - m)$.

Now we shall define the specialized terms called sensitivity and specificity. *Sensitivity* is defined as the probability that a screening test declares the patient diseased given that the patient has the disease. Mathematically, the estimate of sensitivity for the above table is $1 - c/(a + c) = a/(a + c) = 1 - c/m$. So sensitivity is 1-probability of a false positive.

Specificity is the probability that the screening test declares the patient well given that the patient the patient does not have the disease (based on the gold standard). Mathematically, the specificity estimate is $1 - b/(b + d) = d/(b + d) = 1 - b/(n - m)$. So specificity is 1-probability of a false negative.

Ideally, a test should have high sensitivity and specificity. However, measurement error and imperfect discrimination rules prevent perfection (i.e., specificity = 1 and sensitivity = 1). But just as there is a tradeoff of type I and type II error when n is fixed, but the threshold is allowed to change sensitivity, and specificity can be changed to increase one at the cost of the other. So it is usually important to decide which type error is the most serious for the application and make the tradeoff accordingly. Friis and Sellers (1999) provide more detail regarding screening tests.

The curve that shows the tradeoff between specificity and sensitivity is called the receiver operating characteristic (ROC) curve. Useful references on diagnostic testing that include discussion of ROC curves are Pepe (2004), Zhou et al. (2002), Krzanowski and Hand (2009), Gönen (2007) and Broemeling (2007).

6.11 SPECIAL TESTS IN CLINICAL RESEARCH

Superiority testing is the standard testing approach in clinical trials and involves testing a null hypothesis that the treatment is no different from the control or worse than the control versus a one-sided alternative that the treatment is better or superior to the control. This is simply a

one-sided hypothesis with power function requirement at a fixed δ. So the approach is the same as in the Tendril DX lead example. Noninferiority and equivalence are different and require more detailed explanations.

Briefly for noninferiority, the null hypothesis becomes that the treatment is worse than the control by at least a δ, called the noninferiority margin. The alternative is that treatment may be better, but is at least within the margin required to say that it is not inferior. Noninferiority tests are one sided. Equivalence testing means that we want to show that the treatment and control are essentially the same (i.e., within a margin of equivalence δ). So, equivalence tests are two-sided tests that would simply reverse the null and alternative hypotheses if there were no margin (i.e., $\delta = 0$). The existence of a positive margin and the reversal of the null with the alternative make equivalence testing a little complicated, and it deserves a more detailed discussion also.

6.11.1 Superiority Tests

Not much needs to be said for superiority. It is the standard test that fits in naturally to the Neyman–Pearson approach. The one-sided two-sample *t*-test as described in Section 6.2.

6.11.2 Equivalence and Bioequivalence

Bioequivalence and equivalence are the same in terms of the formal approach to hypothesis testing. The only difference is that bioequivalence means that two drug formulations must be essentially the same in terms of their pharmacokinetic and pharmacodynamic (PK/PD) characteristics. This is common when developing a new formulation of a treatment or developing a generic replacement for an approved drug whose patent has expired.

When doing equivalence or bioequivalence testing, the conclusion you want to reach is that the two treatments are nearly the same. This is like trying to "prove" the null hypothesis. For a parameter of interest, we want to show that the difference in the estimates for the subjects on each treatment is within an acceptable range called delta. The Neyman–Pearson approach fixes the level of the test for the null hypothesis of no difference and tries to use the data to reject this hypothesis. If we reject, we have accepted the alternative because we controlled the type II error with an adequately large sample size.

In equivalence testing, we want to accept the null hypothesis. To do this in the Neyman–Pearson framework so that the type II error is controlled, we simply switch the roll of the null and alternative hypotheses. Often, in equivalence testing, it is feasible to do cross-over designs, which remove subject-to-subject variability by allowing each subject to act as their own control.

Example: You want to show that a generic drug or a new formulation of an approved drug is basically the same as the approved drug with respect to PK and PD characteristics. At Auxilium Pharmaceuticals Inc., we had a testosterone gel that was approved and trademarked as Testim®. We wanted to see if we could show that a new formulation with a better odor was equivalent in terms of the PK parameters, area under the curve (AUC), time of maximum concentration (Tmax), and value of maximum concentration (Cmax). For each of the parameters, there is a test of bioequivalence that can be performed. We designed a cross-over trial to perform these tests.

Steps in Equivalence Testing

1. Pose a clinically important difference δ.
2. State a pair of null hypotheses: H_{0L}: $d < -\delta$ and H_{OH}: $d > \delta$, where d is the observed mean difference. The alternative hypothesis is then H_1: $-\delta \leq d \leq \delta$.
3. Choose a significance level α.
4. Find the appropriate critical value (usually from the standard normal or the *t*-distribution).
5. Calculate the appropriate test statistics for the two tests of null hypotheses.
6. Compare these test statistics to their critical values, and if both null hypotheses are rejected, you have rejected nonequivalence or accepted equivalence at the level α.

In the case where the data are normally distributed, we can use Schuirmann's two one-sided *t*-tests. The same idea can be used with other tests when the data are not normally distributed. We next describe Schuirmann's test.

When each test of the null hypotheses is a one-sided Student's *t*-test, it is called Schuirmann's two one-sided *t*-tests (TOST). A simple

way to do the test is to construct a two-sided confidence interval for the mean difference, and if δ lies outside the interval, you reject non-equivalence. To ensure that the two one-sided tests each have level α, you must choose a symmetric $100(1 - 2\alpha)$ % confidence interval. This is a little counterintuitive, because for example, it is a 90% confidence interval that is used to construct a test at the 5% significance level. However, this is right, because we must reject both *t*-tests to claim equivalence.

6.11.3 Noninferiority Tests

Noninferiority is a one-sided test that a new treatment is not clinically significantly worse than a particular established treatment. Significantly worse is defined by a chosen δ just as was needed to demonstrate equivalence.

Steps in Noninferiority Testing

1. Select a clinically important difference δ.
2. State as the null hypothesis H_0: $d \geq \delta$, where $d = M_n - M_s$, and M_n is the mean for the new treatment, and M_s is the mean for the old one. Then the alternative hypothesis H_1 is that $d < \delta$.
3. Choose a significance level α.
4. Determine the critical value for the appropriate test.
5. Calculate the test statistic (d or a scaled version of it).
6. Reject H_0 if the test statistic exceeds the critical value.

6.12 REPEATED MEASURES ANALYSIS OF VARIANCE AND LONGITUDINAL DATA ANALYSIS

In clinical trials, measurements are taken on key variables at several patient visits to the site. If a change from baseline at the end of the trial is all that is of interest, conventional analysis of variance (ANOVA) or covariance can be used. However, if one is interested in how the results change over several visit (i.e., are interested in trends), then the multiple measurements on the same subject at different time points introduces correlations that conventional methods do not account for. When we

are interested in the time evolution of the measurements for many patients over a short number of visits, say 3 to 6, we are doing longitudinal analysis, and the measurements over time for a particular patient are called repeated measures.

The correlation structure within a patient must be modeled and estimated parametrically from the data. Common parametric structures for the correlation matrix are AR(1), Toeplitz, and compound symmetry, among others. These patterns correspond to statistical dependency models. For example, AR(1) is a first-order autoregressive time series model, where $Y(t) = \rho Y(t-1) + \varepsilon(t)$, $-1 < \rho < 1$, and $\varepsilon(t)$ is an independent random variable with mean 0 and constant variance for all times t. Sometimes, if there is sufficient data, the covariance can be estimated without modeling a particular correlation structure. In software packages, such as SAS, Proc Mixed is to declare the covariance to be unspecified.

In SAS, repeated measures analysis of variance can be done using the GLM procedure or the procedure "Mixed," but the two procedures handle various similar statements differently, and some cases can only be done with Proc Mixed. The Mixed Procedure is intended to do mixed effects analysis of variance, where mixed effects means that some of the effects can be treated as fixed, but other may be best modeled as random effects.

Why might we be interested in random effects? In many clinical trials, many different centers from different parts of the country or in different countries enroll patients for the trial. However, sometimes there is significant variation between the sites. We may want to see if these differences do exist, and so to do that, we model the site as a factor in the ANOVA model. Usually, it makes sense to consider the sites chosen as though they represented a random sample from the population of all potential sites. In such cases, the site becomes a random effect. Other factors may also in a similar way need to be modeled as random effects.

This topic is fairly advanced and beyond the scope of the course. But it is such an important part of clinical trials analysis. Also, missing data is a common practical problem, and mixed models provide a way to handle missing data that is sometimes appropriate. The following references provide detailed treatment of longitudinal data analysis and missing data modeling and analysis. Hardin and Hilbe (2003), Verbeke and Molenberghs (1997), Hand and Crowder (1996), and Little and

Rubin (2002) are my recommendations. Little and Rubin deal specifically with missing data and the various approaches to handling them based on the type of missing data. Hardin and Hilbe (2003) take the generalized estimating function approach, which is an alternative way of dealing with longitudinal data that we did not cover. The book gives a thorough and very readable treatment even for nonstatisticians.

6.13 META-ANALYSIS

Two problems may occur when conducting clinical trials.

1. Often a study may not have sufficient sample size to reach definitive conclusions.
2. Two or more studies may have conflicting results (not because there was anything wrong with any of the studies, but rather because type I and type II errors can occur even when the study is well powered).

A technique called meta-analysis is being used more often recently to combine information in order to reach stronger conclusions that are also more likely to be correct than what any single study might tell us. This can be done either by combining estimates or p-values in an appropriate way. Care is required in the choice of studies to be combined. Also publication bias (the bias due to a tendency to only publish positive results) is a common problem. To remedy this, for FDA regulated trials, the FDA requires posting trial information on the Internet (for all phase III trials), including all trial results and data after the trial is completed. This certainly will help to eliminate publication bias.

Hedges and Olkin (1985) was the pioneering work on formal statistical approaches to meta-analysis using frequentist approaches. Stangl and Berry (2000) provide thorough coverage of the Bayesian approach to meta-analysis. In this section, we will illustrate the use of Fisher's test for combining p-values to strengthen inference from several tests.

Fisher's test is based on the following results: Under the null hypothesis in each of k hypothesis tests, the individual p-values have a uniform distribution on [0, 1]. If we let U represent a random variable with this uniform distribution, then let $L = -2 \ln(U)$ where "ln" denotes

the natural logarithm function. Then L has a chi-square distribution with 2 *df*.

Now suppose we have k independent tests, and we let $L_k = -2\ln(V)$, where V is the product of the k independent uniforms. So $L_k = -2\ln(U_1, U_2 \ldots U_k) = -2\ln(U_1) - 2\ln(U_2) - \ldots -2\ln(U_k)$. L_k is the sum of k independent chi-square random variable with 2 *df*, and hence it is known to have a chi-square distribution with *df* equal to the sum of the *df* for the chi-square random variables being summed. So L_k is chi-square with $2k$ *df*. V is the probability that all k null hypotheses are true, which, under the independence assumption, is the product of the individual p-values. Because L_k is a simple transformation of V with a known chi-square distribution, it is more convenient to work with L_k rather than V.

We first illustrate this with a consulting application that I provided to a medical device company. The company conducted a clinical trial in the United States and some countries in Europe. The device was a cutting balloon catheter used for angioplasty. The manufacturer believed that the restenosis rate would be lower for the cutting balloon compared with conventional balloon angioplasty. Historically, the conventional approach had a disappointing 40% restenosis rate. Since the manufacturer expected the new method would have about a 25% rate, which would clearly be a clinically significant improvement, they used these assumptions to determine the necessary sample size.

Initially, the plan was to get FDA approval, which only required a study in the United States. But recruitment was going slower than they had hoped. So they chose to expand the trial to several sites in European countries. Unfortunately, the results were not consistent across the various countries. See Table 6.3.

We see that country E (which is the United States) had the lowest rate and it is below the anticipated 25%. Ironically had the company waited until the required number patients were treated in the United States, they would have had a successful trial. But even though countries C and D also have rate significantly lower than 40%, countries A and B do not, raising the question as to why. Using country as a main effect, an ANOVA clearly shows a significant difference between countries. The most likely explanation is difference in the techniques of the physicians in the European countries, where they may have had less experience with the cutting balloon catheter, or differences in the severity of the disease across the various countries.

Table 6.3
Balloon Angioplasty Restenosis Rates by Country

Country	Restenosis rate % (no. of failures/no. of patients)
A	40% (18/45)
B	41% (58/143)
C	29% (20/70)
D	29% (51/177)
E	22% (26/116)

Table 6.4
Cutting Balloon Angioplasty Combined *p*-Value Meta-Analysis by Fisher's Test

Study	Cutting balloon restenosis proportion	Conventional balloon restenosis proportion	*p*-value	$-2\ln(U)$
Grt	173/551	170/559	0.7455	0.5874
Molstad	5/30	8/31	0.5339	1.2551
Inoue	7/32	13/32	0.1769	3.4641
Kondo	22/95	40/95	0.0083	9.5830
Ergene	14/51	22/47	0.0483	6.0606
Nozaki	26/98	40/93	0.022	7.6334
Suzuki	104/357	86/188	0.001	13.8155
Combined			0.000107	42.3994

The client did have several published studies of the use of the cutting balloon for angioplasty. The hope is that combining this data with the pooled results in the clinical trial, a clinically significant improvement over the conventional rate of 40% could be shown and used to improve the case for approval of the treatment.

I conducted a meta-analysis using six independent studies with the cutting balloon along with the company's clinical trial (referred to as GRT). Table 6.4 shows the meta-analysis using Fisher's *p*-value

combination method. The other studies are labeled using the last name of the first author.

We note that the most convincing study is Suzuki, which other than GRT had the largest sample size. Also, although, some of the studies are small, most of the proportions run from 18 to 30%, making the expected 25% very plausible. This meta-analysis is conclusive even though the GRT result and the Molstad (because of small sample size) paper are not convincing. One drawback of Fisher's approach is that it treats each study equally regardless of sample size. There are other ways to combine the p-values where the studies are weighted according to sample size.

Perhaps the simplest reasonable approach would be to just total the number of restenosis events divided by the total sample size generating two proportions that can be compared directly. In this case, the proportions are 351/1214 and 379/1045 for cutting balloon and conventional balloon, respectively. These sample proportions are 28.9 and 36.3%, respectively, a difference of 7.4%.

Using a normal approximation for the two-sample two-sided test, we get an approximate value of 3.56 for the test statistic, assuming the common proportion under the null hypothesis $p_0 = 0.40$. The two-sided p-value is less than 0.002, since for a standard normal distribution $P[Z > 3.1] = 0.001$ and hence $P[|Z| > 3.1] = 0.002$. Since $3.56 > 3.1$, we know the p-value is lower.

Although this analysis may seem compelling, it would not help to get an approval. The FDA may accept results from meta-analysis, but it would require a protocol and control and approval of the clinical trials. Only GRT had a protocol and was a controlled clinical trial, with its protocol accepted by the FDA. So they would not consider this as clear and convincing statistical evidence.

Another example is based on five published studies of blood loss in pigs, comparing those with versus those without pretreatment with the clotting agent NovoSeven®. Table 6.5 shows the p-values for the individual studies and the combined p-value using Fisher's combination test. One advantage of Fisher's method is that information about the data in each study is not needed, and the tests applied in each study need not be the same. For example, in one study, a nonparametric test might be used, while in another, a parametric test is used. All that we need to know is that the studies are comparable and valid, and have the individual p-value in each case.

Table 6.5
Meta-Analysis of Five Studies of Pig Blood Loss

Study	p-value	$-2\ln(p)$
Lynn 1	0.44	1.641961
Lynn 2	0.029	7.080919
Martinowitz	0.095	4.714083
Schreiber 1	0.371	1.983106
Schreiber 2	0.086	6.91614
Total		20.33621
Combined	0.026	

In this case, we see that the combined p-value is only slightly lower than the Lynn 2 study.

The first major statistical reference on meta-analysis is Hedges and Olkin (1985). Because of the increased popularity of meta-analyses in medical research, there have been a number of excellent books appearing in recent years. This includes Rothstein et al. (2005), Hartung et al. (2008), Whitehead (2002), Stangl and Berry (2000), and Borenstein et al. (2009). Higgins and Green (2008) is a text that provides a summary of systematic reviews of intervention studies covering numerous meta-analyses for the Cochrane Group. Michael Borenstein's company has commercial software to do meta-analysis. Another good recent reference is Egger et al. (2001).

6.14 EXERCISES

1. DB3 gives baseline serum theophylline levels for patients with emphysema. Perform an equivalence test to learn if the data are free from a sex bias, that is, if mean baseline level is equivalent for men and women. There are $n_1 = 6$ women and $n_2 = 10$ men. A difference in means of at least 2 indicates a bias. The sample means and standard deviations for women and men are $m_1 = 12.67$, $s_1 = 3$ for the women, and $m_2 = 9.68$ and $s_2 = 3.65$ for the men. You can assume that the standard deviation is the same for men and women, and hence use a pooled estimate of the standard deviation for the t-test.

2. How are equivalence tests different from standard hypothesis tests?

3. What is the difference between equivalence testing and non-inferiority?

4. What is a pooled standard deviation and when can it be applied?

5. Describe the 1 : 1 correspondence between hypothesis tests and confidence intervals. How do confidence intervals give you more information than *p*-values?

6. Define the following quantities:

(a) Hypothesis test

(b) Null hypothesis

(c) Alternative hypothesis

(d) Significance level

(e) Power of the test

(f) Power function

(g) *p*-value

(h) Type I error

(i) Type II error

7. In a factory, an occupational medicine physician who was conducting a medical research study found the mean blood level of the clerical workers was 11.2 based on a sample. State the null and alternative hypotheses when testing to see if the population of clerical workers has a mean blood level of 11.2.

8. Describe the difference between a one-tailed and a two-tailed test and describe situations where one is more appropriate than the other.

9. Define specificity and sensitivity and relate them to the type I and type II error rates.

10. What are meta-analyses? Why might they be needed?

11. Based on the data in Table 6.1, do you think it is plausible that the true mean difference in temperature between New York and Washington would be 3°F? Would the power of the test be higher, lower, or the same if the true mean difference were 5°F? Does the power depend on the true mean difference? If so, why?

Table 6.6
Antibody Changes from Vaccine Given to 20 Healthy Volunteers

	Antibody concentration to type III GBS	
Volunteer no.	Before immunization	After immunization
1	0.4	0.4
2	0.4	0.6
3	0.5	0.8
4	0.5	0.6
5	0.4	0.5
6	0.5	0.5
7	0.5	0.6
8	0.4	0.5
9	0.4	0.4
10	0.6	0.7
11	0.7	10.2
12	0.7	1.1
13	0.8	0.9
14	0.9	1.2
15	1.0	1.9
16	1.0	0.9
17	1.1	2.1
18	1.0	2.0
19	1.6	8.1
20	2.1	3.8

12. A vaccine against type III group B *Streptococcus* (GBS) was tested on 20 healthy volunteers. Table 6.6 shows the results on the antibodies before and after immunization.

(a) What type of test would you apply?

(b) Would a bootstrap test be better than a paired t-test?

(c) Should the test be one-sided or two-sided? Provide justification for your answer.

CHAPTER 7

Correlation, Regression, and Logistic Regression

In this chapter, we will cover correlation, discussing the Pearson product moment correlation coefficient. The Pearson correlation coefficient (due to Karl Pearson) is a common measure of association that is interrelated with simple linear regression and goes back to the beginning of the twentieth century. It is a natural parameter of the bivariate normal distribution. So its properties and interpretation apply to two variables whose joint distribution is at least, approximately, a bivariate normal distribution. An example of a nonparametric measure of association will be discussed in Chapter 9.

Specifically, there is a mathematical relationship between the slope of the regression line in simple linear regression (only one independent variable) and the correlation coefficient. This will be shown when we cover simple linear regression. Multiple regression is an extension of linear regression to two or more independent variables, and the multiple correlation coefficient is an extension of the square of the Pearson correlation coefficient.

Logistic regression is similar to multiple regression, but whereas in multiple regression the dependent variable is a continuous numerical variable, in logistic regression, the dependent variable is binary (it can be an outcome like success or failure). The expected value of the

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

dependent variable conditional on the independent variables (which can be discrete or continuous) is a probability p, with possible values on the interval [0, 1]. Logistic regression is covered separately in Section 7.6.

7.1 RELATIONSHIP BETWEEN TWO VARIABLES AND THE SCATTER PLOT

The Pearson correlation coefficient that we will discuss in Section 7.2 measures linear association. While it may detect some forms of curved relationships, it is not the best measure for those associations. The linear association may be positive as in the equation

$$Y = 5X - 10. \tag{7.1}$$

Here X and Y are related with a positive slope of and a Y intercept of −10. We will see that this relationship with the addition of an independent random component will give a positive correlation. This simply means that as X increases, Y tends to increase. If Equation 7.1 held exactly, we would drop the word "tends." However, the addition of a random component means that if the random component is negative, the observed value of Y at $X = X_1$ could be smaller than the observed value of Y at X_0, where $X_0 < X_1$. In most cases, the data will not fall perfectly on a straight line, and so we define the difference $Y - \hat{Y}$ to be the residual at X. For example, if the fitted line happens to be $\hat{Y} = 3.5X + 2$, and at $X = 2$, we observe $Y = 8.7$, then $Y - \hat{Y} = 8.7 - (3.5(2) + 2) = 8.7 - 9 = -0.3$. So the residual at $X = 2$ is –0.3. For all the data point (X_i, Y_i) for $i = 1, 2, \ldots n$, we compute the residuals. We then square the residuals and take their sum. This is called the mean square error. Note that in this case, the slope "b" for the fitted line is 3.5, and the intercept "a" is 2. Had we used a different value for "b" and "a," we would have gotten different residuals and hence a different mean square error. The method of least squares is a common way to fit "b" and "a." It simply amounts to finding the value of "b" and "a" that makes the mean square error the smallest. This minimum is unique in many instances, and the resulting values for "b" and "a" are called the least squares estimates of the slope and intercept, respectively. Note that this minimum will be greater than 0 unless all the points fall exactly on a straight line.

Table 7.1 is a listing of systolic and diastolic pressure for 48 elderly men. Figure 7.1 is a scatter plot of this data.

A scatter plot is used to portray the relationship between two variables. It displays the relationship by marking the data on a grid of (X, Y) pairs. This is a plot in Cartesian coordinates of the measurements X and Y for the individual subjects. In the case of Figure 7.1, X is the

Table 7.1
Systolic and Diastolic Blood Pressure for a Sample of 48 Elderly Men

Subject number	Systolic blood pressure	Diastolic blood pressure
01	140	78
02	170	101
03	141	84
04	171	92
05	158	80
06	175	91
07	151	78
08	152	82
09	138	81
10	136	80
11	173	95
12	143	84
13	117	75
14	141	83
15	120	76
16	163	89
17	155	97
18	114	76
19	151	90
20	136	87
21	143	84
22	163	75
23	141	81
24	163	94

Table 7.1
(Continued)

Subject number	Systolic blood pressure	Diastolic blood pressure
25	145	81
26	151	83
27	134	85
28	178	99
29	128	73
30	147	78
31	146	80
32	160	91
33	173	79
34	143	87
35	152	69
36	137	85
37	146	83
38	162	83
39	158	77
40	152	86
41	152	93
42	106	67
43	147	79
44	111	71
45	149	83
46	137	77
47	136	84
48	132	79

systolic blood pressure and Y is the corresponding diastolic blood pressure taken at the same time. If the two variables are highly positively correlated, the pattern of dots will closely resemble a straight line with a little scatter.

In Figure 7.1, we can perhaps visualize a straight line running through the data but mainly what we observe is a tendency for the

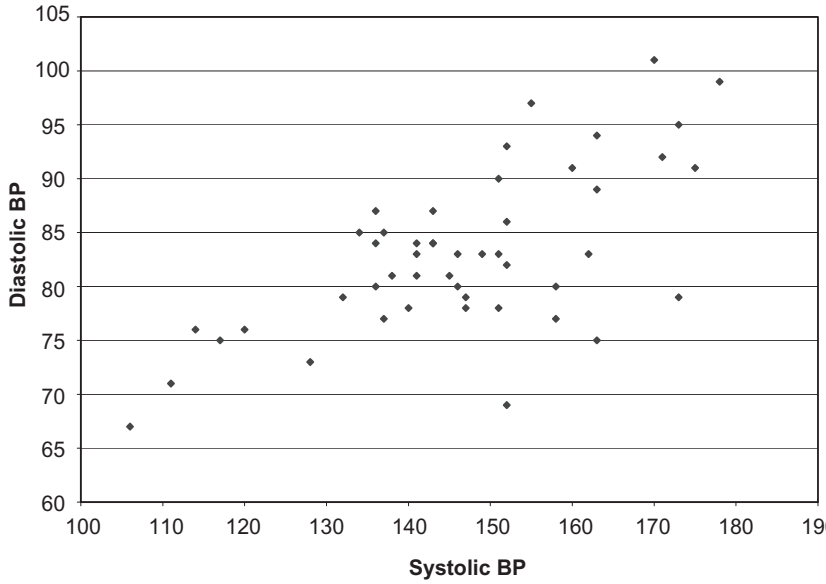


Figure 7.1. Scatter diagram of systolic and diastolic blood pressure (data from Table 7.1).

diastolic pressure to be higher than its average when the systolic pressure is higher than its average and lower than its average when the systolic pressure is below its average. Next in Section 7.2, we will see how the Pearson correlation describes aspects of the scatter plot, and in Section 7.3, how a regression line can be plotted through the data using the least squares criterion.

7.2 PEARSON'S CORRELATION

The Pearson correlation coefficient is a parameter that has a natural place in the bivariate distribution. When we have a sample such as in the scatter plot for systolic and diastolic blood pressure, we can obtain a sample estimate. Generally, ρ is used to denote the parameter, and r to denote the sample estimate of ρ.

$$r = \sum_{i=1}^{n}\left(X_i - \hat{X}\right)\sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)\Big/\sqrt{\sum_{i=1}^{n}\left(X_i - \hat{X}\right)^2\sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2}\,.$$

Here, $\hat{Y}$ and $\hat{X}$ are their respective sample means, and X_i and Y_i are the respective blood pressure readings for the ith subject. This formula is best for understanding the meaning because it shows r to be the ratio of the sample estimate of the covariance between X and Y divided by the square root of the product of their variances. To see it, divide numerator and denominator by n. In the denominator, rewrite n as $\sqrt{nn}$. Put one n under the sums involving X, and one under the sums involving Y. Then the denominator is an estimate of the square root of the product of sample variances, and the numerator is a sample estimate of the covariance.

A more complicated computational formula calculates r faster, and is mathematically equivalent to the expression above. Both r and ρ have the property that they can take on any value in [−1, 1], but cannot take on a value outside that interval.

One common hypothesis test that is conducted when the data is suspected to be correlated is to test that the population correlation coefficient $\rho = 0$ versus the two-sided alternative that $\rho \neq 0$. This test is the same as the test that the slope of the regression line is 0. Under the null hypothesis that $\rho = 0$, the quantity $r(\sqrt{n-2})/\sqrt{1-r^2})$ has a t-distribution with $n - 2$ degrees of freedom.

In this case, if we reject the null hypothesis of no correlation, we can conclude that the two variables are related. But it does not address the issue of why they are related. Often, we study relationships between variables because we suspect a causal link. The simple test for correlation cannot provide us with information on causation. Sound theory is required to make the causal link. In many situations, we must at least have the value for X occur before Y in time for X to be able to cause Y, and in such situations, we can rule out the possibility that Y causes X. Over the past 20 years, a great deal of research in statistical modeling has led to advances in finding models and plausible assumptions where a significant relationship can imply a causal relationship. In this branch of statistics, which is sometimes called causal inference, the names of Pearl, Rubin, and Robins stands out. Some articles that may be of interest are Robins (1999), Hernán et al. (2000), and Hernán et al. (2005). There are also the books, Pearl (2000, 2009), Rubin (2006), and van der Laan and Robins (2003, 2010).

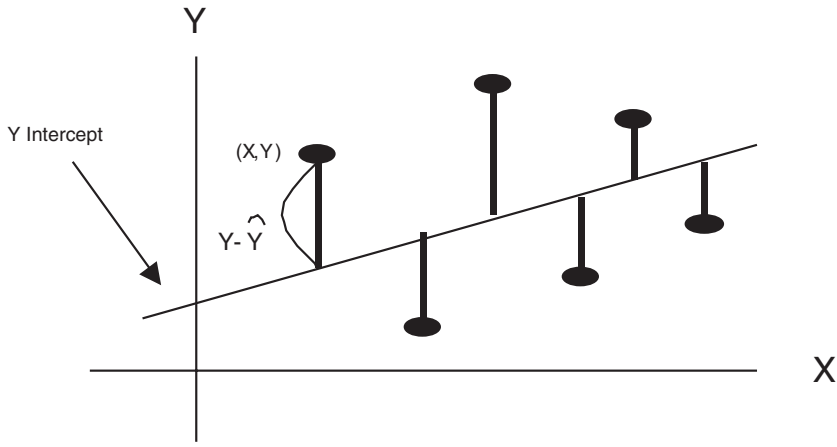


Figure 7.2. Scatter plot for six observations illustrating regression line and residuals.

7.3 SIMPLE LINEAR REGRESSION AND LEAST SQUARES ESTIMATION

A scatter plot of six points with a line fit through it is illustrated in Figure 7.2, taken from figure 12.3 from Chernick and Friis (2003)

The least squares solution for the slope and intercept is the one (the solution is not always unique) that picks a value for "*b*," the slope estimate, and "*a*," the intercept estimate, so that. Now *b* is related to r as we shall show. The least squares estimate of b is

$$b = \sum_{i=1}^{n} \left(X_i - \hat{X}\right)\left(Y_i - \hat{Y}\right) \Big/ \sum_{i=1}^{n} \left(X_i - \hat{X}\right)^2.$$

Let

$$S_y = \sqrt{\sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2 \Big/ (n-1)} \quad \text{and} \quad S_y = \sqrt{\sum_{i=1}^{n} \left(X_i - \hat{X}\right)^2 \Big/ (n-1)}.$$

Then

$$b = (S_y / S_x) r.$$

The least squares estimate of "a" is obtained by solving

$$\hat{Y} = a + b\hat{X} \quad \text{or} \quad a = \hat{Y} - b\hat{X}.$$

To illustrate solving a simple linear regression problem, we use weight and height data obtained for 10 individuals. The data and the calculations are illustrated in Table 7.2.

We first calculate b and a.

$$\sum (X - \hat{X})(Y - \hat{Y}) = 201 \text{ and}$$

$$\sum (X - \hat{X})^2 = 9108.90.$$

$$\text{So } b = 201/9108.90 = 0.0221.$$

$$\text{Then} \qquad a = \hat{Y} - b\hat{X} = 63 - (0.0221)154.10 = 59.59.$$

Now let us see how we would get confidence intervals for new values of Y when $X = x$. First let us define a few terms.

1. Sum of squares error: $SSE = \Sigma(Y_i - \hat{Y})^2$
2. Standard error of estimate: $S_{y.x} = \sqrt{[SSE/(n-2)]}$

Table 7.2
Calculations for Inference about the Predicted *Y*-Value and the Slope of the Regression Line

Subject ID	X = Weight (lbs)	$X - \hat{X}$	$(X - \hat{X})^2$	Y = Height (in)	Predicted height Y_p	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
01	148	–6.1	37.21	64	62.87	1.13	1.29
02	172	17.9	320.41	63	63.40	–0.40	0.16
03	203	48.9	2391.21	67	64.08	2.92	8.52
04	109	–45.1	2034.01	60	62.00	–2.00	4.01
05	110	–44.1	1944.81	63	62.02	0.97	0.95
06	134	–20.1	404.01	62	62.56	–0.56	0.31
07	195	40.9	1672.81	59	63.90	–4.90	24.05
08	147	–7.1	50.41	62	62.84	–0.84	0.71
09	153	–1.1	1.21	66	62.98	3.02	0.15
10	170	15.9	252.81	64	63.35	0.65	0.42
Total	1541		9108.9				49.56

3. Standard for $\hat{Y}$ given $X = x$:

$$SE(\hat{Y}) = S_{y.x}\sqrt{n^{-1} + (x - \hat{X})^2 / \sum (X_i - \hat{X})^2}\,.$$

A $100(1 - \alpha)\%$ confidence interval for predicted value of Y given $X = x$ is then $[Y\hat{} - t_{n-2}(\alpha)SE(Y\hat{}),\ Y\hat{} + t_{n-2}(\alpha)SE(Y\hat{})]$, where $t_{n-2}(\alpha)$ is the $100(1 - \alpha/2)$ percentile of a t-distribution with $n - 2$ degrees of freedom.

A confidence interval for a prediction of Y is sometimes called a prediction interval. Now let's go through the steps above to get a prediction interval for Y given $X = 110$ for the example in Table 7.2. SSE = 49.56 (see table). Then $S_{y.x} = \sqrt{(49.56/8)} = 2.73$. So,

$$SE(Y\hat{}) = 2.73(\sqrt{10^{-1}} + (110 - 154.1)^2 / (108.9) = 0.56.$$

Hence, a 95% prediction interval for Y given $X = 110$ is

$$[62.02 - 2.306(0.56), 62.02 + 2.306(0.56)] = [60.73, 63.31].$$

To test the hypothesis H_0 $\beta = 0$, where β is the slope parameter of the regression equation, we use the test statistic $t = (b - \beta)/SE(b) = b/SE(b)$, since the hypothesized value for $\beta = 0$. Now $SE(b) = S_{y.x} / \left[\sum (X_i - \hat{X})^2\right]$. For our example, $SE(b) = 2.73/\sqrt{9108.9} = 0.0286$. So, $t = 0.77$. We refer to a t-distribution with 8 degrees of freedom to determine the p-value, and we cannot conclude that β is significantly different from 0. Since ρ is a simple multiple of β, we also cannot conclude that ρ is significantly different from 0. In this case, the p-value is greater than 0.2, since it is two-sided, and the 90th percentile of a t-distribution with 8 degrees of freedom is 1.397 and $t = 0.77 << 1.397$.

One important point about simple linear regression is the term linear. Linear here means that the relationship is linear in the parameters and not necessarily in the independent variable. So although we often think of the linear regression equation as $Y = \beta X + \alpha$, this is both linear in the independent variable X, as well as the parameters α and β. However, the equations $Y = \beta X^2 + \alpha$, or $Y = \beta \ln(X) + \alpha$, also fit the simple linear regression model, although they involve nonlinear functions of the independent variable X.

The term nonlinear regression is defined as a form of regression where the equation for Y is nonlinear in the parameters that we wish to estimate. So, for example,

$$Y = g(x, \theta) = \theta_1 + \theta_2 \exp(\theta_3 x),$$

is a simple nonlinear regression because it is nonlinear in the parameter θ_3, and it cannot be transformed into a linear regression. Multiple linear regression is also linear in the parameters. There is a nonlinear regression analog to simple nonlinear regression. However, we will not cover nonlinear regression, and the interested reader should refer to Gallant (1987) and Bates and Watts (1988), which are both authoritative texts on nonlinear regression.

7.4 SENSITIVITY TO OUTLIERS AND ROBUST REGRESSION

Outliers are unusual or extreme observations within a given data set. We might expect laboratory data and other measured data taken on humans to be normally distributed, with approximately 95% of the cases falling within two standard deviations of the mean. Nevertheless, particularly in large samples, extreme values may occur. This could be due to the actually occurrence of an extreme value from the normal distribution, or it could be a measurement, coding, or data entry error. In small samples, this is also possible with all the same explanations. However, the chance of an extreme outcome from a normal distribution is much less likely to occur in small samples.

For the simple linear regression problem discussed in the previous section, we showed how to compute the slope and intercept parameters as a least squares solution. Since the method involves minimizing the sum of squared residuals, these parameter estimates are very sensitive to outliers. This is analogous to the sensitivity of the sample mean to outliers. The sample mean is the least squares estimate of the mean from a sample of independent identically distributed observations. Because the sum of squares is minimized, outliers pull the estimate toward their value, and hence execute great influence than observations near the true population mean.

In regression, the slope is pulled up or down depending on the direction of the outlier. Robust regression methods are used to minimize the influence of outliers at the price of statistical efficiency. However, when outliers are possible, the sacrifice in efficiency is often

more than made up for by the reduction in the bias that the outlier(s) may cause. Later, we shall see that there are also diagnostics that can be used in regression to tell when the least squares estimates are influenced by outliers.

There are two strategies for dealing with outliers in regression. One is to detect and remove the outliers. The other is to use a robust regression procedure in place of least squares. Robust regression is sometimes preferred because it is viewed as accommodating outliers, whereas the removal of an outlier really is a statement that the data point has no value toward the estimation of the parameter.

Deciding that the outlier is an erroneous observation is not something that you can know by just looking at the data, and so removal of outliers should only be done when, after checking the source for generating the data, an actual error is identified. Outliers in regression also have greater influence on the slope when they are near the upper or lower limits on the x-axis. These outliers are called leverage points. In general, any point near the upper or lower limits on the x-axis is a leverage point. But if the leverage point does not affect the estimate very much when it is removed, it is not an outlier with respect to the bivariate distribution of X and Y.

One robust regression method is to find the estimated coefficients that minimize the sum of absolute errors. By doing this, the outliers have less influence than when the deviations are squared. In the case of the mean, a robust sample estimate is the median. It turns out that the median minimizes the sum of absolute deviations of the observations from the estimate. So taking the mean absolute error for regression parameter estimates is analogous to using the median as an estimate of the mean for a simple random sample. There are many other robust regression procedures. We will not cover them here. See Huber (1981) or Maronna et al. (2006) for the details.

I choose a very dramatic example from the 2000 presidential election votes counted in the state of Florida. Although this is not a medical example, it is a very familiar example that makes the case very well. The number of votes received by Patrick Buchanan in Palm Beach County was very high relative to other counties, and hence represents an outlier.

You may recall that the Gore campaign contested the voting results in Florida due to several irregularities that they believe cost

Gore votes. They contended that due to the closeness of the race between Bush and Gore, the contested votes could swing the state over from Bush to Gore, and hence the election would go to Gore, since Florida had enough electoral votes to change the outcome of the votes in the electoral college. One irregularity was the butterfly-shaped ballot in Palm Beach County. The democrats theorized that the unusual butterfly shape of the ballot could confuse voters, and they could mark Buchanan's box thinking that they had voted for Gore.

In this case we could put the data on a scatter plot with counts by county plotted for Gore versus Buchanan, Bush versus Buchanan, or even Ralph Nader versus Buchanan, and in each case, Palm Beach County will be a huge outlier. The data were made public on the Internet, and many statisticians analyzed the data in a variety of ways. Table 7.3 shows the vote by county for Gore, Bush, and Buchanan. There are a total of 67 counties in Florida.

There are many things that can be observed from the data, and the scatter plots will help greatly. First, Gore and Bush were the main party candidates and received the lion's share of the votes. Buchanan and Nader were alternative party candidates, and Nader received far more votes than Buchanan. The number of votes from county to county varies quite a bit for all candidates simply because the population size of the counties varies so much. Dade (where Miami is located), Broward, and Palm Beach are the three largest counties in Florida.

Palm Beach is a heavily populated by registered Democrats, and so Gore won the county by a large margin 268,945 to 152,846. Not including Palm Beach County, Buchanan's votes ranged from 9 in Glades to 1010 in Pinellas. Gore's vote totals ranged from 788 in Lafayette to 386,518 in Broward County. Bush's vote totals ranged from 1316 in Liberty to 289,456 in Dade County. In Palm Beach County, Buchanan got 3407 votes. This is more than three times the amount of votes he got in any other county! By comparison, in Broward and Dade counties, Buchanan only got 789 and 561, respectively. Figure 7.3 shows Gore's votes versus Buchanan's, and Figure 7.4 shows Bush's votes versus Buchanan's.

This seems to present a prima facie case that there is some irregularity going on in Palm Beach County, and that it is likely that many of the votes for Buchanan were not intended for him. But then

Table 7.3
2000 Presidential Election Vote by County in Florida for Gore, Bush, and Buchanan

County	Gore votes	Bush votes	Buchanan votes	County	Gore votes	Bush votes	Buchanan votes
Alachua	47,300	34,062	262	Lake	36,555	49,965	289
Baker	2392	5610	73	Lee	73,530	106,123	306
Bay	18,850	38,637	248	Leon	61,425	39,053	282
Bradford	3072	5413	65	Levy	5403	6860	67
Brevard	97,318	115,185	570	Liberty	1011	1316	39
Broward	386,518	177,279	789	Madison	3022	3038	29
Calhoun	2155	2873	90	Mantee	49,169	57,948	272
Charlotte	29,641	35,419	182	Marion	44,648	55,135	563
Citrus	25,501	29,744	270	Martin	26,619	33,864	108
Clay	14,630	41,745	186	Monroe	16,483	16,059	47
Collier	29,905	60,426	122	Nassau	6952	16,404	90
Columbia	7047	10,964	89	Okaloosa	16,924	52,043	267
Dade	328,702	289,456	561	Okeechobee	4588	5058	43
De Soto	3322	4256	36	Orange	140,115	134,476	446
Dixie	1825	2698	29	Osceola	28,177	26,216	145
Duval	107,680	152,082	650	Palm beach	268,945	152,846	3407
Escambia	40,958	73,029	504	Pasco	69,550	68,581	570
Flager	13,891	12,608	83	Pinellas	200,212	184,884	1010

Table 7.3
(Continued)

County	Gore votes	Bush votes	Buchanan votes	County	Gore votes	Bush votes	Buchanan votes
Franklin	2042	2448	33	Polk	74,977	90.101	538
Gadsden	9565	4750	39	Putnam	12,091	13,439	147
Gilchrist	1910	3300	29	Santa Rosa	12,795	36,248	311
Glades	1420	1840	9	Sarasota	72,854	83,100	305
Gulf	2389	3546	71	Seminole	58,888	75,293	194
Hamilton	1718	2153	24	St. Johns	19,482	39,497	229
Hardee	2341	3764	30	St. Lucie	41,559	34,705	124
Hendry	3239	4743	22	Sumter	9634	12,126	114
Hernando	32,644	30,646	242	Suwanee	4084	8014	108
Highlands	14,152	20,196	99	Taylor	2647	4051	27
Hillsborough	169,529	189,713	845	Union	1399	2326	26
Holmes	2154	4985	76	Volusia	97,063	82,214	396
Indian Ruver	19,769	28,627	105	Wakulla	3835	4511	46
Jackson	6868	9138	102	Walton	5637	12,176	120
Jefferson	3038	2481	29	Washington	2796	4983	88
Lafayette	788	1669	10				

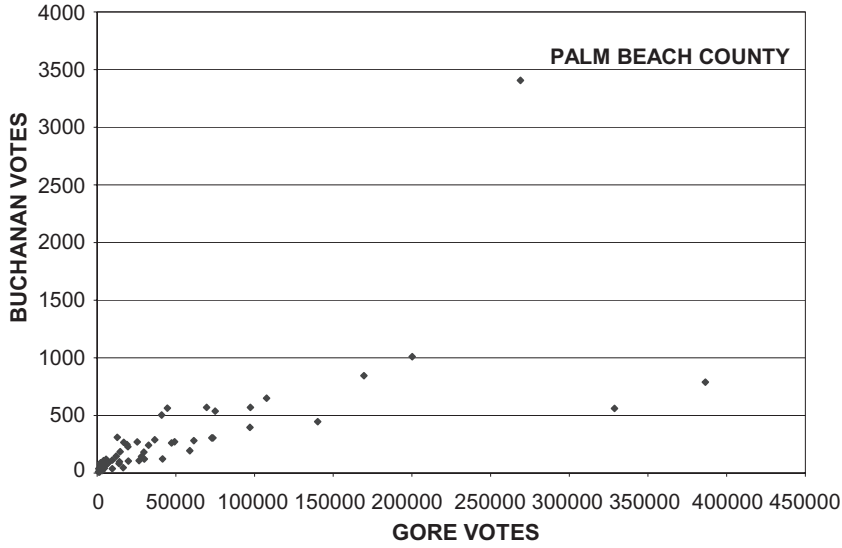


Figure 7.3. County vote totals in Florida; Gore versus Buchanan.

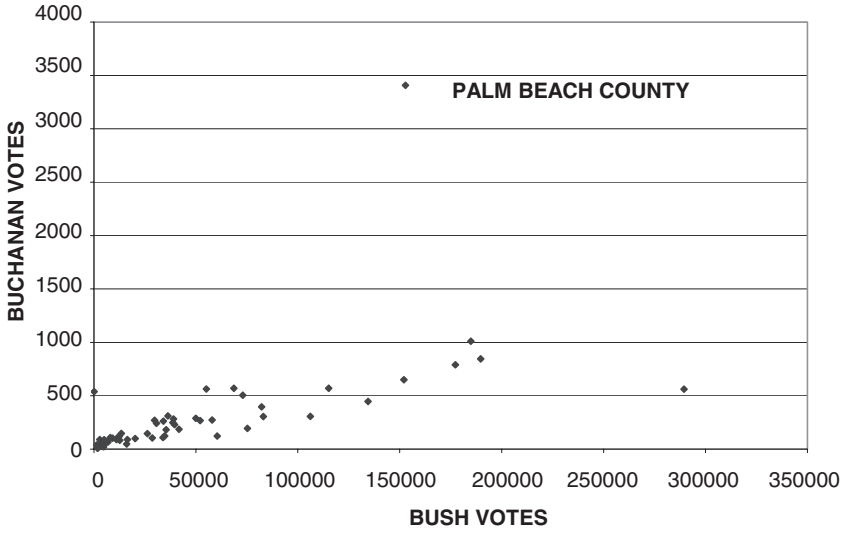


Figure 7.4. County votes totals in Florida: Bush versus Buchanan.

why worry, Buchanan didn't come close to winning the state or even Palm Beach County, and 3407 votes is not a lot. Some were probably intended to go to Buchanan. In fact, we can do a regression analysis by excluding Palm Beach and fitting a regression line to the Gore versus Buchanan data, and predict the number of votes Buchanan would be expected to have given that Gore had 268,945. This will be an estimate of how many out of the 3407 might be legitimate, and the remainder would belong to Gore, Bush, or Nader.

Since we do not know how to split it up, someone playing devil's advocate could say that Bush probably would have gotten something like the proportion that he otherwise got, and the differential would not be enough to swing the election to Gore. But Bush only won by approximately 2000 votes, so if all those votes were for Gore, it could swing the election to Gore's favor.

If we just did a little data mining, we would have seen this anomaly even if we didn't know about the butterfly ballot. But now the butterfly ballot becomes important because (1) Palm Beach is primarily Democratic; and (2) the ballot makes it easy to mistake Buchanan for Gore, but not Buchanan for Bush. Now adding up all the votes from the counties we get:

Gore 2,907,342 Bush 2,828,127.

So Gore would have actually won by close to 80,000 votes. But this does not include the absentee ballots, many of which came from the troops overseas. So the absentee ballots swung the vote to Bush. Although we can estimate how many additional votes we think Gore should have gotten from Palm Beach County there is uncertainty in our estimates, and in the end, we would not be highly confident that Gore won Florida. A better resolution, as I see it, would be to have Florida do a reelection.

Some would argue that it would not be fair to allow registered voters to vote if they hadn't voted on election day. Now those voters could be excluded because we have records of every registered voter who cast a ballot. Of course even this would not replicate the results because some voters could change their mind and some that legitimately voted for Nader or Buchanan could switch to Gore or Bush. There is also the issue of the absentee ballots. There really is no

good resolution to the problem. Also the Bush campaign said that for every county that Gore contests, Bush could also find counties that he would contest. Perhaps the Supreme Court's decision was correct. If there is no good way to correct a mistake you should stay with the results you have. It was the right decision, but their reasoning was wrong.

7.5 MULTIPLE REGRESSION

The differences between simple linear regression and multiple linear regression are

1. One independent predictor variable versus two or more independent predictor variables
2. The bivariate correlation squared is replaced by the multiple correlation coefficient R^2.
3. In multiple linear regression, the correlation matrix replaces the correlation coefficient.
4. Partial correlations can be defined in multiple regression.

Recall that as mentioned in the section on simple linear regression, the form of multiple regression that we are referring to in this section is linear regression, which involves an equation that is linear in the parameters and not necessarily the independent variables. Multiple nonlinear regression is not a topic for this text, but Gallant (1987) is an excellent text that concentrates on nonlinear regression (both simple and multiple).

Not written in the equation above is the additive independent error term denoted by ε. This error term has mean 0 and a variance σ^2 that is constant (does not change as the independent variables change). Under these assumptions, the least squares estimates of the regression parameters are minimum variance unbiased estimators. Also, if ε has a normal distribution, the parameter estimates are maximum likelihood estimates. This property also holds for simple linear regression. The property that the least squares estimates are minimum variance among unbiased estimates is called the Gauss–Markov theorem. A proof can

be found in Draper and Smith (1998, p. 136). The maximum likelihood result can also be found in Draper and Smith (1998, p. 137).

In practice, once we consider multiple regression, there is an issue of how many candidate variables should be included in the regression. Also, some of the variables that we think affect the dependent variable may be related to each other, and so some different selections of subsets of the variables may produce essentially the same predictions. However, in such cases, we have a phenomenon called multicollinearity.

When this happens, it is not a good idea to include all the variables. This is because there may be different sets of values that could be used for the parameters to almost identically fit the data. When this is the case, the estimates are unstable, meaning that slight changes in the data could produce large changes in the regression parameters. Consequently, multicollinearity must be avoided.

There are diagnostics for determining when multicollinearity or near multicollinearity occurs. Belsley et al. (1980) cover this in detail. Another way to avoid multicollinearity is to use one of the many possible procedures for selecting a subset of the independent variables. Among the possibilities are best subset selection (requiring an evaluation of all possible subsets, which can be a lot of possibilities), forward selection (adding variables in one at a time based on an F to enter criterion), backward selection (start with all variables in the model and remove one at a time based on an F to exit criterion), and stepwise selection (at each stage, when a proper subset of the variables is in the regression model F to enter and F to exit, criteria are looked at to decide if the next step should be to add or drop a variable, and which variable to remove [add]).

Other texts on regression cover these methods in detail, but are not important to cover in this text. These methods are all available in most statistical packages that include multiple regression.

We will illustrate multiple regression by again using the Florida 2000 Presidential Election results. We will attempt to predict Buchanan's votes in Palm Beach on the basis of the data from all the other counties, but not simply use Bush's or Gore's or Nader's votes in a simple linear regression. Rather, we will look at a multiple regression model using Bush, Gore, and Nader, and the possible subsets of these. We hope to get a better prediction by using more than one predictor, but we also realize that these vote totals are positively correlated because of the

variability of the size of the counties, and hence all candidate votes increase together because the increase is primarily due to the larger size of county.

Using the software package SAS, we looked at three of the possible multiple regression models. Let N_1 = Gore's total votes in the county, N_2 = Bush's total, N_3 = Nader's total, and M = Buchanan's total votes (the dependent variable). The three models are as follows:

1. $M = \beta_1 N_1 + \beta_2 N_2 + \beta_3 N_3 + \alpha.$
2. $M = \beta_2 N_2 + \beta_3 N_3 + \alpha.$
3. $M = \beta_2 N_2 + \beta_3 N_3 + \beta_{23} N_2 N_3.$

In model 1, the coefficient β_1 was not statistically significant. So model 1 was dispensed with, and only models 2 and 3 remained under consideration. The SAS code used to obtain the results is given in italics as follows:

```
data florida;
    input county$ gore bush buchanan nader;
    cards;
    alachua 47300 34062 262 3215
    baker 2392 5610 73 53
    '
    .
    .
    walton 5637 12176 120 265
    washingtn 2796 4983 88 93
    ;
    run;

    data florid2;
        set florida;
        if county = 'palmbch' then delete;
        nbinter = nader*bush;
```

```
run;

proc reg;
model buchanan = nader bush gore;
run;

proc reg;
model buchanan = nader bush;
run;

proc reg;
model buchanan = nader bush nbinter;
run;
```

The data statement at the beginning creates the SAS data set "florida," with "county" as a character variable, which is indicated by "$" after it in the input statement, and "gore, bush, buchanan and nader" as numerical variables representing the vote totals for that candidate in the given county. The input statement tells how to assign the data that will be read. The cards statement indicates that the data read according to the input statement is to follow. The symbol ";" at the end of the data indicates the completion of reading the data. The statement "run" indicates the finish of the data step.

The next statement is a new data step used to modify the original data set. The set statement means to copy the data set florida into florid2. The "if statement" deletes the line corresponding to Palm Beach county, so that the model will be constructed without including Palm Beach. The statement "nbinter = nadir*bush" creates a variable equal to the product of Nader's total with Bush's total. This variable will be used as the interaction term in the third regression.

The first regression generates model 1, where we can test the significance of Gore's total when included with Bush and Nader. This is part of the standard SAS output for this procedure. The second regression is for the model that includes Bush and Nader's votes only to predict Buchanan's total. The third regression incorporates an interaction term between Bush and Nader.

The output is now presented in bold face, as follows:

Model: MODEL1 (using votes for Nader, Bush, and Gore to predict votes for Buchanan)
Dependent Variable: BUCHANAN

Analysis of Variance

Source	*df*	Sum of Squares	Mean Square	*F*-Value	Prob > *F*
Model	3	2777684.5165	925894.82882	114.601	0.0001
Error	62	500914.34717	8079.26366		
Total	65	3278598.8636			
	Root MSE	89.88472	R^2	0.8472	
	Dep Mean	211.04545	Adj R^2	0.8398	
	C.V.	42.59022			

Parameter Estimates

Variable	*df*	Parameter Estimate	Standard Error	*T* for H_0: Parameter = 0	Prob > |T|
INTERCEP	1	54.757978	14.29169893	3.831	0.0003
NADER	1	0.077460	0.01255278	6.171	0.0001
BUSH	1	0.001795	0.00056335	3.186	0.0023
GORE	1	–0.000641	0.00040706	–1.574	0.1205

Model: MODEL2 (using votes for Nader and Bush to predict votes for Buchanan)
Dependent Variable: BUCHANAN

Analysis of Variance

Source	*df*	Sum of Squares	Mean Square	*F*-Value	Prob > *F*
Model	2	2757655.9253	1378827.9626	166.748	0.0001
Error	63	520942.93834	8268.93553		
Total	65	3278598.8636			
	Root MSE	90.93369	R^2	0.8411	

	Dep Mean	211.04545	Adj R^2	0.8361
	C.V.	43.08725		

Parameter Estimates

Variable	*df*	Parameter Estimate	Standard Error	*T* for H_0: Parameter = 0	Prob > |*T*|
INTERCEP	1	60.155214	14.03642389	4.286	0.0001
NADER	1	0.072387	0.01227393	5.898	0.0001
BUSH	1	0.001220	0.00043382	2.812	0.0066

Model: MODEL3 (using votes for Nader and Bush plus an interaction term Nader*Bush to predict votes for Buchanan)
Dependent Variable: BUCHANAN

Analysis of Variance

Source	*df*	Sum of Squares	Mean Square	*F*-Value	Prob > *F*
Model	3	2811645.8041	937215.26803	124.439	0.0001
Error	62	466953.05955	7531.50096		
Total	65	3278598.8636			
	Root MSE	86.78422	R^2	0.8576	
	Dep Mean	211.04545	Adj R^2	0.8507	
	C.V.	41.12110			

Parameter Estimates

Variable	*df*	Parameter Estimate	Standard Error	*T* for H_0: Parameter = 0	Prob > |*T*|
INTERCEP	1	36.353406	16.7731503	2.261	0.0273
NADER	1	0.098017	0.01512781	6.479	0.0001
BUSH	1	0.001798	0.00046703	3.850	0.0003
NBINTER	1	−0.000000232	0.000000009	−2.677	0.0095

For each model, the value of R^2 describes the percentage of the variance in the votes for Buchanan that can be explained by the predictor variables. This is a measure of the goodness of fit for the model. The adjusted R^2 is slightly smaller and takes into account the fact that the estimates have greater variability in prediction due their correlation in estimation from a common data set.

Both the R^2 and adjusted R^2 are highest in model 3. The R^2 and adjusted R^2 in models 1 and 2 are almost the same. But model 2 is preferable to 1 because Gore's coefficient is not statistically significant. Each model is highly predictive, as indicated by the p-value for the overall F-test, which is 0.0001 in each case.

It appears that model 3 is the best. So we will use model 3 to predict Buchanan's total in Palm Beach County. Here are the predictions that each model would give.

Model 1: 587.710 votes for Buchanan

Model 2: 649.389 votes for Buchanan

Model 3: 659.236 votes for Buchanan

We see that none of the models predict more than 660 votes for Buchanan. Not mentioned in the section on simple linear regression were the simple linear regression models. Without going into the details, which can be found in Chernick and Friis (2003), the prediction for the simple linear regression models ranged from 600 to 1076.

Recall that Palm Beach actually recorded 3407 votes for Buchanan. This is more than three times the amount obtained by any of the predictions. Subtracting the predictions from 3407, we see that Buchanan received between 2331 = 3407 − 1076 and 2807 = 3407 − 600 that we believe were mistakes. Our best estimate is 3407 − 660 = 2747. In any case, if these votes should have gone to Gore, this swing would have a significant impact on the results.

7.6 LOGISTIC REGRESSION

Logistic regression is a method used to predict binary outcomes on the basis of one or more predictor variables. The goals are the same as with linear regression. We attempt to construct a model to best describe the

relationship between a response variable and one or more explanatory variables. The difference that distinguishes logistic regression from other forms of regression is that there are only two possible outcomes, and the job is to estimate the probabilities of the two possible outcomes or the odds of the outcome of interest.

Because we have a dichotomous response variable, we use a very different methodology from the one employed in ordinary linear regression. The text by Hosmer and Lemeshow (2000) is one of the most readable texts devoted to logistic regression and providing instructive examples.

In this section, we provide one simple example along with its solution. For logistic regression, we have predictor variables $X_1, X_2, \ldots X_k$ and are interested in

$E[Y| X_1, X_2, \ldots X_k]$, where Y is the dichotomous outcome variable. This expectation is a probability because Y only takes on the values 0 and 1, and so the conditional expectation is the conditional probability that $Y = 1$. For simplicity, we will go through the notation when there is only one predictor variable X in the model. Then we let $\pi(x) = E[Y|X = x]$. Now because Y is dichotomous and $\pi(x)$ is a probability, it is constrained to belong to (0, 1). The possible values for X may be unconstrained (i.e., may be anywhere between $-\infty$ and $+\infty$)

Then if we want the parameters α and β for the right-hand side of the equation to be of the linear form $\alpha + \beta x$ when $X = x$, then the left-hand side cannot be constrained to a bounded interval such as (0, 1). So we define the logit transformation $g(x) = \ln[\pi(x)/\{1 - \pi(x)\}]$. First we note that the transformation $\omega(x) = \pi(x)/\{1 - \pi(x)\}$ takes values from (0, 1) to (0, ∞). Then applying the logarithm transforms, it takes values from (0, ∞) to ($-\infty$, ∞).

So the logistic regression model is $g(x) = \alpha + \beta x$. The observed values of g(X) will have an additive random error component. We can express this on the probability scale by inverting the transformation to get $\pi(x) = \exp(\alpha + \beta x)/[1 + \exp(\alpha + \beta x)]$. To see this requires a little basic algebra as follows: $\exp(\ln[x]) = x$ since the natural logarithm and the exponential function are inverse functions of each other. Now $\exp(g[x]) = \exp((\alpha + \beta x)) = \exp\{\ln[\pi(x)/(1 - \pi[x])]\} = \pi(x)/\{1 - \pi(x)\}$. So we now solve the equation $\pi(x)/\{1 - \pi[x]) = \exp((\alpha + \beta x))$ for $\pi(x)$. So multiplying both sides by $1 - \pi(x)$, we get $\pi(x) = \{1 - \pi(x)\}$ $\exp((\alpha + \beta x))$. Distributing $\exp((\alpha + \beta x))$ on the right-hand side gives us $\pi(x) = \exp((\alpha + \beta x)) - \pi(x)\exp((\alpha + \beta x))$, and then by adding $\pi(x)$

$\exp((\alpha + \beta x))$ to both sides, we have $\pi(x) + \pi(x)\exp((\alpha + \beta x)) = \exp((\alpha + \beta x))$. Now, we factor out $\pi(x)$ from the left-hand side of the equation and get $\pi(x)[1 + \exp((\alpha + \beta x))] = \exp((\alpha + \beta x))$. Finally, we divide both sides by $[1 + \exp((\alpha + \beta x))]$ and get

$$\pi(x) = \exp((\alpha + \beta x))/[1 + \exp((\alpha + \beta x))].$$

Our objective in logistic regression is to estimate the parameters α and β to provide the "best fit" in some statistical sense. Now, in ordinary linear regression, when the error terms are normally distributed with mean equal to zero and a constant variance, least squares, and maximum likelihood are the same. In the logistic regression model, however, maximum likelihood and least squares are not equivalent because the error term is not normally distributed. Now we proceed to see where maximizing the likelihood will take us.

Suppose the data consists of the pair (x_i, y_i) for $i = 1, 2, \ldots, n$. The x_is are the observed values for X, and the y_is are the observed Y-values. Remember that the y_is are dichotomous and only can be 0 or 1. The likelihood function is then

$$L(x_1, y_1, x_2, y_2, \ldots x_n, y_n) =$$
$$\pi(x_1)^{y_1}[1-\pi(x_1)]^{1-y_1}\,\pi(x_2)^{y_2}[1-\pi(x_2)]^{1-y_2}\ldots\pi(x_n)^{y_n}[1-\pi(x_n)]^{1-y_n}.$$

The solution is obtained by taking partial derivatives with respect to α and β to obtain the two equations $\Sigma[y_i - \pi(x_i)] = 0$ and $\Sigma x_i[y_i - \pi(x_i)] = 0$. The parameters α and β enter these equations through the relationship $\pi(x_i) = \exp((\alpha + \beta x_i))/[1 + \exp((\alpha + \beta x_i))]$. These equations must be solved numerically since they are not linear in α and β. It is also not obvious that the solution is unique.

For the fine details, see Hosmer and Lemeshow (2000) or Hilbe (2009). The logistic regression model is a special case of the generalized linear model due to Nelder. The generalized linear model is linear in the regression parameters but replaces the response Y with a function called the link function. In the case of logistic regression, the logit function is the link function. If you want to learn more about generalized linear model, including other examples, consult McCullagh and Nelder (1989).

The data in Table 7.4 was adapted from Campbell and Machin (1999) by Chernick and Friis (2003), and is used here to illustrate a

Table 7.4
Hemoglobin Level (Hb), Packed Cell Volume (PCV), Age, and Menopausal Status for 20 Women*

Subject number	Hb (g/dL)	PCV (%)	Age (years)	Menopause (0 = no, 1 = yes)
1	11.1	35	20	0
2	10.7	45	22	0
3	12.4	47	25	0
4	14.0	50	28	0
5	13.1	31	28	0
6	10.5	30	31	0
7	9.6	25	32	0
8	12.5	33	35	0
9	13.5	35	38	0
10	13.9	40	40	0
11	15.1	45	45	1
12	13.9	47	49	0
13	16.2	49	54	1
14	16.3	42	55	1
15	16.8	40	57	1
16	17.1	50	60	1
17	16.6	46	62	1
18	16.9	55	63	1
19	15.7	42	65	1
20	16.5	46	67	1

*From Chernick and Friis (2003, p. 286, table 12.10).

logistic regression analysis. The purpose of this data is to fit a logistic regression model to see if the odds of becoming anemic differ for women under 30 years of age compared with women over 30. Female patients with hemoglobin levels below 12 g/dL were classified as anemic.

We see from the data that two out of the five women under 30 were anemic, while only 2 of the 15 women over 30 were anemic. None of the women experiencing menopause were anemic. It is because during menstruation, younger, nonmenopausal women have blood and hemoglobin loss, while postmenopausal women would not. So it was hypoth-

esized that the nonmenopausal women would be at greater risk for anemia than the postmenopausal women. By risk we mean the probability of being anemic given whether or not you are postmenopausal. So we are saying that we expect that just conditioning on nonmenopausal or postmenopausal, we would expect the conditional probability to be higher for nonmenopausal women.

Another common way to look at the difference in risks such as anemia when comparing two groups like this is the odds ratio, say O_1/O_2, where $O_1 = \pi_1/(1 - \pi_1)$ and $O_2 = \pi_2/(1 - \pi_2)$. O_1 and O_2 are called the odds—say 1 denotes nonmenopausal women and 2 denotes postmenopausal women. Relative risk is π_1/π_2. So when π_1 and π_2 are small, $1 - \pi_1$ and $1 - \pi_2$ are close to 1, and the odds ratio and relative risk are nearly the same. But when they are not small, the two measures can differ. Lachin (2000) is an excellent text on biostatistics that emphasizes relative risks and odds ratios. So it is a great source to use to clear up any confusion you might have.

Campbell and Machin only used the age dichotomized at 30, and estimated that the regression parameter for age group was 1.4663, with a standard error of 1.1875. The Wald test is the analog in logistic regression to the *t*-test for significance of the parameter. The value of the Wald statistic was 1.5246, which translates to a *p*-value of 0.2169. So at least for the two age groups, there was not a statistically significant difference. However, age could still be an important factor if the cut point should be different or if age is left on a continuous scale. Also, it may be that there is too much patient to patient variability for 20 women to be an adequate sample size. Also, age is correlated with menopause. So it may be that age would be far more important if the dichotomous menopause variable were not included in the model.

The result of performing the logistic regression using the actual ages, as was done by Chernick and Friis (2003), gives a coefficient of –0.2077, with a standard deviation of 0.1223, indicating a possible decrease in the risk of anemia with increasing age. The Wald statistic is 2.8837, corresponding to a *p*-value of 0.0895. This is significant at the 10% level, but not at the 5% level. It could well be that we would find greater significance with a larger sample of women. The choice of 30 to dichotomize was probably a bad choice. We note that 6 of the 15 women over 30 were not menopausal, and the coefficient of 1.4663 was in the opposite direction of what the alternative hypothesis would suggest.

7.7 EXERCISES

1. Define the following terms:
   (a) Association
   (b) The Pearson correlation coefficient
   (c) Simple linear regression
   (d) Multiple linear regression
   (e) Nonlinear regression
   (f) Scatter plot
   (g) Slope of the regression line in simple linear regression

2. What assumptions are needed for the Pearson correlation coefficient to be a meaningful measure of the relationship between two variables?

3. What is the mathematical relationship between the correlation coefficient and the slope of the simple linear regression line? Can the slope be negative and the correlation be positive? If the correlation is zero, what is the value of the slope?

4. Regarding outliers:
   (a) How would you define an outlier?
   (b) Does an outlier always imply an error in the data?
   (c) Give an example of an outlier that represented an error in the data.
   (d) Give an example where the outlier is more important to the research than the other observations.

5. What is logistic regression? How is it different from ordinary linear regression?

6. How does multiple linear regression differ from simple linear regression?

7. What is the definition of the multiple correlation coefficient R^2?

8. How is R^2 useful in evaluating the goodness of a model?

9. What is the equivalent to R^2 in simple linear regression?

10. What is multicollinearity? Why does it pose problems estimating regression parameters?

11. What is stepwise regression? Why is it used?

12. Refer to Table 7.5. A psychiatric epidemiologist studied information he collected on the anxiety and depression levels for 11 subjects. Produce a scatter diagram for anxiety score on the x-axis and depression score on the y-axis.

Table 7.5
Anxiety and Depression Scores for 11 Subjects

Subject ID	Anxiety score	Depression score
1	24	14
2	9	5
3	25	16
4	26	17
5	35	22
6	17	8
7	49	37
8	39	41
9	8	6
10	34	28
11	28	33

13. Again referring to Table 7.5, calculate the following: (a) mean anxiety score, (b) mean depression score, (c) standard deviations for depression and anxiety scores, and (d) Pearson correlation between anxiety score and depression score.

14. An experiment was conducted to study the effect of increasing the dosage of a certain barbiturate. Three readings were recorded at each dose. Refer to Table 7.6.

(a) Plot the scatter diagram (scatter plot)

(b) Determine by least squares the simple linear regression line relating dosage X to sleeping time Y.

(c) Provide a 95% two-sided confidence interval for the slope.

(d) Test that there is no linear relationship at the 0.05 level.

15. Fit the model and predict the sleeping time for a 12 μM/kg?>

(a) The Pearson product moment correlation coefficient

(b) A test result as to whether or not the correlation coefficient is significantly different from 0 at the 0.05 significance level.

(c) The same test as (b) but at the 0.01 significance level.

Table 7.6
Dosage Versus Sleeping Time

Sleeping time Y (hours)	Dosage X (μM/kg)
4	3
6	3
5	3
9	10
8	10
7	10
13	15
11	15
9	15
$\Sigma Y = 72$	$\Sigma X = 84$
$\Sigma Y^2 = 642$	$\Sigma X^2 = 1002$
$\Sigma XY = 780$	

16. Table 7.7 shows the Math Olympiad scores for 33 math students at Churchville Elementary School in Churchville, Pennsylvania in 2002. We are interested in how well the first score (test 2) predicts the students next score (test 3). Plot the scatter diagram for this data. Compute the Pearson correlation coefficient and the square of the correlation coefficient. Calculate the mean score for test 2 and the mean score for test 3.

17. Math Olympiad and regression toward the mean. The least squares regression equation for exam score 3, y as a function of exam score 2, x, is:

$$y = 0.4986x + 1.0943.$$

The possible scores for exam 2 are 0, 1, 2, 3, 4, and 5. For each possible score, use the above regression equation to predict the score for exam 3 for a student who got that score on exam 2. Fill in the predicted scores for Table 7.8.

18. Having computed the average scores on exams 2 and 3, you know that in both cases, the average is somewhere between 2 and 3. So scores on exam 2 of 0, 1, and 2 are below average, and scores of 3, 4, and 5 are above average. Compare the scores on exam 2 with their predicted score for exam 2 scores of 0, 1, and 2. Are the predicted scores lower or higher than the exam 2 score? Now, for scores 3, 4, and 5, are the predicted scores

Table 7.7
Math Olympiad Scores for Churchville Students

Student number	Score on exam 2	Score on exam 3
1	5	4
2	4	4
3	3	1
4	1	3
5	4	4
6	1	1
7	2	3
8	2	2
9	4	4
10	4	3
11	3	3
12	5	5
13	0	1
14	3	1
15	3	3
16	3	2
17	3	2
18	1	3
19	3	2
20	2	3
21	3	2
22	0	2
23	3	2
24	3	2
25	3	2
26	3	2
27	2	0
28	1	2
29	0	1
39	2	2
31	1	1
32	0	1
33	1	2

Table 7.8
Predicting Student's Olympiad 3 Exam Score Based on Olympiad 2 Exam Score

Exam 2 score	Predicted exam 3 score
0	
1	
2	
3	
4	
5	

higher or lower than the exam 2 score? What changes when we move from below the average to above the average? The result is a mathematical property called regression toward the mean. This occurs in any regression problem. Some people thought it was a tendency to move toward mediocrity. But that is a fallacy called the regression fallacy.

CHAPTER 8

Contingency Tables

Contingency tables are cross-tabulations of one categorical variable versus another. They are used to test hypotheses about association between the variables or differences among proportions. We will see that the chi-square test is an approximate test for association when the data set is large enough. Large enough means that each cell in the table is filled with a reasonable number of counts (5 as a minimum is a good rule of thumb).

On the other hand, Fisher's exact test and its generalizations achieve the exact significance level, but require an added assumption that the row sums and the column sums are fixed at their observed levels when comparing the existing table with other possible arrangement that occur under the null hypothesis of no association (sometimes referred to as independence).

The simplest table is the 2×2, where each variable can have only two categories. However, in general, we have the $R \times C$ table where R is the number of categories for the row variable, and C is the number of categories for the column.

8.1 2 × 2 TABLES AND CHI-SQUARE

For an example of a 2×2 table, consider the case: Let us consider whether or not there is a difference in the preference for western

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

Table 8.1
Preference for Type of Medical Care by Gender

	Type of medical care preference		
Gender	Western medicine	Alternative medicine	Row total
Men	49 (39.5) A	51 (60.5) B	100 (100) A + B
Women	30 (39.5) C	70 (60.5) D	100 (100) C + D
Column total	79 (79) A + C	121 (121) B + D	200 (200) N = A + B + C + D Grand total

Expected frequencies are shown in parentheses.

medicine versus alternative medicine between men and women. Suppose we have a questionnaire where we ask the respondents if they prefer western medicine or alternative medicine, and we tell them that they must choose one over the other. We also keep track of the gender of all the respondents. Table 8.1 illustrates the 2×2 table that corresponds to the results of such a survey.

The table above presents the observed frequencies from a survey of men and women. Out of 100 men, 49 favored western medicine, and 51 favored alternative medicine. Out of 100 women, 30 favored western medicine, and 70 favored alternative medicine. The expected frequencies represent the total (not necessarily an integer because it is an average) in each of the four cells under the assumption of no association. The cell counts are represented algebraically as A for men favoring western medicine, B for men favoring alternative medicine, and C for women favoring western medicine, and D for women favoring alternative medicine.

The expected value under the null hypothesis of no difference or independence for the cell containing A, we shall denote as $E(A)$, and similarly for B, C, and D. $E(A)$ is the product of the two proportions involving A multiplied by N (the grand total). So $E(A) = N[(A + B)/N][(A + C)/N] = N(A + B)(A + C)/N^2 = (A + B)(\mathrm{A} + \mathrm{C})/N$. By the same argument, $\mathrm{E}\ (\mathrm{B}) = (\mathrm{B} + \mathrm{A})(B + D)/N$, and $E(C) = (C + D)(C + A)/N$, and $E(D) = (D + C)(D + B)/N$. Applying these formula in the example in Table 8.1, the values in parentheses are obtained using $A = 49$, $B = 51$, $C = 30$, and $D = 70$.

In general, Karl Pearson's chi-square test determines the goodness of fit of the observed data to the expected value under a model. This is a general asymptotic result that applies to a wide variety of problems, including testing for independence between two variables, as in the current example. The asymptotic distribution in the general case of an $R \times C$ contingency table is the central chi-square distribution with $(R-1)(C-1)$ degrees of freedom. So in the 2×2 table, $R = 2$ and $C = 2$, and hence the degrees of freedom is 1. The chi-square statistic is given by the formula

$$\chi^2 = \sum (O_i - E_i)^2 / E_i,$$

where O_i is the observed total in for cell I, and E_i is the expected total for cell i.

$$\chi^2 = (49-39.5)^2/39.5 + (30-39.5)^2/39.5 + (51-60.5)^2/60.5 + (70-60.5)^2/60.5 = 7.55.$$

For $\alpha = 0.05$, the critical value for a chi-square random variable with 1 degree of freedom is 3.84. So, since $7.55 >> 3.84$, the choice of type of medical care does differ between men and women.

8.2 SIMPSON'S PARADOX IN THE 2 × 2 TABLE

Sometimes, as for example in a meta-analysis, it may be reasonable to combine results from two or more experiments, each of which produces a 2×2 table. We simple cumulate for the corresponding cells in each table the sum of the counts over all the tables.

However, this creates an apparent paradox, known as Simpson's paradox. Basically, Simpson's paradox occurs when we see an apparent association in each of the individual tables, but not in the combined table or the association reverses!

To understand this better, we take the following example from Lloyd (1999, pp. 153–154). For this analysis (a fictitious example used to illustrate the issue), a new cancer treatment is given, experimentally, to the patients in hospital A, who have been categorized as either terminal or nonterminal.

Before we analyze the patients based on their terminal/nonterminal category, we naively think that we can see a difference in survival

simply based on the treatment without regard to their status. Hospital A follows the patients, and records the results after 2 years of follow-up, hoping to see better survival under the new treatment. The results are given in Table 8.2.

By examining the table, the results seem clear. There were 221 patients receiving the new treatment, and 221 the old. But the old treatment appears better because 177 survived, compared with only 117 under the new treatment. The survival rate for patients under the old treatment is 80.1%, and only 52.9% for the new treatment. This is puzzling to the investigators, because they thought that the new treatment was better!

The investigators think about it, and now they say to themselves "maybe the greater survival in the old treatment group could be due to an imbalance of terminally ill patients." Since terminally ill patients are likely to die regardless of treatment, it is possible that the observed difference is explained, because many more patients were terminally ill in the new treatment group.

They decide to split the data into two groups and generate two 2×2 tables. Table 8.3 is the table for the terminal patients, and Table 8.4 for the nonterminal patients.

Table 8.2
All Patients: Survival Versus Treatment

Treatment	Survived 2 years	Died within 2 years	Total
New	117	104	221
Old	177	44	221
Total	294	148	442

Table 8.3
Terminal Patients Only: Survival Versus Treatment

Treatment	Survived 2 years	Died within 2 years	Total
New	17	101	118
Old	2	36	38
Total	19	137	156

Table 8.4
Nonterminal Patients Only: Survival Versus Treatment

Treatment	Survived 2 years	Died within 2 years	Total
New	100	3	103
Old	175	8	183
Total	275	11	286

Here, the picture is quite different. The survival rate is much lower in the terminal patients (as we should expect), and many more terminal patients got the new treatment compared with the old 118 versus 38. In the terminal group, the survival rate for those getting the new treatment is 14.4% compared with only 5.2% for patients on the old treatment. For the nonterminal patients, the new treatment group has a survival rate of 97.1%, slightly higher than the 95.6% for the old treatment group.

So here is the paradox. The new treatment is apparently better in both subgroup analyses (although probably not statistically significantly better for the nonterminal patients). So based on the subgroup analysis, the new treatment might get regulatory approval. However, if we only had the combined results, we would be convinced that the new treatment is inferior to the old treatment.

So why does Simpson's paradox occur and how do we resolve it? First, notice the imbalance between the subgroups. Only 156 patients were terminal compared with 286 in the nonterminal group. Also, in the terminal group, there were far more patients getting the new treatment (118), while only 38 patients got the old treatment. These imbalances mask the benefit of the new treatment when the data is combined. Also notice that the survival rates are so drastically different for terminal and nonterminal patients.

A total of 275 out of 286 nonterminal patients survived (96.15%), whereas only 19 out of 156 survived among the terminal patients only (12.18%). So the combination of the two groups makes no sense. It is like adding apples with oranges. In reality, the combined table is meaningless and presents a distorted picture.

In such cases, we would not combine these two studies in a meta-analysis, as they are estimating radically different success probabilities.

So the investigators were right in thinking that the status of the disease had a confounding effect on the result in the combined table, and the analysis should have done only on the separate groups. Thus, Simpson's paradox is not a true paradox, but rather a misunderstanding about the proportions in the tables.

Another way to deal with this to avoid the occurrence of Simpson's paradox would be stratification. Make sure that there are sufficiently large numbers of terminal and nonterminal patients. Also through randomization we can make sure that an equal number in each group get the new treatment as get the old. The stratification can force any ratio of nonterminal to terminal; a 1 to 1 balance is not necessary, but an approach that creates a near 1 to 1 balance will do the job.

8.3 THE GENERAL $R \times C$ TABLE

The $R \times C$ table is a generalization of the 2×2, where the column variable can have two or more categories denoted by C, and the row variable can also have two or more categories denoted by R. The chi-square statistic has the same form, but as mentioned earlier, the asymptotic distribution under the null hypothesis is a central chi-square, with $(R - 1)(C - 1)$ degrees of freedom, compared with 1 for the 2×2 table.

To illustrate, we will look at an example of a 3×3 table. The data is a sample taken from a registry of women with breast cancer. The research problem is to see if there is a relationship with the ethnicity of the patient and the stage of the cancer. The three ethnicities considered are Caucasian, African American, and Asian. The three stages are called *in situ*, local, and distant. The data in the 3×3 table is given next in Table 8.5.

The chi-square statistic is again obtained by taking the observed minus expected squared divide by the expected in each of the nine cells and summing them together. We see that the Asians seem to be very different from their expected value under the independence model. Also, the in situ stage has all ethnicities, with totals very different from their expected values. The chi-square statistic is 552.0993. For the chi-square with degrees of freedom = $(3 - 1)(3 - 1) = 2 \times 2 = 4$. A value of 16.266 corresponds to a p-value of 0.001. So the p-value for a chi-square value of 552.0993.

Thus far, all the tables we have studied had plenty of counts in each cell. So the chi-square test is highly appropriate and gives results very

Table 8.5
Association Between Ethnicity and Breast Cancer Stage From a Registry Sample*

	Stage of breast cancer			
Ethnicity	*In situ*	Local	Distant	Total
Caucasian	124 (232.38)	761 (663,91)	669 (657.81)	1554
African American	36 (83.85)	224 (239.67)	301 (237.47)	561
Asian	221 (64.87)	104 (185.42)	109 (183.71)	434
Total	381	1089	1079	2549

*Note: Count with expected count in parentheses.

close to other asymptotic and exact nonparametric tests. However, when some cells are sparse (i.e., 0–4 counts in those cells), Fisher's exact test for 2×2 tables and its generalization to the R × C table is a better choice. That is the topic of the next section.

8.4 FISHER'S EXACT TEST

In contingency tables, the counts in each cell may be totally random, and hence the row and column totals are not restricted. However, there are cases where the row totals and column totals (called marginal totals or margins) are fixed in advance. In such cases, it makes sense to consider as the sample space all possible tables that yield the same totals for each row and each column. The distribution of such tables under the null hypothesis of independence is known to be a hypergeometric distribution.

So one could ask under the null hypothesis is our observed table likely to occur or not based on the known hypergeometric distribution. This idea goes back to Fisher (1935) for 2×2 tables, and can easily be generalized to any $R \times C$ table. This idea has been applied even when the rows and column need not be the same as in the observed table, with the argument being that it still makes sense to condition on the given values for the row and column totals. In fact, the Fisher exact test gives nearly the same results as the chi-square when the chi-square is appropriate as an approximation, and the chi-square test does not

Table 8.6
Basic 2 × 2 Contingency Table

	Column 1	Column 2	Row totals
Row 1	x	$r - x$	r
Row 2	$c - x$	$N - r - c + x$	$N - r$
Column totals	c	$N - c$	N

involve any conditioning on the marginal totals. As a historical note, Conover (1999) points out that the same idea appeared in Irwin (1935) and Yates (1934). So it is not clear whether or not Fisher should be credited as the originator.

Now, let us see, in the case of the 2 × 2 table how the hypergeometric distribution occurs under the null hypothesis. Let N be the total number of observations. The totals for the two rows are r and $N - r$ for our data set. Similarly we have column totals of c and $N - c$. Table 8.6 shows the complete picture.

Because the values r, c, and N are fixed in advance, the only random variable remaining is x, by our notation the entry in the cell for row 1 and column 1. Now, x can vary from 0 to the minimum of c and r. This restriction happens because the sum of the two columns in row 1 must be r, and the sum of the two rows in column 1 must be c.

As we provide different values for x, we get different 2 × 2 contingency tables. So the possible values of x determine all the possible 2 × 2 tables with the margins fixed. For the null hypothesis of independence, the probability of p_1 that an observation falls in row 1 is equal to the probability p_2 that an observation falls in row 2 regardless of what column it is in. The same argument can be made for the columns. The random variable $T = x$ has the hypergeometric distribution that is for $x = 0, 1, \ldots, \min(r, c)$ $P(T = x) = C(r, x)C(N - r, c - x)/C(N, c)$ and $P(T = x) = 0$ for any other value of x where $C(n, m) = n!/[(n - m)!\, m!]$ for any $m \leq n$.

A one-sided p-value for the test for independence in a 2 × 2 table is calculated as follows:

1. Find all 2 × 2 tables with the same row and column totals of the observed table where cell (1, 1), row 1 and column 1, has a total less than or equal to x from the observed table, and a probability less than or equal to the observed probability.

2. Use the above formula for the hypergeometric distribution to calculate the probability of these tables under the null hypothesis.
3. Sum the probabilities for all such tables.

This gives a one-sided p-value. To get a p-value for the other side, sum all the probabilities of tables where cell (1, 1) has values greater than or equal to x, and probability lower than the observed probability. The two-sided p-value is just the sum of the two one-sided p-values minus the observed probability, because the sum would count the observed probability twice.

We now illustrate an example of testing skills at detecting order. Now suppose as hypothesized by Agresti (1990, p. 61) that an experiment was conducted to test the null hypothesis of random guesses versus the alternative of skill in detecting order. Agresti's well-known text was revised (see Agresti [2002]). The patient is given a medicine and water. She is told that four of the cups have the water poured first, and four had the medicine place in the cup first, and then the water was added. The cups are numbered 1–8. Because she knows that four are water first, and four are medicine first, the table is constrained to have both row margins and both column margins totaling four. The table looks like Table 8.7.

Here, the conditioning is uncontroversial, because the experimenter told the patient the row and column constraints. There are now only five possible tables: (1) perfect guessing $x = 4$, (2) two mistakes out of eight guesses when $x = 1$, (3) four mistakes out of eight guesses when $x = 2$, (4) six mistakes out of eight guesses when $x = 3$, and (5) eight mistakes out of eight guesses when $x = 4$. Clearly, the more mistakes that there are, the further we are from the alternative hypothesis. The

Table 8.7
Patient Taking Medicine Experiment: Possible 2×2 Tables

Placed in cup first	Patient guesses medicine	Patient guesses water	Row totals
Medicine	x	$4 - x$	4
Water	$4 - x$	x	4
Column totals	4	4	8

Table 8.8
Patient Taking Medicine Experiment: Observed Table

Placed in cup first	Patient guesses medicine	Patient guesses water	Row totals
Medicine	3	1	4
Water	1	3	4
Column totals	4	4	8

null hypothesis expects four mistakes. This test is naturally one sided, since the patient claims skill, meaning she get more right than four out of eight, and less than four simply says that she is no better or worse at guessing than by chance. We do the experiment and get a result where she guesses 1 "water first" wrong and 1 "medicine first" wrong. Her table looks like Table 8.8.

There is one table more extreme than the observed, and that is the perfect guess table with $x = 4$. The *p-value* for this experiment is the sum of the probabilities that $x = 3$ and $x = 4$. So $p = C(4, 3)C(4, 1)/C(8, 4) = (4!/[3!\ 1!])(4!/[1!\ 3!])/(8!/[4!\ 4!]) = 4(4)\ 4!/8\ 7\ 6\ 5) = 16/70 = 8/35 = 0.229$. Perfect guessing has probability $1/70 = 0.0142$. So the *p*-value for the experiment is $0.229 + 0.014 = 0.243$. This is not statistically significant. Only a perfect score would have been significant for a sample size of eight, with four of each mixture.

Fisher's exact test is an example of nonparametric procedures that go by various names: permutation tests, randomization tests or rerandomization tests. For a modern treatment of these procedures, see Good (2000). In Fisher's original book (Fisher 1935), the exact problem is presented except that instead of a patient taking medicine, it is a lady tasting tea. The experiment itself is covered in detail in Salsburg (2001), a beautiful story of the history of the development of statistics through the twentieth entury.

8.5 CORRELATED PROPORTIONS AND MCNEMAR'S TEST

When considering the paired *t*-test, we recognized the advantage of reducing variance through the use of correlated observations. Since we

were looking at mean differences, it was positive correlation that helped. McNemar's test is used for correlated categorical data. We can use it to compare proportions when the data are correlated. Even if there are more than two categories for the variables, McNemar's test can be used if there is a way to pair the observations from the groups.

As an example, suppose that we have subjects who are attempting to quit smoking. We want to know which technique is more effective: a nicotine patch or group counseling. So we take 300 subjects who get the nicotine patch and compare them to 300 subjects who get the counseling. We pair the subjects by characteristics that we think could also affect successful quitting and pair the subjects accordingly.

For example, sex, age, level of smoking, and number of years you have smoked may affect the difficulty for quitting. So we match subjects on these factors as much as possible. Heavy smokers who are women and have smoked for several years would be matched with other women who are heavy smokers and have smoked for a long time. We denote by 0 as a failure to quit, where quitting is determined by not smoking a cigarette for 1 year after the treatment. We denote by 1 a success at quitting.

The possible outcomes for the pairs are (0, 0), (0, 1), (1, 0), and (1, 1). We will let the first coordinate correspond to the subject who receives the nicotine patch, and the second coordinate his match who gets counseling instead. The pairs (0, 0) and (1, 1) are called concordant pairs because the subjects had the same outcome. The pair (0, 0) means they both failed to quit, while the pair (1, 1) means that they both were able to quit. The other pairs (0, 1) and (1, 0) are called discordant pairs because the matched subjects had opposite outcomes.

The concordant pairs provide information indicating possible positive correlation between members of the pair without providing information about the difference between proportions. Similarly, the discordant observations are indicative of negative correlation between the members of the pair. The number of 1s and 0s in each group then provides the information regarding the proportions.

What we mean is that if I only tell you a pair is concordant and not whether it is (0, 0) or (1, 1), you know that they are correlated but do not know the actual outcome for either subject in the pair. The same idea goes for the discordant observations. Although you know the results are opposite, indicating a possible negative correlation, and we know we have added a success and a failure to the total, we do not

Table 8.9
Outcomes for Pairs of Subjects Attempting to Stop Smoking

	Counseling failure	Counseling success	Nicotine patch total
Nicotine patch failure	$N = 143$ (0, 0)	$R = 48$ (0, 1)	$N + R = 191$
Nicotine patch success	$Y = 92$ (1, 0)	$Z = 17$ (1, 1)	$Y + Z = 109$
Counseling total	$N + Y = 235$	$R + Z = 65$	$N + Y + R + Z = 300$

know if it is added to the nicotine patch group or the group counseling group.

Table 8.9 is a 2×2 table that counts the number of each of the four possible pairs for this matched experiment.

Under the null hypothesis that success does not depend on the treatment, we would expect the discordant observations (0, 1) and (1, 0) to be approximately equal. So the expected total given the discordant total $R + Y$ would be $(R + Y)/2$. McNemar's test statistic is $T = (R - [R + Y]/2)^2/\{(R + Y)/2\} + (Y - [R + Y]/2)^2/\{(R + Y)/2\}$. This is just like the chi-square statistic summing "the observed minus expected squared divided by expected." After some algebra we see for the 2×2 table, this simplifies to $(R - Y)^2/[2(R + Y)]$, and so the test is equivalent to testing for large values of $W = (R - Y)^2/(R + Y)$. For a more detailed account, see Conover (1999, p. 166). In this example, we get $(92 - 48)^2/(92 + 48) = (44)^2/140 = 0.1936/140 = 13.82$. Now under the null hypothesis, T is asymptotically chi-square with 1 degree of freedom. $T = W/2 = 6.91$. Consulting the chi-square table, we see that the p-value is slightly less than 0.01.

For more than two categories in each group the idea of concordant and discordant pairs extends, and McNemar's test can be applied to an $R \times 2$ table.

8.6 RELATIVE RISK AND ODDS RATIO

Relative risks and odds ratios are important in medical research and are common in epidemiology studies as well. For a detailed discussion of these concepts, see Lachin (2000), and, from the epidemiologists

Table 8.10
Assessment of Relative Risk in a 2 × 2 Table

	Outcome		
Factor	Present	Absent	Total
Yes	a	b	$a + b$
No	c	d	$c + d$
RR (relative risk) = $[a/(a + b])/[c/(c + d)] = a(c + d)/[c(a + b)]$			

perspective, Friis and Sellers (1999). Both concepts are germane to contingency tables where proportions are considered. Relative risk is used in cohort studies. Suppose we have a population of subjects who have different exposure to risk factors for a particular disease.

We have a record of their medical history and information on past exposure. Then we follow these patients to see if they get the disease. The occurrences of new cases of the disease, in this population, are compared between groups with different exposure to see if the exposure affects the incidence rates. Table 8.10 shows the concept of relative risk in a 2 × 2 table.

The relative risk can vary from 0 to ∞. It is the ratio of the proportion of the cases where the disease occurs, given the factor is present to the proportion of cases where the disease occurs, given the factor is absent. It shows how many times the risk increases or decreases according to whether or not the factor is present. So a relative risk of 2 for lung cancer when the subject is a smoker compared with a nonsmoker is interpreted as smoking doubles your risk of getting lung cancer. Table 8.11 shows a 2 × 2 table for lung cancer with smoking as a factor from a cohort study.

So we see from Table 8.11 that the relative risk is 6.53, which means that you are over 6.5 times more likely to get lung cancer if you smoke than if you don't smoke. So a relative risk greater than 1 in this case means that the factor increases your chances of getting the disease.

Consequently we often test that the relative risk is different from 1 or perhaps greater than 1 if we anticipate a negative effect from the factor. In Chapter 10, when we cover survival analysis, we will see how the survival models allow us to obtain tests of hypotheses on the relative risk or construct confidence intervals for it. This is often done when the Cox proportional hazard model is fit to the survival data. In such a

Table 8.11
Assessment of Relative Risk in a 2 × 2 Table from Lung Cancer Cohort Study

	Lung cancer		
Smokers	Present	Absent	Total
Yes	98	202	300
No	35	665	700
RR (relative risk) = $a(c + d)/[c(a + b)]$ = 98(700)/[35(300)] = 6.53			

Table 8.12
2 × 2 Assessment of an Odds Ratio

Factor	Cases	Controls
Yes	a	B
No	c	D
Total	$a + c$	$b + d$
OR (odds ratio) = $(a/c)/(b/d) = ad/(bc)$		

Table 8.13
2 × 2 Assessment of an Odds Ratio

Smoker	Lung cancer cases	Controls
Yes	18	15
No	9	12
Total	29	27
OR (odds ratio) = 18(12)/[(15)9] = 1.6		

situation, the relative risk is the same as the hazard ratio. This will be explained in Chapter 10.

A closely related concept is the odds ratio. Suppose we want to do a case-control study to look at the risk that smoking cases with respect to lung cancer rather than a cohort study. For such a study, we would have a 2 × 2 table of the form given as Table 8.12.

Now Table 8.13 shows the calculation for a particular small case-control study.

In this study, we see that the interpretation of the odds is that the odds are 1.6 times greater that a nonsmoker who is otherwise similar to the smoker. Odds represent the ratio of the probability of occurrence to the probability that the disease does not occur, and the odds ratio is just a ratio of the odds for cases divided by the odds for controls. This is a little different from relative risk but conveys a similar message. When the risk (probability of occurrence of the disease) is low, the odds ratio provides a good approximation to the relative risk.

We looked at point estimates for relative risk and odds ratios, but we did not show you how to get confidence intervals. Confidence intervals for these estimates are found in Lachin (2000, p. 24 for relative risk). We will discuss relative risk again in the context of survival analysis in Chapter 10.

8.7 EXERCISES

1. Define the following:

(a) Chi-square test

(b) Contingency table

(c) Odds ratio

(d) Relative risk

(e) Cohort study

(f) Case-control study

(g) Test for independence in 2×2 table

2. In a survey study, subjects were asked to report their health as excellent, good, poor, and very poor. They were also asked to answer whether or not they had smoked at least 250 cigarettes in their lifetime. Suppose Table 8.14 represents the outcome of the survey.

Determine if there is a relationship between cigarette usage and reported health status at the 5% significance level, one sided. What is the p-value for the chi-square test? Why is it appropriate to use the chi-square test?

3. For the same subjects in the survey in the above table, the subjects were asked if they were currently smoking, had quit smoking or never smoked. Table 8.15 shows the survey results on smoking status versus health assessment.

Is reported health status related to smoking status? Test at the 5% level one-sided.

Table 8.14
Relationship Between Reported Health Status and Smoking Usage

Reported health status	Smoked 250 or more cigarettes	Smoked 0 to 249 cigarettes	Total
Excellent	30	72	102
Good	350	485	835
Poor	121	138	259
Very poor	39	20	59
Total	540	715	1255

Table 8.15
Relationship Between Reported Health Status and Smoker vs. Non-smoker Categories

Reported health status	Current smoker	Quit smoking	Never smoked	Total
Excellent	20	42	40	102
Good	250	300	285	835
Poor	117	72	70	259
Very poor	29	18	12	59
Total	416	432	407	1255

Why is the chi-square test appropriate here?

4. In the same survey, the subjects also were asked to classify themselves according race, with the choices African American, Asian, Hispanic, Native Americans, or European American. Twelve subjects failed to respond. Table 8.16 shows race versus smoking status.

Is race related to smoking status? Test at 5%. Should you do the test one-sided or two-sided? Is the chi-square an appropriate test? Does it matter that 12 subjects did not respond? If a much higher percentage of the subjects did not respond what might invalidate the analysis?

5. Again using the same survey data, suppose we have statistics at baseline regarding the subjects drinking status, as well as their smoking status. Given the results in the following table, what would you conclude about the relationship between smoking status and drinking status? Test at the 5% level (Table 8.17).

Table 8.16
Relationship Between Race and Smoking Status

Race	Current smoker	Quit smoking	Never smoked	Total
African American	40	34	40	114
Asian	33	34	65	142
Hispanic	117	92	70	279
Native American	20	28	12	60
European American	202	232	214	648
Total	412	430	401	1243

Table 8.17
Relationship Between Smoking and Alcohol Consumption

Smoking status	Current alcohol user	Light alcohol user	Former alcohol user	Never used alcohol	Total
Heavy	40	24	10	40	114
Moderate	33	30	49	30	142
Light	30	72	20	38	160
None	110	158	232	339	839
Total	213	284	311	447	1255

6. A clinical trial is conducted at an academic medical center. Diabetic patients were randomly assigned to a new experimental drug to control blood sugar levels versus a standard approved drug using a $1:1$ randomization. Two hundred patients were assigned to each group, and the 2×2 table (Table 8.18) shows the results.

Test at the 5% level to determine if the new drug is more effective. Is it appropriate to apply the chi-square test? Why would it be difficult to do Fisher's test without a computer? How many contingency tables are possible with the given row and column marginal totals?

7. A study involving 75 patients who at one time used sodium aurothiomalate (SA) as a treatment for rheumatoid arthritis. The purpose was to examine whether or not the toxicity of SA could be linked to the patients sulphoxidation capacity assessed by the sulphoxidation index (SI). For SI, a value of 6.0 was used to separate impaired sulphoxidation ($SI > 6.0$), with unim-

Table 8.18
Fasting Blood Glucose Levels (normal vs. out of normal range) vs. Drug Treatment Group

Fbg level at follow-up	Patients with investigational drug	Patients with approved drug	Total
Normal	119	21	140
Not normal	81	179	260
Total	200	200	400

Table 8.19
Relationship Between High or Normal Sulphoxidation Index and Major Toxicity Reaction

		Major adverse reaction (toxicity)		
		Yes	No	Total
Impaired sulphoxidation (SI > 6.0)	Yes	32	9	41
	No	12	22	34
	Total	44	31	75

paired sulphoxidation (SI ≤ 6.0). The results are given in the 2×2 table below (Table 8.19).

(a) Using Fisher's exact test, determine whether or not impaired sulphoxidation affects toxicity.

(b) Perform a chi-square test on the 2×2 table.

(c) Are the results of the tests similar?

CHAPTER 9

Nonparametric Methods

Most of the statistical methods that we have covered in this book have involved parametric models. The bootstrap, Fisher's exact test, and the chi-square test are the exceptions. A parametric model is one that involves probability distributions that depend on a few parameters (usually five or less). For example, when we assume a normal distribution the parameters, it depends on are the mean and variance. We then use the data to estimate the parameters, and we base our inference on the sampling distribution for these estimates based on the parametric model. But in many practical situations, the parametric model may be hard to justify, or may be found to be inappropriate when we look at the sample data.

Nonparametric methods on the other hand only assume that the distribution function F is continuous. Ranking the data or considering permutations of the data are two ways to construct test statistics that are distribution free under the null hypothesis. Distribution free mean that the distribution of the test statistic is known exactly when the null hypothesis is assumed, and does not depend on the form or parameters of the original data. So, for example, the sign test has a binomial distribution with $p = 1/2$ under the null hypothesis, and Fisher's exact test has a specific hypergeometric distribution when the null hypothesis is true.

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

There is a price paid for this, however. That price is that some information in the data is ignored. In the case of rank tests, we ignore the numerical values of the observations and just consider how they are ordered. So when a parametric model is consistent with the data, the maximum likelihood estimates make efficient use of the data and provide a more powerful test than its nonparametric counterpart, which cannot exploit the information in the distribution's family.

9.1 RANKING DATA

We use rank tests when we want to make inferences about two or more populations and we don't have a good parametric model (that theoretical or empirical work would suggest). Suppose, for example, that we have samples from two populations. Our null hypothesis is that the two distributions are identical. In this case, we pool the observations and order the pooled data from the smallest value to the largest value. This is like temporarily forgetting the population the data points were taken from. Under the null hypothesis, this shouldn't matter, since the distributions are the same. So the data should be well mixed (i.e., there will not be a tendency for the sample from population 1 to have mostly high ranks or mostly low ranks). In fact, we would expect that the average rank for each population would be nearly the same. On the other hand, if the populations were different, then the one with the larger median would tend to have more of the higher ranks than the one with the lower median.

This is the motivation for the Wilcoxon rank-sum test. Comparing the average of the ranks is very similar to comparing the sum of the ranks (if the sample sizes are equal or nearly so). The Wilcoxon rank-sum test (equivalent to the Mann–Whitney test) compares the sum of the ranks from, say, population one, and compares it with the expected value for that rank sum under the null hypothesis. This test is the topic of the next section. It can be generalized to three or more populations. The generalization is called the Kruskal–Wallis test.

9.2 WILCOXON RANK-SUM TEST

The Wilcoxon rank-sum test is a nonparametric analog to the unpaired *t*-test. See Conover (1999) for additional information about this test.

Table 9.1
Left Leg Lifting Among Elderly Males Getting Physical Therapy: Comparing Treatment and Control Groups

Unsorted scores		Scores sorted by rank	
Control group	Treatment group	Control group (rank)	Treatment group (rank)
25	26		16 (1)
66	85	18 (2)	
34	48	25 (3)	
18	68		26 (4)
57	16	34 (5)	
$N_1 = 5$	$N_2 = 5$		48 (6)
		57 (7)	
		66 (8)	
			68 (9)
			85 (10)
		$T1 = \Sigma R = 25$	$T2 = \Sigma R = 30$

As previously mentioned, the test statistic is obtained by ranking the pooled observations, and then summing the ranks of, say, the first population (could equally well have chosen the second population).

We will illustrate the test first scores of a leg-lifting test for elderly men Table 9.1.

For this problem, the sum of all the ranks is $(N_1 + N_2)(N_1 + N_2 + 1)/2 = (5 + 5)\ (5 + 5 + 1)/2 = 10(11)/2 = 55$, since $N_1 = N_2 = 5$. Now, since $N_1/(N_1 + N_2)$ = probability of randomly selecting a patient from group 1 and $N_2/(N_1 + N_2)$ = probability of randomly selecting a patient from group 2, if we multiply $N_1/(N_1 + N_2)$ by $(N_1 + N_2)(N_1 + N_2 + 1)/2$, it gives the expect rank-sum for group 1. This is $N_1(N_1 + N_2 + 1)/2$, which in our example is $5(11)/2 = 27.5$. From the tables for the Wilcoxon test, we see that a rank-sum less than 18 or greater than 37 will be significant at the 0.05 level for a two-side test. Since $T1 = 25$, we cannot reject the null hypothesis.

As a second example, we take another look at the pig blood loss data. Table 9.2 shows the data the pooled rankings.

The *p*-value for this test is greater than 0.20, since the 80% confidence interval for T1 is [88, 122], which contains 112. Now, although

Table 9.2
Pig Blood Loss Data (mL)

Control group pigs (pooled ranks)	Treatment group pigs (pooled ranks)
786 (9)	543 (5)
375 (1)	666 (7)
4446 (19)	455 (3)
2886 (16)	823 (11)
478 (4)	1716 (14)
587 (6)	797 (10)
434 (2)	2828 (15)
4764 (20)	1251 (13)
3281 (17)	702 (8)
3837 (18)	1078 (12)
Sample mean = 2187.40 (rank-sum = 112)	Sample mean = 1085.90 (rank-sum = 98)
Sample *SD* = 1824.27	Sample *SD* = 717.12

the Wilcoxon test cannot reject the null hypothesis that the distributions are the same, the *t*-test (one sided) rejected the null hypothesis that the means are equal. Why do we get conflicting results? First of all, we made two very dubious assumptions when applying the *t*-test. They were (1) both populations have normal distributions, and (2) the distributions have the same variances. Standard tests for normality such as Wilk–Shapiro or Anderson–Darling would reject normality, and the control group standard deviation is about 2.5 times larger than the treatment group. The *F*-test for equality of variances would likely reject equality of variances.

The *t*-test is therefore not reliable. So it should not be a surprise that the test could give erroneous results. Since neither assumption is needed for a nonparametric rank test, it is more trustworthy. The fact that the result is nonsignificant may just be an indication that the sample size is too small. Recall also that the Wilcoxon test is not the most powerful.

We shall now look at another simpler analog to the two-sample independent *t*-test. It is called the sign test, and just looks at the sign of the difference between the two means.

9.3 SIGN TEST

Suppose we are testing the difference in the "center" of two populations that are otherwise the same. This is the same situation that we encountered with the Wilcoxon rank-sum test, except here we will be considering paired observations. So the sign test is an analog to the paired *t*-test. Another test called the Wilcoxon signed-rank test is a little more complicated and more powerful because it uses the idea of ranking the data as well as considering the sign of the paired difference. For simplicity, we will only cover the signed test, and the interested reader can go to Conover (1999) or any of the other many books on nonparametric statistics to learn more about the signed-rank test.

Now, the idea behind the sign test is that we simply look at the paired differences and record whether the difference is positive or negative. We ignore the magnitude of the difference and hence sacrifice some of the information in the data. However, we can take our test statistic to be the number of cases with a positive sign (or we could choose the number with a negative sign). We are assuming the distribution is continuous. So the difference will not be exactly zero. If we choose to do the test in practice when the distribution is discrete, we can simply ignore the cases with 0 as long as there are not very many of them. Whether we choose the positive signs or the negative signs, under the null hypothesis that the distributions are identical, our test statistic has a binomial distribution with parameter n = the number of pairs (or the number of pairs with a nonzero difference in situations where differences can be exactly 0) and $p = P(X > Y) = 1/2$, where $X - Y$ is the paired difference of a randomly chosen pair and the test statistic is the number of positive differences (or $P(X < Y)$ if $X - Y$ is the paired difference and the test statistic is the number of negative pairs).

Now under the alternative hypothesis that the distributions differ in terms of their center, the test statistic is binomial with the same n and $p = P(X > Y)$. However, the parameter p is not equal to 1/2. So the test amounts to the coin flipping problem. Is the coin fair? A coin is fair if it is just as likely to land heads as tails. We are asking the same question about positive signs for our paired differences. So if we compute an exact binomial confidence interval for the proportion of positive paired differences a two-sided test at the 5% significance level amounts to determining whether or not a two-sided 95% confidence interval for p contains 1/2.

Table 9.3
Daily Temperatures for Two Cities: Paired Nonparametric Sign Test

Day	Washington mean temperature (°F)	New York mean temperature (°F)	Paired difference	Sign
1. January 15	31	28	3	+
2. February 15	35	33	2	+
3. March 15	40	37	3	+
4. April 15	52	45	7	+
5. May 15	70	68	2	+
6. June 15	76	74	2	+
7. July 15	93	89	4	+
8. August 15	90	85	5	+
9. September 15	74	69	5	+
10. October 15	55	51	4	+
11. November 15	32	27	5	+
12. December 15	26	24	2	+

If we want to do a one-sided test with the alternative that $p > 1/2$, then we compute a 95% confidence interval of the form (a, 1] and reject the null hypothesis if $a > 1/2$, or for the opposite side and confidence interval of the form [0, b) with $b < 1/2$. Table 9.3 shows how the sign test is applied comparing average temperatures in New York and Washington paired by date.

We note that the number of plus signs is 12, which is the highest possible, favoring Washington as being warmer than New York. So since it is the most extreme; the probability of 12 pluses is the one-sided p-value. So $p = (1/2)^{12} = 0.000244$. Since the binomial distribution is symmetric about 1/2 under the null hypothesis ($p = 1/2$), the two-sided p-value is just double the one-sided p-value which is 0.000488.

9.4 SPEARMAN'S RANK-ORDER CORRELATION COEFFICIENT

Thus far, we know about Pearson's correlation coefficient, which is suitable for bivariate normal data, and its estimate is related to the slope

estimate from the least squares line. But how do we measure a monotonic relationship that is not linear or the data is very nonnormal? Spearman's rank-order correlation coefficient is a nonparametric measure of such relationships.

Suppose we have a relationship given by $Y = \sqrt{X}$ measured with no error and defined for all $X \geq 0$. Recall that Pearson's correlation can be between -1 and 1, and is only equal to 1 or -1 if there is a perfect linear relationship. Now this square root function is a monotonic function but is nonlinear. So the Pearson correlation would be less than 1. In such cases, we would prefer that a correlation measure for a perfect monotonic functional relationship would equal 1 if it is an increasing function such as the square root or -1 for a negative exponential (i.e., $Y = \exp(-X)$). Spearman's rank correlation coefficient does that. In fact, there are two nonparametric correlation measures that have been devised to satisfy this condition for perfect monotonic relationships and be properly interpretable for any continuous bivariate distribution. Spearman rank correlation "ρ" and Kendall's "τ" introduced by Spearman (1904) and Kendall (1938), respectively. We shall only discuss Spearman's ρ.

Spearman's ρ in essence is calculated by the same formula as Pearson's correlation, but with the measured values replaced by their ranks. What exactly do we mean by this? For each X_i, replace the value by the rank of X_i when ranked with relationship to the set of observed Xs with rank 1 for the smallest values in increasing order up to rank n for the largest of the Xs. Do the same for each Y_i. Then take the ranked pairs and compute the correlation for these pair just like you would with Pearson's correlation coefficient. For example, suppose we consider the pair (X_5, Y_5), and X_5 is ranked third out of 20, and Y_5 sixth out of 20, Then we replace the pair with (3, 6) their ranked pair.

The computational formula for Spearman's rank correlation is

$$\rho = \left\{\sum_{i=1}^{n} R(X_i)R(Y_i) - n([n+1]/2)^2\right\} \Big/ \left[\left\{\sum_{i=1}^{n} R(X_i)^2 - n([n+1]/2)^2\right\}\right] \left\{\sum_{i=1}^{n} R(Y_i)^2 - n([n+1]/2)^2\right\}.$$

In the case of no ties, this formula simplifies to

$$\rho = 1 - 6T/[n(n^2 - 1)\},$$

where $T = \sum_{i=1}^{n}[R(X_i) - R(Y_i)]^2$.

As an example, let us look at the correlation between the temperature in Washington and New York over the 12 months of the year. The reason the paired t test worked so well was because most of the variation was due to seasonal effects that were removed through the paired differences. This variation will translate into high correlation between X_i, the temperature in New York on the 15th of month i, with Y_i, the temperature Washington, DC on the 15th of the ith month. Using the ranks, we show in Table 9.4 how the Spearman correlation is calculated in this case.

Table 9.4
Daily Temperatures for Two Cities: Spearman Rank Correlation

Day	Washington mean temperature (°F)	New York mean temperature (°F)	Ranked pairs	Term
	Y (rank)	X (rank)	$[r(x), ry)]$	$[r(y_i) - r(x_i)]^2$
1. January 15	31 (2)	28 (3)	(3, 2)	1
2. February 15	35 (4)	33 (4)	(4. 4)	0
3. March 15	40 (5)	37 (5)	(5, 5)	0
4. April 15	52 (6)	45 (6)	(6, 6)	0
5. May 15	70 (8)	68 (8)	(8, 8)	0
6. June 15	76 (10)	74 (10)	(10, 10)	0
7. July 15	93 (12)	89 (12)	(12. 12)	0
8. August 15	90 (11)	85 (11)	(11, 11)	0
9. September 15	74 (9)	69 (9)	(9, 9)	0
10. October 15	55 (7)	51 (7)	(7, 7)	0
11. November 15	32 (3)	27 (2)	(2, 3)	1
12. December 15	26 (1)	24 (1)	(1.1)	0
T				2
$\rho = 1 - 6T/[n(n^2 - 1)]$				$= 1 - 12/[12(143)]$ $= 142/143$ $= 0.9930$

9.5 INSENSITIVITY OF RANK TESTS TO OUTLIERS

Of course, with univariate data outliers are the extremely large or extremely small observations. For bivariate data, it is less obvious what should constitute an outlier, as there are many directions to consider. Observations that are extreme in both dimensions will usually be outliers, but not always. For example, if data are bivariate normal, the contours of constant probability are ellipses whose major axis is along the linear regression line. When the data are highly correlated, these ellipses are elongated.

If a bivariate observation falls on or near the regression line, it is a likely observation, and if the correlation is positive, and if X and Y are both large or both small, we may not want to consider such observations to be outliers. The real outliers are the points that are far from the center of the semi-minor axis. Another measure, called the influence function, determines a different direction, namely the direction that most highly affects the estimate of a parameter. For the Pearson correlation, the contours of constant influence are hyperbolae. So outliers with respect to correlation are values that are far out on the hyperbolic contours.

We noticed previously that outliers affect the mean and variance estimates, and they can also affect the bivariate correlation. So, confidence intervals and hypothesis tests can be invalidated by outliers. However, nonparametric procedures are designed to apply to a wide variety of distributions, and so should not be sensitive to outliers. Rank tests clearly are insensitive to outliers because a very large value is only one rank higher than the next largest, and this does not at all depend on the magnitude of the observations or how far separated they are.

As an illustration, consider the following data set of 10 values, whose ordered values are 16, 16.5, 16.5, 16.5, 17, 19.5, 21, 23, 24, and 30. The largest value, 30, clearly appears to be an outlier. The sample mean is 20, and half the range is 7, whereas the number 30 is 10 units removed from the mean. The largest and second largest observations are separated by six units, but in term of ranks, 24 has rank 9 and 30 has rank 10, a difference in rank that is the same as between 23 and 24, which have ranks 8 and 9, respectively.

Table 9.5
Pig Blood Loss Data (Modified)

Control group pigs	Treatment group pigs
786	643
375	666
3446	555
1886	823
465	1816
580	997
434	2828
3964	1351
2181	902
3237	1278
Sample mean = 1785.40	Sample mean = 1185.90

9.6 EXERCISES

1. Table 9.5 provides a modification of the pig blood loss data as an exercise for the Wilcoxon rank-sum test

Do the results differ from the standard two-sample *t*-test using pooled variances? Are the resulting *p*-values similar? Compare the *t*-test and the Wilcoxon rank-sum test for a one-side alternative that the treatment group has a lower blood loss average than the control group.

2. Apply the Wilcoxon rank-sum test to the data in Table 9.6 on the relationship between the number of patients with schizophrenia and the season of their birth by calling fall and winter as group 1, and spring and summer as group 2. The four individual seasons represent data points for each group. Ignore the possibility of a year effect.

Do we need to assume that births are uniformly distributed? If we knew that there were a higher percentage of births in the winter months, how would that affect the conclusion?

3. Based on Table 9.7, which is a modification of the temperature data for New York and Washington, apply the sign test to see if the difference in the temperatures is significant.

4. Using the data from Table 9.7 in exercise 3, compute the Spearman rank correlation coefficient between the two cities

Table 9.6
Season of Birth for 600 Schizophrenic Patients over 6 Years

Season		Number of patients
Year 1	Fall	20
Winter		35
Spring		20
Summer		25
Year 1	Total	100
Year 2	Fall	32
Winter		38
Spring		10
Summer		15
Year 2	Total	95
Year 3	Fall	27
Winter		43
Spring		13
Summer		17
Year 3	Total	105
Year 4	Fall	19
Winter		36
Spring		18
Summer		28
Year 4	Total	101
Year 5	Fall	33
Winter		36
Spring		14
Summer		21
Year 5	Total	104
Year 6	Fall	23
Winter		41
Spring		22
Summer		9
Year 6	Total	95

Table 9.7
Temperature Comparisons between Two Cities

Day	Washington mean temperature (°F)	New York mean temperature (°F)	Paired difference	Sign
1. January 15	31	38	–7	–
2. February 15	35	33	2	+
3. March 15	40	37	3	+
4. April 15	52	45	7	+
5. May 15	70	68	2	+
6. June 15	76	74	2	+
7. July 15	93	89	4	+
8. August 15	90	85	5	+
9. September 15	74	69	5	+
10. October 15	55	51	4	+
11. November 15	32	27	5	+
12. December 15	26	24	2	+

Table 9.8
Aggressiveness Scores for 12 Sets of Identical Twins Based on Birth Order

Twin set	1st born aggressiveness score	2nd born aggressiveness score	Paired difference	Sign of paired difference
1	85	88	–3	–
2	71	78	–7	–
3	79	75	4	+
4	69	64	5	+
5	92	96	–4	–
6	72	72	0	0
7	79	64	15	+
8	91	89	2	+
9	70	62	8	+
10	71	80	–9	–
11	89	79	10	+
12	87	75	12	+

Table 9.9
Aggressiveness Scores for 12 Sets of Identical Twins Based on Birth Order

Twin set	1st born aggressiveness score (rank) x	2nd born aggressiveness score (rank) y	Ranked pair (x. y)	Term $r(x_i)r(y_i)$
1	85 (8)	88 (10)	(8, 10)	80
2	71 (3.5)	78 (7)	(3.5, 7)	24.5
3	79 (6.5)	75 (5.5)	(6.5, 5.5)	35.75
4	69 (1)	64 (2.5)	(1, 2.5)	2.5
5	92 (12)	96 (12)	(12, 12)	144
6	72 (5)	72 (4)	(5, 4)	20
7	79 (6.5)	64 (2.5)	(6.5, 2.5)	16.25
8	91 (11)	89 (11)	(11, 11)	121
9	70 (2)	62 (1)	(2, 1)	2
10	71 (3.5)	80 (9)	(3.5, 9)	31.5
11	89 (10)	79 (8)	(10, 8)	80
12	87 (9)	75 (5.5)	(9, 5.5)	49.5

5. Given the aggressiveness scores for the twins shown in Table 9.8, apply the sign test to see of the there is a difference depending on order of birth.

Remember that when the paired difference is 0, we ignore the case. So in this case, we treat the 11 sets without ties (8 pluses and 3 minuses). Are the results statistically significant at the 5% level (two-sided)? What is the two-sided p-value? What is the p-value for the one-sided alternative that the first born is more aggressive?

6. Using Table 9.9, compute the Spearman rank correlation coefficient for the aggressiveness scores. Does this suggest that both twins tend to be similar in degree of aggressiveness?

CHAPTER 10

Survival Analysis

Survival analysis is based on data where patients are followed over time until the occurrence of a particular event such as death, relapse, recurrence, or some other event that is of interest in to the investigator. Of particular interest is the construction of an estimate of a survival curve which is illustrated in Figure 10.1. Survival probability at t represents the probability that an event does not occur by time t.

In the figure, the *x*-axis shows time in years and the *y*-axis survival probability. In this case, the function $S(t)$ is a Weibull curve with $S(t) = \exp(-(\lambda t)^{\beta})$ and $\lambda = 0.4$ and $\beta = 2.0$. In a clinical study, $S(t)$ represents the probability that the time from initiation in the study ($t = 0$) until the occurrence of the event for an arbitrary patient is greater than a specified value t. The curve represents the value for this probability as a function of t. Data on the observed time to the event for each patient is the information to use to estimate the survival curve for all t in $(0, \infty)$. See Section 10.4.2 for more details on the Weibull family of survival curves.

The survival curve or the comparison of two or more survival curves is often important in determining the effectiveness of a new treatment. It can be used for efficacy as in the case of showing that an anticoagulant is effective at reducing stroke for patients with atrial fibrillation. More often, it is used as a safety parameter, such as in the

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

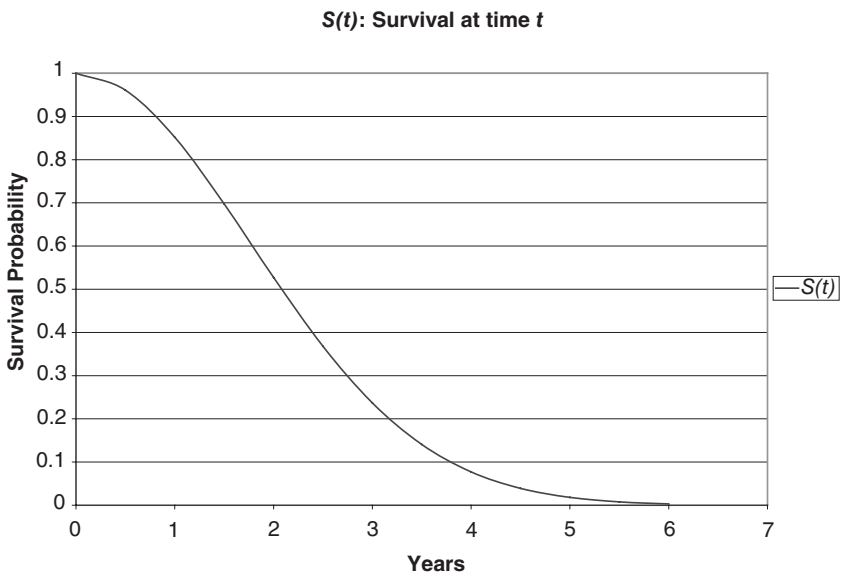


Figure 10.1. A typical survival curve.

determination of a particular adverse event that the treatment is suspected to cause. The term survival analysis came about because it was originally used when mortality was the outcome, but it can be used for time-to-event data for any event. More generally, the curve does not necessarily have to be a function of time. It is even possible for time to be replaced by a variable that increases with time, such as the cost of a worker's compensation claim where the event occurs when the claim is closed.

10.1 TIME-TO-EVENT DATA AND RIGHT CENSORING

What characterizes survival data is that some patients have incomplete results. In a particular study, there is a time at which the study ends and the data must be analyzed. At that point, some of the patients may not have experienced the event (either because they will never have the event or because the event will occur some time later). The data for these patients should not be thrown out because that would (1) ignore valuable information about the time to event, since these patients time

to event must at least be longer than the time from study initiation until the termination of the study (called the censoring time); and (2) leaving them out biases the estimate of parameters, such as median or mean survival time, since the censored observations are more likely to be the longer times than those that were not censored. So from (2), we see that the median time-to-event is underestimated if the censored data are ignored. Other censoring could occur if the patient becomes lost to follow-up prior to the date of completion for the study.

What makes survival analysis different is the existence of incomplete data on some patients whose time to event is right censored (i.e., cut off at the end of the study). The key to the analysis is to find parametric, semi-parametric, or nonparametric ways to estimate the survival curve utilizing both the complete and incomplete observations. This will often allow for a less biased median survival time estimate. The remainder of the chapter will cover various methods.

The first method is the life table. Although the methods we describe here are straightforward, there are many practical difficulties. One of these is the problem of unreported events. This is a very big problem with medical devices. Attempts have been made to address the issue of bias in estimates due to underreporting. But these methods must rely heavily on assumptions about the underreporting. The article by Chernick et al. (2002) covers the issue in detail.

10.2 LIFE TABLES

The survival curve $S(t)$ is defined to be equal to the probability that $T > t$ where T is the random variable representing the time to the event. The data is Table 10.1 is taken from Altman (1991, p. 367). In this example events are restricted to the time $(0, L]$ with events occurring after time L, right censored.

We notice from the table that patients are accrued over time for slightly less than 6 months. The study is terminated at 18 months after the first patient is enrolled in the study. Four patients died during the trial six were either living at the end of the trial or lost to follow-up. Specifically, patients 1, 5, 7, and 10 died, patients 3, 6, 8, and 9 completed the study alive and patients 2 and 4 were lost to follow-up. This table provides us with exactly all we need to construct the various types of survival curves.

Table 10.1
Survival Times for Patients

Censor code*	Patient no.	Time at entry (months)	Time to death or censoring (months)	Survival time (months)
1	1	0.0	11.8	11.8
0	2	0.0	12.5	12.5†
0	3	0.4	18.0	17.6†
0	4	1.2	4.4	3.2†
1	5	1.2	6.6	5.4
0	6	3.0	18.0	15.0†
1	7	3.4	4.9	1.5
0	8	4.7	18.0	13.3†
0	9	5.0	18.0	13.0†
1	10	5.8	10.1	4.3

*Death occurred = 1, censoring = 0, $L = 18.0$.

†Censored observation.

Life tables give survival probability estimates for intervals of time whereas survival curves are continuous over time (although their nonparametric estimates are step functions that only change when events occur). Life tables must be used when the only information that is available is the number of events occurring in the intervals. If we have the exact times when each event occurs, and all the times when censoring occurs, we can estimate the survival curve by parametric or nonparametric methods.

We can also create a life table by choosing time intervals and counting the number of events and censoring times that occur in each specified interval. However, the use of life tables when we have the exact times for the events and censoring is inefficient, since it ignores some of the available information about survival (namely, where in the interval each event occurs). In addition to the interval survival probability, the life table provides an estimate of the cumulative survival probability at the end of the time interval for each interval. Whether we are estimating cumulative survival over time or for life table intervals, there is a key equation that is exploited. It is shown as Equation 10.1.

$$S(t_2) = P(t_2 \mid t_1)S(t_1) \text{ for any } t_2 > t_1 \geq 0, \tag{10.1}$$

Table 10.2
Life Table for Patients From Table 10.1

Time interval I_j	No. of deaths in I_j	No. withdrawn in I_j	No. at risk in I_j	Avg. No. at risk in I_j	Est. prop. of deaths in I_j	Est. prop. Surv. at end of I_j	Est. cum. surv. at end of I_j
[0, 3)	1	0	10	10	0.1	0.9	0.9
[3, 6)	2	1	9	8.5	0.235	0.765	0.688
[6, 9)	0	0	6	6	0.0	1.0	0.688
[9, 12)	1	0	6	6	0.167	0.833	0.573
[12, 15)	0	3	5	5	0	1.0	0.573
[15, 18)	0	2	2	2	0	1.0	0.573
[18, ∞)	0	0	0	0	—	—	—

where $S(t)$ = survival probability at time $t = P(T > t)$, t_1 is the previous time of interest, and t_2 is some later time of interest (for a life table, t_1 is the beginning of the interval, and t_2 is the end of the interval.

For the life table, we must use the data as in Table 10.1 to construct the estimates that we show in Table 10.2. In the first time interval, say [0, a], we know that $S(0) = 1$ and $S(a) = P(a|0)S(0) = P(a|0)$. This is gotten by applying Equation 10.1, with $t_1 = 0$ and $t_2 = a$, and substituting 1 for $S(0)$. The life table estimate was introduced by Cutler and Ederer (1958), and therefore it also is sometimes called the Cutler–Ederer method. We exhibit the life table as Table 10.2, and then will explain the computations.

In constructing Table 10.2 from the data displayed in Table 10.1, we see that including event times and censoring times, the data range from 1.5 to 17.6 months. Note that since time of entry dies not start at the beginning of the study, the time to event is shifted by subtracting the time of entry from the time of the event (death or censoring). We choose to create 3-month intervals out to 18 months. The seven intervals comprising all times greater than 0 are: (0, 3), [3, 6), [6, 9), [9, 12), [12, 15), [15, 18) and [18, ∞). Intervals denoted [a, b) include the number "a" and all real numbers up to but not including "b." Intervals (a, b) include all real numbers greater than "a" and less than "b" but do not include "a" or "b." In each interval, we need to

determine the number of subjects who died during the interval, the number withdrawn during the interval, the total number at risk at the beginning of the interval, and the average number at risk during the interval.

To understand Table 10.2, we need to explain the meaning of the column heading.

Column 1 is labeled "Time interval" and is denoted I_j for the jth interval.

Column 2 contains the number that died in the jth interval and is denoted D_j.

Column 3 contains the number that withdrew during the jth interval and is denoted W_j.

Column 4 contains the number at risk at the start of the jth interval and is denoted N_j.

Column 5 is the average number at risk during the jth interval and is denoted N_j'.

Column 6 is the estimated proportion of deaths during the interval and is denoted as q_j.

Column 7 is the estimated proportion of subjects surviving the interval and is denoted by p_j.

Column 8 is the cumulative probability of surviving the interval.

We note that the deaths are determined just by counting the deaths with event time falling in the interval. The withdrawals are simply determined by counting the number of censoring times falling in the interval. The number at risk at the beginning of the interval is just the total at time 0 minus all deaths and withdrawals that occurred from time 0 up to but not including time "a" where "a" is the beginning time for the interval.

Now the average number remaining over the jth interval is $N_j' = N_j - (W_j/2)$. We then get the estimated proportion that are dead, to be $q_j = D_j/N_j'$. Then the estimated proportion surviving the interval is $p_j = 1 - q_j$. Remember the key recursion in Equation 10.1? It gives $S_j = p_j S_{j-1}$. This recursive equation allows S_1 to be determined from the known value S_0 after calculating p_1. Then S_2 is calculated using S_1 and p_2, and this continues up to the time of the last event or censor time.

This is the Cutler–Ederer method, and just about any life table is generated in a very similar fashion.

10.3 KAPLAN–MEIER CURVES

The Kaplan–Meier curve is a nonparametric estimate of the survival function (see Kaplan and Meier 1958). It is computed using the same conditioning principle that we used for the life table. However, here we estimate the survival at every time point, but only do the iterative computations at the event or censoring times. The estimate is taken to be constant between points. It has sometimes been called the product limit estimator, because at each time point, it is calculated as the product of conditional probabilities. Next, we describe in detail how the curve is estimated.

10.3.1 The Kaplan–Meier Curve: A Nonparametric Estimate of Survival

For all time from 0 to t_1, where t_1 is the time of the first event, the Kaplan–Meier survival estimate is $S_{km}(t) = 1$. At time t_1, $S_{km}(t_1) = S_{km}(0)(n_1 - D_1)/n_1$, where n_1 is the total number at risk, and D_1 is the number that die (have an event) at time t_1. Since $S_{km}(0) = 1$, $S_{km}(t_1) = (n_1 - D_1)/n_1$. For the example in Table 10.3, below we see that $S_{km}(t_1) =$

Table 10.3
Kaplan–Meier Survival Estimates for Example in Table 10.1

Time	No. of deaths in D_j	No. withdrawals W_j	No. at risk n_j	Est. prop. of deaths q_j	Est. prop. surviving $p_{j=1} - q_j$	Est. cumulative survival $S_{km}(t_j)$
$t_1 = 1.5$	1	0	10	0.1	0.9	0.9
$t_2 = 4.3$	1	1	9	0.125	0.875	0.788
$t_3 = 5.4$	1	0	6	0.143	0.857	0.675
$t_4 = 11.8$	1	0	6	0.167	0.833	0.562
$18 > t > 11.8$	0	5	5	0	1.0	0.562
$t \geq 18$	0	0	0	0	—	—

$(10 - 1)/10 = 0.9$. At the next death time t_2, $S_{km}(t_2) = S_{km}(t_1)(n_2 - D_2)/n_2$. For n_2, we use the value of N_2 in Table 10.2, and get $S_{km}(t_2) = (0.9)(8 - 1)/8 = 0.9(7/8) = 0.9(0.875) = 0.788$. Note that $n_2 = 8$ because there was one withdrawal between time t_1 and t_2. The usual convention is to assume "deaths before losses." This means that if events occur at the same time as censored observations, the censored observations are left in the patients at risk for each event at that time and removed before the next event occurring at a later time.

We notice a similarity in the computations when comparing Kaplan–Meier with the life table estimates. However the event times do not coincide with the endpoints of the intervals and this leads to quantitative differences. For example, at t = 4.3, the Kaplan–Meier estimate is 0.788, whereas the life table estimate is 0.688. At t = 5.4, the Kaplan–Meier estimate is 0.675 whereas the life table is 0.688. At and after t = 11.8 the Kaplan–Meier estimate is 0.562, and the life table estimate is 0.573. Although there are numerical differences qualitatively, the two methods give similar results.

10.3.2 Confidence Intervals for the Kaplan–Meier Estimate

Approximate confidence intervals at any specific time t can be obtained by using Greenwood's formula for the standard error of the estimate and the asymptotic normality of the estimate. For simplicity, let S_j denote $S_{km}(t_j)$. Greenwood's estimate of variance is $V_j = S_j^2[\sum_{i=1}^{j} q_i/(n_i p_i)]$. Greenwood's approximation for the 95% confidence interval at time t_j is $\left[S_j - 1.96\sqrt{V_j}, S_j + 1.96\sqrt{V_j}\right]$.

Although Greenwood's formula is computationally easy through a recursion equation, the Peto approximation is much simpler. The variance estimate for Peto's approximation is $U_j = S_j^2(1 - S_j)/n_j$. Peto's approximation for the 95% confidence interval at time t_j is $\left[S_j - 1.96\sqrt{U_j}, S_j + 1.96\sqrt{U_j}\right]$.

Dorey and Korn (1987) have shown that Peto's method can give better lower confidence bounds than Greenwood's, especially at long follow-up times where there are very few patients remaining at risk. In the example in Table 10.3, we shall now compare the Peto 95% confidence interval with Greenwood's at time $t = t_3$. For Greenwood, we

need to calculate V_3, which requires recursively calculating V_1 and V_2 first.

$V_1 = (0.9)^2[0.1/\{10[0.9]\})] = (0.9)(0.01) = 0.009$. Then

$$V_2 = S_2^2[q_2/(n_2 p_2) + V_1/S_{j-1}^2] = (0.788)^2[0.125/\{8(0.875) + 0.009/(0.9)^2\}]$$
$$= 0.621(0.0179 + 0.0111) = 0.621(0.029) = 0.0180.$$ Finally,

$V_3 = (0.675)^2[0.143/\{7(0.857)\} + 0.018/(0.788)^2] = 0.4556[0.143/6] = 0.0109$. So the 95% Greenwood confidence interval is

$$\left[0.675 - 1.96\sqrt{0.0109}, 0.675 + 1.96\sqrt{0.0109}\right]$$
$$= [0.675 - 0.2046, 0.675 + 0.2046] = [0.4704, 0.8796].$$

For Peto's estimate of variance, U_3, we simply calculate

$$U_3 = S_3^2(1 - S_3)/n_3 = (0.675)^2(1 - 0.675)/7$$
$$= (0.675)^2(0.325)/7 = 0.4556.(0.0464) = 0.0212\ .$$

So Peto's estimate is

$$\left[0.675 - 1.96\sqrt{0.0212}, 0.675 + 1.96\sqrt{0.0212}\right]$$
$$= [0.675 - 0.285, 0.675 + 0.285] = [0.390, 0.960].$$

In this example, we see that Peto's interval is much wider and hence more conservative than Greenwood's. However, that does not necessarily make it more accurate. Both methods are just approximations, and we cannot say that one is always superior to the other.

10.3.3 The Logrank and Chi-Square Tests: Comparing Two or More Survival Curves

To compare two survival curves in a parametric family of distributions, such as the negative exponential or the Weibull distribution, we only need to test for differences in the parameters. However, for a nonparametric estimate, we look for departures in the two Kaplan–Meier curves. The logrank test is a nonparametric test for testing equality of two survival curves against the alternative of some difference. Details about the test can be found in the original work of Mantel (1966) or in texts such as Lee (1992, pp. 109–112) or Hosmer et al. (2008).

Rather than go into the detail of computing the logrank test for comparing the two survival curves, we can conduct a similar test that

Table 10.4
Computation of Expected Numbers for the Chi-Square Test in the Breast Cancer Example

Remission time T	Number of remissions at T d_T	Number at risk in treatment group n_1	Number at risk in control group n_2	Expected frequency in treatment group E_1	Expected frequency in control group E_2
15	1	5	5	0.5	0.5
18	1	4	4	0.5	0.5
19	2	3	3	1.0	1.0
20	1	3	1	0.75	0.25
23	1	2	0	1.0	0.0
Total	—	—	—	3.75	2.25

has an asymptotic chi-square distribution with $k - 1$ degrees of freedom, where k is the number of survival curves being compared. For comparing two curves, the test statistic is chi-square with 1 degree of freedom under the null hypothesis. The chi-square statistic as usual takes the form $\sum_{i=1}^{k}(O_i - E_i)^2/E_i$, where n is the number of event times.

The expected values E_i are computed by pooling the survival data and computing the expected numbers in each group based on the pooled data (which is the expected number when the null hypothesis is true, and we condition on the total number of events at the event time points and sum up the expected numbers. Our example is from a breast cancer trial.

In the breast cancer study, the remission times for the treatment group, getting cyclophosphamide, methatrexate, and fluorouracil (CMF), are 23 months, and four patients censored at 16, 18, 20, and 24 months. For the control group, remission times were at 15, 18, 19, 19, and 20, and there were no censoring times. Table 10.4 shows the chi-square calculation for expected frequencies in the treatment and control groups in a breast cancer trial.

Based on the table above, we can compute the chi-square statistic, $(1 - 3.75)^2/4.75 + (5 - 2.25)^2/2.25 = 1.592 + 3.361 = 4.953$. From the chi-square table with 1 degree of freedom, we see that a value of 3.841 corresponds to a p-value of 0.05 and 6.635 to a p-value of 0.01. Hence, since $3.841 < 4.953 < 6.635$, we know that the p-value for this test is

between 0.01 and 0.05, and the survival curves differ significantly at the 5% level.

The logrank test is very similar except that instead of E_i in the denominator, we compute $V = \sum_{i=1}^{m} v_i$, where m is the number of time points for events from the pooled data, and $v_i = n_{1i}n_{2i}d_i(n_i - d_i)/[n_i^2(n_i - 1)]$, where n_{1i} = number at risk in group 1 at time t_i, n_{2i} = number at risk at time t_i in group 2, $n_i = n_{1i} + n_{2i}$, and d_i = combined number of deaths (events pooled from all groups) that have occurred by time t_i. For two groups, the logrank test also has an approximate chi-square distribution with 1 degree of freedom under the null hypothesis. A nice illustration of the use of the logrank test with the aid of SAS software can be found in Walker and Shostak (2010). Additional examples of two-sample and k-sample tests can be found in many standard references on survival analysis, including, for example, Hosmer et al. (2008).

10.4 PARAMETRIC SURVIVAL CURVES

When the survival function has a specific parametric form, we can estimate the survival curve by estimating just a few parameters (usually 1 to 4 parameters). We shall describe two of the most common parametric models, the negative exponential and the Weibull distribution models.

10.4.1 Negative Exponential* Survival Distributions

The negative exponential survival distribution is a one-parameter family of probability models determined by a parameter λ, called the rate parameter or failure rate parameter. It has been found to be a good model for simple product failures, such as the electric light bulb. In survival analysis, we have several related functions. For the negative exponential model, the survival function $S(t) = \exp(-\lambda t)$, where $t \geq 0$ and $\lambda > 0$. The distribution function $F(t) = 1 - S(t) = 1 - \exp(-\lambda t)$, $f(t)$ is the density function, which is the derivative of $F(t)$, $f(t) = \lambda\exp(-\lambda t)$. The hazard function $h(t) = f(t)/S(t)$. For the negative exponential model,

* Also simply referred to as the exponential distribution.

Table 10.5
Negative Exponential Survival Estimates for Patients From Table 10.3

Time T	Number of deaths Dj	Number of withdrawals Wj	Number at risk n_j	Est. prop, of deaths q_j	Est. prop. surviving p_j	KM survival estimate	Negative exp. survival estimate
1.5	1	0	10	0.1	0.9	0.9	0.940
4.3	1	1	8	0.125	0.875	0.788	0.838
5.4	1	0	7	0.143	0.857	0.675	0.801
11.8	1	0	6	0.167	0.833	0.562	0.616
18	0	5	5	0	1	0.562	0.478

$h(t) = \lambda\exp(-\lambda t)/\exp(-\lambda t) = \lambda$. In this case, we will fit an exponential model to the data used to fit the Kaplan–Meier curve in Table 10.3. Table 10.5 compares the estimated negative exponential survival curve with the Kaplan–Meier estimate.

The exponential survival curve differs markedly from the Kaplan–Meier curve, indicating that the negative exponential does not adequately fit the data.

10.4.2 Weibull Family of Survival Distributions

The Weibull model is more general and involves two parameters λ and β. The negative exponential is the special case of a Weibull model, when $\beta = 1$. The Weibull is common in reliability primarily because it is the limiting distribution for the minimum of a sequence of independent identically distributed random variables. In some situations, a failure time can be the first of many possible event times, and hence is a minimum. So under common conditions, the Weibull occurs as an extreme value limiting distribution similar to the way the normal distribution is the limiting distribution for sums or averages.

For the Weibull model $S(t) = \exp(-(\lambda t)^{\beta})$, $F(t) = 1 - \exp(-(\lambda t)^{\beta})$, $f(t) = \lambda\beta(\lambda t)^{\beta-1}\exp[-(\lambda t)^{\beta}]$, and $h(t) = \lambda\beta(\lambda t)^{\beta-1}$. For the Weibull model, $\lambda > 0$ and $\beta > 0$.

10.5 COX PROPORTIONAL HAZARD MODELS

The Cox proportional hazards regression model is called semi-parametric because it includes regression parameters for covariates (which may or may not be time dependent), but in terms of the baseline hazard function, it is completely general (hence not parametric). So part of the modeling is parametric, and another part is nonparametric, hence the term semi-parametric. In SAS®, the model can be implemented using the procedure PHREG, or STCOX in STATA. An excellent and detailed treatment with SAS applications can be found in Walker and Shostak (2010, pp. 413–428). A similar treatment using the STATA software package can be found in Cleves et al. (2008).

The purpose of the model is to test for the effects of a specific set of k covariates on the event times. These covariates can be numerical or categorical. In the case of categorical variables, such as treatment groups, the model can estimate relative risks for the occurrence of an event in a fixed interval when the patient gets treatment A versus when the patient gets treatment B.

For example, in the RE-LY trial to compare three treatments, two doses of dabigatran and warfarin as a control, the Cox model was used to estimate the relative risk of the patient getting a stroke during the trial while on one treatment versus another. This ratio was used to test for superiority or noninferiority of the dabigatran doses versus warfarin with respect to stroke or systemic embolism as the event. The model was also used for other types of event, with major bleeding being a primary safety endpoint.

The model is defined by its hazard function $h(t) = \lambda(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m)$, where m is the number of covariates the X_i are the covariates and is the baseline hazard function (t represents time). We only consider $t \geq 0$. It is called a proportional hazards model because $h(t)$ is proportional to $\lambda(t)$, since $h(t)/\lambda(t)$ is a constant (does not depend on t) that is determined by the covariates. The parameters β_i are estimated by maximizing the partial likelihood. The estimation procedure will not be described here, but its computation requires the use of numerical methods and high-speed computers.

There are many books on survival analysis that cover the Cox model, and even some solely dedicate to the method. A recent text providing an up-to-date theoretical treatment is O'Quigley (2008), which includes over 700 references. Other texts worthy of mention are

Cox and Oakes (1984), Kalbfleisch and Prentice (1980, 2002), Therneau and Grambsch (2000), Lachin (2000), Klein and Moeschberger (2003, paperback 2010), Hosmer and Lemeshow (1999), Cleves et al. (2008), Klein and Moeschberger (2003), and Hosmer et al. (2008). There have been a number of extensions of the Cox model, including having the covariates depend on time. See Therneau and Grambsch (2000) if you want a lucid and detailed account of these extensions. Parametric regression models for survival curves can be undertaken using the SAS procedure LIFEREG and the corresponding procedure STREG in STATA.

10.6 CURE RATE MODELS

The methods for analysis of cure rate models are similar to those previously mentioned, and require the same type of survival information. However, the parametric models previously described all have cumulative survival curves tending to zero as time goes to infinity. For cure rate models, a positive probability of a cure is assumed. So the cumulative survival curve for a cure rate model converges to $p > 0$ as time goes to infinity, where p is called the cure probability, cure fraction or cure rate. Often the goal in these models is to estimate p.

For nonparametric methods such as the Kaplan–Meier approach, p is difficult to detect. It would be the asymptotic limit as t gets larger, but the Kaplan–Meier curve gives us no information about the behavior of the survival curve beyond the last event time or censoring time (whichever is last). So to estimate the cure rate requires a parametric mixture model.

The mixture model for cure rates was first introduced by Berkson and Gage (1952). The general model is given by the following equation:

$$S(t) = p + (1 - p)S_1(t)$$

where p is the cure probability, and $S_1(t)$ is the survival curve for those who are not cured. $S_1(t)$ is the conditional survival curve given the patient is not cured. The conditional survival curve can be estimated by parametric or nonparametric methods. For an extensive treatment of cure rate models using the frequentist approach, see Maller and Zhou

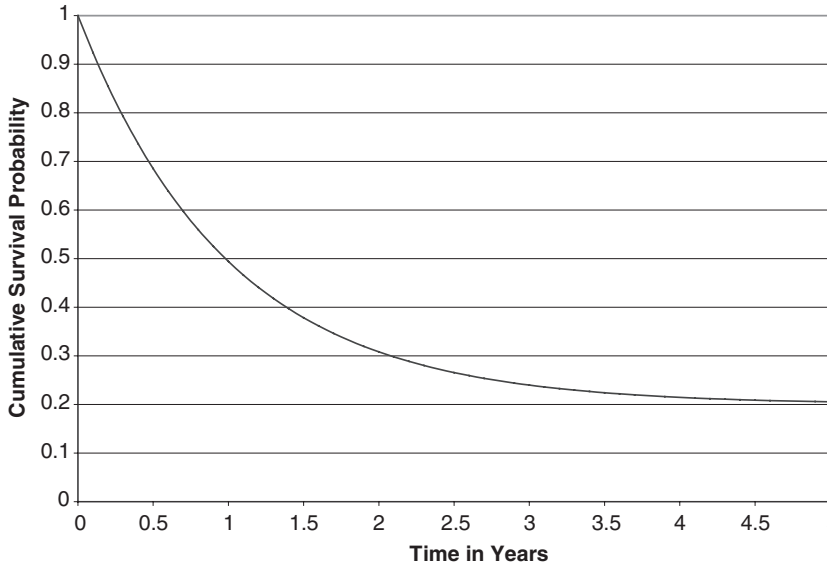


Figure 10.2. Exponential cure rate model with cure rate $p = 0.20$ and exponential rate parameter $\lambda = 1$.Sente videm patum ad inam nonvere timorio rterumunina nihi, catum

(1996). The Bayesian approach to cure rate models can be found in Ibrahim et al. (2001).

We illustrate a parametric mixture survival curve with an exponential survival curve with rate parameter $\lambda = 1$, for the conditional survival curve $S_1(t)$ and with survival probability $p = 0.2$. This curve is shown in Figure 10.2.

Although cure rate modeling began with Berkson and Gage in the 1950s, much of the literature came about in the 1990s when computing became much faster and the EM algorithm for the frequency approach and MCMC methods for Bayesian approaches became easy to implement. Until recently, the free software WinBUGS was the main option for doing MCMC methods for the Bayesian approach to modeling. However, very recently in SAS Version 9.2, MCMC methods have been added as a procedure in SAS/STAT. Users of SAS software may find this more convenient.

10.7 EXERCISES

1. Define the following:

(a) Life table

(b) Kaplan–Meier curve

(c) Negative exponential survival distribution

(d) Cure rate model

(e) Chi-square test to compare two survival curves

2. If the survival function $S(t) = 1 - t/b$ for $0 \le t \le b$, where b is a fixed positive constant, calculate the hazard function. When is the hazard function lowest? Is there a highest rate?

3. Suppose + denotes a censoring event, and that the event times in months for group1 are [8.1, 12, 17 33+, 55, and 61] while for group 2 they are [32, 60, 67, 76+, 80+, and 94]. Test to see if the survival curves are different using the chi-square test.

4. Suppose the survival time since a bone marrow transplant for eight patients who received the transplant is 3, 4.5, 6, 11, 18.5 20, 26, and 35. No observations were censored.

(a) What is the median survival time for these patients?

(b) What is the mean survival time?

(c) Construct a life table where each interval is 5 months.

5. Using the data in example 4:

(a) Calculate a Kaplan–Meier curve for the survival distribution

(b) Fit a negative exponential model.

(c) Compare b with a.

(d) Is the negative exponential survival distribution a good fit in this case?

6. Modify the data in example 4 by making 6, 18.5, and 35 censoring times

(a) Estimate the median survival time.

(b) Why would an average of all the survival times (excluding the censoring times) be inappropriate?

(c) Would an average including the censoring times be appropriate?

7. Now using the data as it has been modified in exercise 6, repeat exercise 5a.

8. Listed below are survival and censoring times (using the + sign for censoring) for six males and six females.

Males: 1, 3, 4+, 9, 11, 15

Females 1, 3+, 6, 9, 10, 11+

(a) Calculate the Kaplan–Meier curve for males

(b) Calculate the Kaplan–Meier curve for females

(c) Test for a difference between the male and female survival curves using the chi-square test.

(d) Compute the logrank statistic and perform the same test as in *c* using this statistic? Do you reach the same conclusion as in *c*? Are the chi-square and logrank test statistics close in value? Are the *p*-values nearly the same?

9. What assumptions are required for the Cox proportional hazard model? Why is it called a semi-parametric method?

10. Suppose a cure model is known to have $S_1(t) = \exp(-0.5t)$.

Recall $S(t) = p + (1 - p)S_1(t)$. Suppose that we know that $S(2) = 0.5259$. Can you calculate the cure rate for this model? If so what is it?

Solutions to Selected Exercises

Chapter 1

1. What is a Kaplan–Meier curve?

A Kaplan–Meier curve is an estimate of cumulative survival over time based on possibly right-censored time-to-event data. It is an estimate obtained without making a parametric assumption about the shape of the survival curve.

3. Why is randomization important in clinical trials?

In clinical trials, we are comparing two or more treatments. Confounding can occur when the subjects in one treatment group has very different characteristics than the other. In one case, there may be a much higher percentage of males in one group than in the other, or one group might tend to have older patients than the other. In such situations, a significant difference in response between the two groups could be due to the difference in treatment, but it also could be due to differences in ages or gender. Randomization tends to balance out these factors, thus eliminating the confounding.

7. What are retrospective studies?

Retrospective studies are any type of study where all the data were generated in the past and are now being used for the purpose of an investigation that was not considered prior to the collection of data.

9. What are controlled clinical trials and why is blinding important?

In clinical trials, we are comparing two or more treatments on human subjects. The trial is considered controlled when randomization is properly used and blinding is

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

included. When the investigator and/or the patient know which treatment group they are in before the completion of the treatment, they could act in a way that creates bias in the estimates. If both the investigator and the patient are unaware of the treatment the patients are more likely to all be treated in the same manner and bias will not creep into the study.

Chapter 2

3. Describe and contrast the following types of sampling designs. Also state when if ever it is appropriate to use the particular designs:

(a) Simple random sample

(b) Stratified random sample

(c) Convenience sample

(d) Systematic sample

(e) Cluster sample

(f) Bootstrap sample

(a) Simple random sampling is just sampling at random without replacement from a well-defined population.

(b) Stratified random sampling is a sampling procedure where the data are divided into groups (strata) that make the subpopulations homogeneous groups. In each strata, a specific number patients are sampled at random without replacement. So it is a collection of simple random samples drawn for each strata. Stratified random sampling is better than simple random sampling when subpopulations are homogeneous, and there are differences between the groups. If the original population is already very homogeneous, there is no benefit to stratification over simple random sampling. It is possible to obtain unbiased estimates of the population mean by either sampling technique, but one estimate will have a lower variance compared with the other depending on the degree of homogeneity within and between the subpopulations.

(c) A convenience sample is any sample that is collected in an operationally convenient way. This is usually not an acceptable way to sample because it is not possible to draw inferences about the population from the sample. This is because inference depends on having known probabilities for drawing elements from the population.

(d) Systematic sampling is an ordered way of selecting elements from the population. So, for example, if you wish to take a 20% sample, you can enumerate the population and draw the first and skip the next four until you have run through the entire population. Systematic sampling can sometimes be easier than random sampling, and if there is no pattern to the ordering it may behave like a simple random sample. However, if there are patterns, such as cycles, the method can be extremely biased. In the 20% sample, suppose that the data

formed a sine wave as you step through the order. If the peak of the cycle occurs at the first case and repeats every five you will only collect the high values, and the mean will be much larger than the population mean. Similarly, if the trough occurs in the first sample and the cycle length is five, you collect only the lowest values from the population, and the sample mean will be too low.

(e) Cluster sampling is another way that may be more convenient than simple random sampling. For example, when the Census Bureau does survey sampling in a city, it may be convenient to sample every house on a particular block since the blocks form a list that can be randomly sampled. In such situations, cluster sampling has advantages.

(f) Bootstrap sampling is not a procedure to sample from the population per se. Instead, we have a sample (presumably random), and bootstrapping is sampling with replacement from this sample to try to infer properties of the population based on the variability of the bootstrap samples. In the ordinary case when the sample size is n, we also take n elements for the bootstrap sample by sampling with replacement from the n elements in the original sample.

6. How does bootstrap sampling differ from simple random sampling?

As described earlier, bootstrap samples with replacement from a random sample, whereas simple random sampling samples without replacement from a population.

7. What is rejection sampling and how is it used?

Rejection sampling is a method for sampling at random without replacement. A common way to sample without replacement is to eliminate the elements from the population as they are selected in sequence and randomly sample each time from the reduced population. With rejection sampling you can achieve the same properties without changing the population you draw the samples from. You simply keep a running list of all the elements that have thus far been sampled, and if the new one is a repeat of one of the old ones, you throw it out and try again always making sure that nothing repeats.

10. Why is the sampling design choice more critical than the size of the sample?

If you make a bad choice of design you can create a large bias that cannot be overcome by an increase in sample size no matter how large you make it. However if the sample size is too small but the design is appropriate, you can obtain unbiased estimates of the population parameters. Increasing the sample size will not prevent us from obtaining an unbiased estimate, and since the accuracy of an unbiased estimate depends only on its variance, the sample size increase will reduce the variance and make the estimate more accurate. So with a good design, we can improve the estimate by increasing the sample size. But no increase in sample size will remove a bias that is due to the poor design.

Chapter 3

1. What does a stem-and-leaf diagram show?

A stem-and leaf diagram has the nice property of describing the shape of the data distribution in a way similar to a histogram but without losing information about the exact value of the cases with a histogram bin.

3. What is the difference between a histogram and a relative frequency histogram?

A histogram has a bar height that equals the number of cases belonging to the bin interval, The relative frequency histogram has the same shape, but the height represents the number of cases belonging to the bin interval divided by the total number n of cases in the entire sample. So the height of the bar represents a proportion or percentage of the data falling in the interval (or a frequency relative to the total).

5. What portion of the data is contained in the box portion or body of a box-and-whiskers plot?

The bottom of the box is the 25th percentile and the top is the 75th percentile. So the box contains 50% of the data.

7. What relationship can you make to the three measures of location (mean, median, and mode) for right–skewed distributions?

For unimodal distributions that are right skewed: mean < median < mode.

10. What is the definition of mean square error?

The mean square error is the average of the squared deviations of the observations from their target. Note that the target is not always the mean. Using this definition one can show that Mean Square Error = B^2 + Variance, where B is the bias (the difference between the mean and the target). When the estimate is unbiased, $B = 0$, and Mean Square Error = Variance.

Chapter 4

1. What is a continuous distribution?

A continuous distribution is a probability distribution with a density defined on an interval, the whole real line, or a set of disjoint intervals.

2. What is important about the normal distribution that makes it different from other continuous distributions?

The normal distribution is a special continuous distribution because of the central limit theorem, which states that for most distributions (continuous or discrete) the average of a set of n independent observations with that distribution has a distribution that is approximately a normal distribution if n is large (usually 30 or more).

6. How are the median, mean, and mode related for the normal distribution?

For any normal distribution by the symmetry property, the mean and median are the same, and the distribution is also unimodal with the mode at the mean. So the three measures are always equal for normal distributions.

10. How is the t distribution related to the normal distribution? What is different about the *t*-statistic particularly when the sample size is small?

Student's *t*-distribution with *n* degrees of freedom is approximately the same as a standard normal distribution when *n* is large (large is somewhere between 30 and 100. When *n* is small, the *t*-distribution is centered at 0, and is symmetric, but the tails drop off much more slowly than for the standard normal distribution (small is from 2 to 30). The smaller the degrees of freedom are, the heavier are the tails of the distribution.

11. Assume that the weight of women in the United States who are between the ages of 20 and 35 years has a normal distribution (approximately), with a mean of 120 lbs and a standard deviation of 18 lbs. Suppose you could select a simple random sample of 100 of these women. How many of these women would you expect to have their weight between 84 and 156 lbs? If the number is not an integer, round off to the nearest integer.

First, let us compute the *Z*-statistic. Suppose *X* is the weight of a girl chosen at random, then her *Z*-statistic is $(X - 120)/18$. By the assumption that *X* is normal or approximately so, *Z* has a standard normal distribution. We want the probability $P[84 \le X \le 156]$. This is the same as $P[(84 - 120)/18 \le Z \le (156 - 120)/18] = P[-2 \le Z \le 2] = 0.9544$. See the table of the standard normal distribution. So the expected number of women would be $0.9544(100) = 95.44$ or 95 rounded to the nearest integer.

Chapter 5

2. What are the two most important properties for an estimator?

The most important properties of a point estimator are its bias and variance. These are the components of the estimator's accuracy.

3. What is the disadvantage of just providing a point estimate?

As noted in problem 2, accuracy is the most important property of an estimator and without knowledge or an estimate of the mean square error (or equivalently the bias and variance) you do not know how good the estimator is.

5. If a random sample of size *n* is taken from a population with a distribution with mean μ and standard deviation σ, what is the standard deviation (or standard error) of the sample mean equal to?

For a random sample of size *n*, the sample mean is unbiased and has a standard deviation of $\sigma/\sqrt{n}$.

8. Explain how the percentile method bootstrap confidence interval for a parameter is obtained.

To obtain a percentile method bootstrap confidence interval with confidence level $100(1-\alpha)\%$, generate many (e.g., 1000) bootstrap samples. Calculate the estimate of interest for each bootstrap sample. Order the estimates from lowest to highest. Find the first integer greater than or equal to $100\alpha/2$. Let's call that m. Then look for the mth bootstrap estimate, call it $E^*_{(m)}$. Then find the first integer greater than $100(1-\alpha/2)$. Call that integer ml. Find the mlth bootstrap estimate in the ordered list and call that $E^*_{(ml)}$. Then the interval $[E^*_{(m)}, E^*_{(ml)}]$ is a two-sided $100(1-\alpha)$ % bootstrap percentile confidence interval for the parameter being estimated.

11. The mean weight of 100 men in a particular heart study is 61 kg, with a standard deviation of 7.9 kg. Construct a 95% confidence interval for the mean.

We assume that the 100 men constitute a random sample of size 100, and that the central limit theorem will apply to the average weight. So first we compute the Z-statistic for the lower and upper bounds of the 95% confidence interval. Call the confidence interval $[L, U]$, then for the Z-scores the interval is $[(L-61)/7.9, (U-61)/7.9]$. To make this a symmetric two-sided interval, we want $(L-61)/7.9$ to be the 2.5 percentile of a standard normal random variable and $(U-61)/7.9$ to be the 97.5 percentile. From our table for the standard normal, we look for the point X with area from 0 to X equal to $0.975/2 = 0.4875$. We see that this gives a value of $X = 2.24$. So $(U-61)/7.9 = 2.24$. By symmetry, $(L-61)/7.9 = -2.24$. We can now solve for U and L. $U = 2.24(7.9) + 61 = 17.696 + 61 = 78.696$ and $L = 61 - 2.24(7.9) = 61 - 17.696 = 43.304$.

Chapter 6

2. How are equivalence tests different from standard hypothesis tests?

The standard Neyman–Pearson method for hypothesis testing make the hypothesis of no significant difference the null hypothesis with the power of the test controlled for the alternative by the necessary sample size. However, in equivalence testing, we want no significant difference to be the alternative. Some people call this "proving the null hypothesis." In the Neyman–Pearson approach, we cannot "prove the null hypothesis" without formally making it the alternative. That is because the type I error is only the probability that we reject the null hypothesis when the null hypothesis is "true." It does not control the probability of accepting the null hypothesis when the null hypothesis is "true." If the sample size is small, it is hard to reject the null hypothesis regardless of whether or not it is "true." So to control that probability, we make the null hypothesis the alternative and then controlling the power of the test controls the probability of accepting the hypothesis when it is "true," since that is precisely the definition of power. By no significant difference we mean that the difference between the two groups is in absolute value less than a defined margin of equivalence δ.

3. What is the difference between equivalence testing and noninferiority?

In equivalence testing, we require that the difference be no greater than a specified δ and no less than $-\delta$. In noninferiority testing, there is no restriction on how much larger the treatment mean is compared to the control mean, but the treatment mean minus the control mean cannot be less than $-\delta$, where δ is now called the noninferiority margin.

8. Describe the difference between a one-tailed and a two-tailed test and describe situations where one is more appropriate than the other.

Sometimes, the alternative is only interesting in one direction. For example, in clinical trials, we are usually only interested in showing that the treatment is superior to the control. This would mean that we want the average treatment effect to be statistical significantly higher than the control. If the average treatment effect is less than the average control treatment effect, it is just as bad as if there were no difference.

10. What are meta-analyses? Why might they be needed?

Meta-analyses are analyses that combine information from several studies on the same or similar endpoints. The purpose is to use the information to draw stronger conclusions about the endpoint than was possible from any individual study. The can be necessary when several small studies show trends that are not statistically significant but are all or most in the same direction. The meta-analysis may be able to provide research results that are significant rather than just a trend.

11. Based on the data in Table 6.1, do you think it is plausible that the true mean difference in temperature between New York and Washington would be 3°F? Would the power of the test be higher, lower, or the same if the true mean difference were 5°F? Does the power depend on the true mean difference? If so, why?

Yes: The observed difference is 3° or more higher in Washington versus New York in January, March, April, July, August, September, October, and November, and is 2° higher in the other 4 months (February, May, June, and December).

Assuming the variance of the difference does not change the power of the test would be higher if the true mean difference were 5° instead of 3°. This is because the greater the separation of the center of the distribution, the less the distributions overlap.

Chapter 7

1. Define the following terms:

(a) Association

(b) The correlation coefficient

(c) Simple linear regression

(d) Multiple linear regression

(e) Nonlinear regression

(f) Scatter plot

(g) Slope of the regression line in simple linear regression

(a) Association is a general term for a relationship between two variables. It includes the Pearson correlation coefficient, Kendall's tau, and Spearman's rho among others.

(b) The correlation coefficient usually refers to the Pearson product moment correlation, which is a measure of the strength of the linear association between two variables. Sometimes Kendall's tau and Spearman's rho are also called correlations. Spearman's rank correlation measures the degree to which X increases as Y increase or X decreases as Y increases. It is 1 when Y is exactly a monotonically increasing function of X, and -1 when Y is exactly a monotonically decreasing function of X.

(c) Simple linear regression is the curve relating two variables X and when $Y = af(X) + b + e$, where a and b are the parameters, and e represents a random noise component. In this formulation, Y is a linear function of the parameters a and b, and f is any function of X. The regression function is $af(X) + b$. If $f(X) = X$, Y is linear in X also, but $f(X)$ could also be $\sqrt{X}$ or X^2 or $\log(X)$.

(d) Multiple linear regression is similar to linear regression except that Y is a function of two or more variables $X_1, X_2, \ldots X_n$, where $n \geq 2$.

(e) Nonlinear regression can have Y be a function of one or more variables. It differs from linear regression in that it must be nonlinear in parameters. So, for example, Y could be $\exp(b)X^a$, or some other complicated expression. $Y = X^a + e$ is nonlinear. But if the noise term were multiplicative, that is, $Y = X^a e$, then it is transformable to a linear regression, since $\ln(Y) = \ln(e) + a \ln(X)$ In this case, we can solve by least squares with a zero intercept restriction. $\ln(e)$ is the additive noise term, and $Z = \ln(Y)$ has a linear regression $Z = aW + \delta$, where $W = \ln(X)$ and $\delta = \ln(e)$. The only parameter now is a, and Z is a linear function of the parameter a. Usually, in nonlinear regression, iterative procedures are needed for the solution, while in linear regression, the least squares solution is obtained in closed form by solving equations that are called the normal equations.

(f) A scatter plot is a graph of pairs (X, Y) that graphically shows the degree of relationship between the variables X and Y and is often the first step toward fitting a model of Y as a function of X.

(g) In simple linear regression, where $Y = af(X) + b + e$. The parameter a is called the slope of the regression line. When $f(X) = X$, the least squares regression line is fit through the scatter plot of the data. The closer the data points fall near the least squares line the higher is the correlation between X and Y, and the better the linear regression line fits the data. The slope of that regression line is the least squares estimate of a, and the Y intercept for the line is the least squares estimate of b.

5. What is logistic regression? How is it different from ordinary linear regression?

Logistic regression involves a response variable that is binary. The predictor variables can be continuous or discrete or a combination of both. Call Y the binary

variable, then the regression of Y on the predictor vector $\underline{X} = (X_1, X_2, \ldots, X_k)$ is the probability that $Y = 1$ given $\underline{X} = \underline{x}$. We let $\pi(\underline{x}) = E[Y|\underline{X} = \underline{x}]$. The logistic regression expresses $g(\underline{x})$ = the logit function of $\pi(\underline{x})$, namely $g(\underline{x}) = \ln[\pi(\underline{x})/\{1 - \pi(\underline{x})\}]$, with the function g linear in the prediction vector $\underline{x}$. Here the coefficient of each $X_i = x_i$ is β_i, $i = 1, 2, \ldots, k$. It differs from ordinary linear regression in that $E(Y|\underline{X} = \underline{x})$ is a probability belonging to the interval [0, 1], and the logit transformation is used to transform it to $(-\infty, \infty)$. It is linear like ordinary linear regression, but only after the logit transformation.

7. What is the definition of the multiple correlation coefficient R^2?

For multiple linear regression, the multiple correlation coefficient is the proportion of the variance in Y that is explained by the estimated regression equation divided by the total variance in Y. It is a measure of goodness of fit to the line, and when $R^2 = 1$, the regression equation explains all of the variance implying that all the data fall exactly on the regression line.

9. What is the equivalent to R^2 in simple linear regression?

In simple linear regression, the square of the Pearson correlation coefficient is analogous to R^2. The square of the correlation is the percentage of the variance in Y explained by the variable X. When the data fall perfectly on a line the correlation equals ± 1, and its square equals 1.

11. What is stepwise regression? Why is it used?

Stepwise regression is a procedure for selecting a subset of a set of proposed predictor variables to include in the regression model. It uses criteria to either add a variable or subtract a variable at any stage until there are no more variables satisfying the drop or add criterion. Stepwise regression is used because often in practice we know a set of variables that are related to the response variable, but we don't know how correlated they are among each other. When there is correlation, a subset of the variables will do better at predicting new values for the response than the full set of variables. This is because the estimated coefficients can be unstable when there is correlation among the variables. This is called the multicollinearity problem, because high correlation means that one of the predictor variable, say X_1, is nearly expressible as a linear combination of other predictor variables. If the multiple correlation between the variables and X_1 the regression coefficients are not unique.

14. An experiment was conducted to study the effect of increasing the dosage of a certain barbiturate. Three readings were recorded at each dose. Refer to Table 7.6.

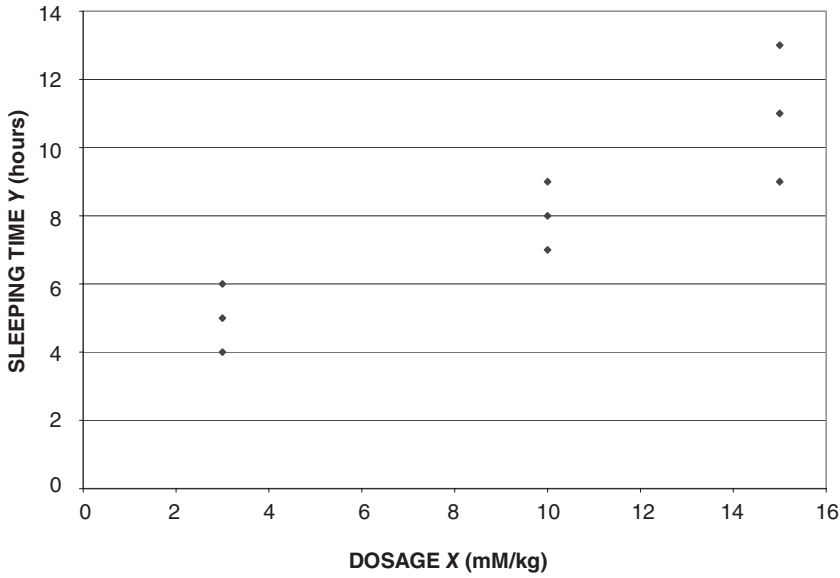
(a) Plot the scatter diagram (scatter plot)

(b) Determine by least squares the simple linear regression line relating dosage X to sleeping time Y.

(c) Provide a 95% two-sided confidence interval for the slope.

(d) Test that there is no linear relationship at the 0.05 level.

(a) Scatter plot



(b) Least squares the simple linear regression line relating dosage X to sleeping time Y.
$E(Y|X) = a + bX$, where the intercept $b = \sum_{i=1}^{n}(X_i - X^\wedge)(Y_i - Y^\wedge)/\sum_{i=1}^{n}(X_i - X^\wedge)^2 = (\sum_{i=1}^{n}\sum_{i=1}^{n} X_i Y_j - nX^\wedge Y^\wedge)/[\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j - n(X^\wedge)^2]$ and $a = Y^\wedge - bX^\wedge$. From Table 7.6 we see $\Sigma Y = 72$ so $Y^\wedge = 72/9 = 8$ and $\Sigma X = 84$ so $X^\wedge = 84/9 = 9.33$.
Now $b = [6048 - 9(72/9)(84/9)]/[7056 - 9(84/9)(84/9)] = 5376/6272 = 0.857143$, and $a = 8 - 0.857(9.33) = 0.00286$. The regression line is therefore $Y = 0{,}857X + 0.00286$.

(c) A two-sided 95% confidence interval for b is obtained by recalling that $SSE = \Sigma(Y_i - Y^\wedge)^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} Y_i Y_j - n(Y^\wedge)^2$. $SS_{y.x} = \sqrt{[SSE/(n-2)]}$ and $SE(b) = S_{y.x}\Big/\left(\sqrt{[\sum(X_i - X^\wedge)^2]}\right)$. Also, $t = (b - \beta)/SE(b)$ has a t-distribution with $n - 2$ degrees of freedom. So the degrees of freedom for this case is 7.
$SSE = 5184 - 9(64) = 4608$ and $SS_{y.x} = \sqrt{4608/7} = 25.66$ and $\mathrm{SE(b)} = 25.66\Big/\sqrt{[7056 - 9(9.33)^2]} = 25.66/79.1995 = 0.324$. Therefore, a two-sided 95% confidence interval for b is $[b - 0.324t_7(0.975), b + 0.324t_7(0.975)]$. From the t tables in the appendix, we see that $t_7(0.975) = 2.365$. So the confidence interval = [0.857 − 0.324(2.365), 0.857 + 0.324(2.365)] = [0.091, 1.62].

(d) Since 0 is not contained in the interval, we would reject the hypothesis that $\beta = 0$.

Chapter 8

2. In a survey study subjects were asked to report their health as excellent, good, poor and very poor. They were also asked to answer whether or not they had

smoked at least 250 cigarettes in their lifetime. Suppose Table 8.14 represents the outcome of the survey.

Determine if there is a relationship between cigarette usage and reported health status at the 5% significance level one-sided. What is the *p*-value for the chi-square test? Why is it appropriate to use the chi-square test?

Yes, we reject the null hypothesis at the 0.05 level. Based on SAS Version 9.2 Proc Freq, chi-square p-value < 0.0001, indicating a highly significant relationship between smoking frequency and health status. Over 70% of the patients in the excellent category smoke fewer than 250 cigarettes. Similarly, 58% of patients in the good health category smoke fewer than 250 cigarettes. In the poor health category, 53% are from the smoke fewer than 250 cigarettes. But in the very poor health category 64% are from the smoke 250 or more category

The chi square test is appropriate because the sample sizes are large and each category has at least 20 counts.

6. A clinical trial is conducted at an academic medical center. Diabetic patients were randomly assigned to a new experimental drug to control blood sugar levels versus a standard approved drug using a 1 : 1 randomization. 200 patients were assigned to each group and the 2 × 2 table (Table 8.18) shows the results.

Test at the 5% level to determine if the new drug is more effective. Is it appropriate to apply the chi-square test? Why would it be difficult to do Fisher's test without a computer? How many contingency tables are possible with the given row and column marginal totals?

Based on both the chi-square test and Fisher's exact test, we see that the drug is very effective. p-value for both test is much less than 0.0001. The chi-square test s appropriate because each cell has at least 21 patients in it. Fisher's test would be difficult to do by hand because there are many contingency tables to look at. But using SAS 9.2, this is not really a problem. There are 141 such tables with the fixed marginal totals.

Chapter 9

2. Apply the Wilcoxon rank-sum test to the data in the following table on the relationship between the number of patients with schizophrenia and the season of their birth by calling fall and winter as group 1 and spring and summer as group 2. The four individual seasons represent data points for each group. Ignore the possibility of a year effect (Table 9.6).

Do we need to assume that births are uniformly distributed? If we knew that there were a higher percentage of births in the winter months how would that affect the conclusion?

The ordered data and ranks are as follows:

9—summer	1
10—spring	2
13—spring	3
14—spring	4
15—summer	5

17—summer	6
18—spring	7
19—fall	8
20—fall	9.5
20—spring	9.5
21—summer	11
22—spring	12
23—fall	13
25—summer	14
27—fall	15
28—summer	16
32—fall	17
33—fall	18
35—winter	19
36—winter	20.5
36—winter	20.5
38—winter	22
41—winter	23
43—winter	24

By groups we have ranks and sum of ranks as follows:

	Group 1 (fall and winter)	Group 2 (spring and summer)
	8	1
	9.5	2
	13	3
	15	4
	17	5
	18	6
	19	7
	20.5	9.5
	20.5	11
	22	12
	23	14
	24	16
Rank sum	209.5	90.5

It appears obvious that group 1 has the larger numbers.
From the table for the Wilcoxon rank-sum test when $n1 = n2 = 12$ we have for $T = 209.50$. The p-value given here was obtained using the following SAS code:

```
data schizo;
input group$ season$ nschizo
;
datalines;
g2 summer 9
g2 summer 15
g2 summer 17
g2 summer 21
g2 summer 25
g2 summer 28
g2 spring 10
g2 spring 13
g2 spring 14
g2 spring 18
g2 spring 20
g2 spring 22
g1 fall 19
g1 fall 20
g1 fall 23
g1 fall 27
g1 fall 32
g1 fall 33
g1 winter 35
g1 winter 36
g1 winter 36
g1 winter 38
g1 winter 41
g1 winter 43
;
run;

ods graphics on;
proc npar1way data = schizo;
class group;
var nschizo;
run;

ods graphics off;
data schizo;
input group$ season$ nschizo
;
datalines;
```

```
g2 summer 9
g2 summer 15
g2 summer 17
g2 summer 21
g2 summer 25
g2 summer 28
g2 spring 10
g2 spring 13
g2 spring 14
g2 spring 18
g2 spring 20
g2 spring 22
g1 fall 19
g1 fall 20
g1 fall 23
g1 fall 27
g1 fall 32
g1 fall 33
g1 winter 35
g1 winter 36
g1 winter 36
g1 winter 38
g1 winter 41
g1 winter 43
;
run;
ods graphics on;


proc npar1way data = schizo;
class group;
var nschizo;
run;

ods graphics off;
```

The result is included in the following SAS output:

```
The SAS System           12:10 Monday, December 6, 2010
2

                The NPAR1WAY Procedure

  Wilcoxon Scores (Rank Sums) for Variable nschizo
           Classified by Variable group
```

```
 Sum of  Expected  Std Dev  Mean
group N Scores  Under H0  Under H0  Score
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
g2  12  90.50  150.0  17.312976  7.541667
g1  12  209.50  150.0  17.312976  17.458333

  Average scores were used for ties.


    Wilcoxon Two-Sample Test

    Statistic  90.5000

    Normal Approximation
    Z  -3.4078
    One-Sided Pr < Z  0.0003
    Two-Sided Pr > |Z|  0.0007

    t Approximation
    One-Sided Pr < Z  0.0012
    Two-Sided Pr > |Z|  0.0024

   Z includes a continuity correction of 0.5.


    Kruskal-Wallis Test

    Chi-Square 11.8111
    DF   1
    Pr > Chi-Square  0.0006
```

Yes, we do need births to be at least uniform over the seasons. This is because we are ranking based on number of births over a season. If there tended to be more births in the fall and winter compared with summer and spring, the higher number of schizophrenic patients could be due to the higher number of total births rather than a tendency for schizophrenics to be born during winter and fall. If there were a higher number of births in the winter, then we could not reach our intended conclusion. We would need to know the number of births in each season and adjust accordingly by looking at proportion of schizophrenic births rather than the total number.

6. Using Table 9.9, compute the Spearman rank correlation coefficient for the aggressiveness scores. Does this suggest that both twins tend to be similar in degree of aggressiveness?

Recall that the formula for Spearman's rank correlation is given as follows:

$$\rho = \frac{\left\{\sum_{i=1}^{n} R(X_i)R(Y_i) - n([n+1]/2)^2\right\}}{\left\{\sum_{i=1}^{n} R(X_i)^2 - n([n+1]/2)^2\right\}\left\{\sum_{i=1}^{n} R(Y_i)^2 - n([n+1]/2)^2\right\}}.$$

In the case of no ties, this formula simplifies to

$$\rho = 1 - 6T/[n(n^2-1)\},$$

where $T = \sum_{i=1}^{n}[R(X_i) - R(Y_i)]^2$.

Since we have ties, we cannot use the simplified formula.

So in this example, the estimate

$$\begin{aligned}\rho &= \{\textstyle\sum_{i=1}^{12} R(X_i)R(Y_i) - 507\}/[\{\sum_{i=1}^{12} R(X_i)^2 - 507\}\{\sum_{i=1}^{12} R(Y_i)^2 - 507\}] \\ &= (535.0 - 507.0)/(649.0 - 507.0)(649.0 - 507.0) \\ &= 28/(142)^2 = 28/20{,}164 = 0.001389.\end{aligned}$$

The correlation is negligible, meaning there is practically no relationship between order of birth and aggressiveness of the twins. A statistical test would show that the correlation is not statistically significantly different from 0. So the twins tend to be similar in the amount of aggressiveness shown in their scores.

Chapter 10

2. If the survival function $S(t) = 1 - t/b$ for $0 \le t \le b$, where b is a fixed positive constant, and $S(t) = 0$ for $t > b$, calculate the hazard function. When is the hazard function lowest? Is there a highest rate?

$F(t) = 1 - S(t) = 1 - (1 - t/b) = t/b$ for $0 \le t \le b$ in this case, and $F(t) = 1$ for $t > b$. Now $f(t) = dF(t)/dt = 1/b$ for all $0 \le t \le b$ and $f(t) = 0$ otherwise. Now the hazard function is defined as $h(t) = f(t)/S(t) = 1/[bS(t)] = 1/b(1 - t/b) = 1/(b - t)$ for $0 \le t \le \underline{b}$. At $t = 0$, the hazard function is $1/b$, and that is its lowest value. For $b > t > 0 h(t) = 1/(b - t) > 1/b$, since $b > b - t$. So the hazard function is increasing. But as $t \to b$, $b - t \to 0$, and $h(t) = 1/(b - t) \to \infty$. So there is no maximum value for the hazard function.

4. Suppose the survival time since a bone marrow transplant for eight patients who received the transplant is 3, 4.5, 6, 11, 18.5 20, 26, and 35. No observations were censored.

(a) What is the median survival time for these patients?

(b) What is the mean survival time?

(c) Construct a life table where each interval is 5 months.

(a) Since the data is complete and there are 8 event times, the median is the average of the 4th and 5th ordered observations, which is (11 + 18.5)/2 = 14.75 months.
(b) The mean survival is just the arithmetic average of the eight event times, which is (3 + 4.5 + 6 + 11 + 18.5 + 20 + 26 + 35)/8 = 15.5 months.
(c) A life table for this data using 5-month intervals is as follows:

Time Interval I_j	No. of deaths in I_j	No. withdrawn in I_j	No.at risk in I_j	Avg. no. at risk in I_j	Est. prop. of deaths in I_j	Est. prop. Surv. at end of I_j	Est. cum. surv. at end of I_j
[0, 5)	2	0	8	8	0.25	0.75	0.75
[5, 10)	1	0	6	6	0.167	0.833	0.585
[10 15)	1	0	5	5	0.200	.8	0.468
[15,20)	1	0	4	4	0.250	0.750	0.341
[20,25)	1	0	3	3	0.333	0.667	0.227
[25,30)	1	0	2	2	0.500	0.500	0.114
[30,35)	0	0	2	2	0.000	1.000	0.114
[35,40)	1	0	1	1	1.000	0.000	0.000
[40, ∞)	0	0	0	0	—	—	—

10. Suppose a cure model is known to have $S_1(t) = \exp(-0.5t)$.

Recall $S(t) = p + (1 - p)\, S_1(t)$. Suppose that we know that $S(2) = 0.5259$. Can you calculate the cure rate for this model? If so what is it?

$S(t) = p + (1 - p)S_1(t)$. Since we know $S_1(t) = \exp(-0.5t)$, we only need to determine $pS(t)$.
We are given $S(2) = 0.5259$, and we can use this information to solve for p.
$S(2) = 0.5259$, on the one hand, and

$$S(2) = p + (1-p)\exp[-(0.5)^2] = p + (1-p)(0.3679)$$
$$= p(1-0.3679) + 0.3679.$$

$$\text{So } p(1-0.3679) = 0.5259 - 0.3679 = 0.1580.$$

$$P = 0.1580/(1-0.3679) = 0.1580/0.6321 = 0.4295$$

So $S(t) = 0.4295 + 0.5705\exp(-0.5t)$, and the probability of cure is 0.4295. This means that approximately 43% of the patients receiving this treatment will be cured of the disease based on this model.

APPENDIX

Statistical Tables

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

Table 1
Percentage Points, *F*-Distribution ($\alpha = 0.05$)

m / n	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34		2.19	2.15	2.11	2.06	2.02	1.97	1.92

(*Continued*)

Table 1
(Continued)

m / n	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	1.87	1.81	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92		
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

SOURCE: Beyer, William H., ed. (1966). *Handbook of Tables for Probability and Statistics*. Cleveland, Ohio: The Chemical Rubber Co., p. 242.
Taken from Chernick and Friis (2003), Appendix A, p. 363, with permission.

Table 2
Studentized Range Statistic

	Upper 5% points																		
n / v	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07	50.59	51.96	53.20	54.33	55.36	56.32	57.22	58.04	58.83	59.56
2	6.08	8.33	9.80	10.88	11.74	12.44	13.03	13.54	13.99	14.39	14.75	15.08	15.38	15.65	15.91	16.14	16.37	16.57	16.77
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	9.72	9.95	10.15	10.35	10.53	10.69	10.84	10.98	11.11	11.24
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

(*Continued*)

Table 2
(Continued)

	Upper 5% points																		
n / v	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
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

SOURCE: Beyer, William H., ed. (1966). *Handbook of Tables for Probability and Statistics*. Cleveland, Ohio: The Chemical Rubber Co., p. 286.
Taken from Chernick and Friis (2003), Appendix B, pp. 364–365, with permission.

Table 3
Quantiles of the Wilcoxon Signed-Rank Test Statistic

	$W_{0.005}$	$W_{0.01}$	$W_{0.025}$	$W_{0.05}$	$W_{0.10}$	$W_{0.20}$	$W_{0.30}$	$W_{0.40}$	$W_{0.50}$	$\frac{n(n+1)}{2}$
$n = 4$	0	0	0	0	I	3	3	4	5	10
5	0	0	0	1	3	4	5	6	7.5	15
6	0	0	1	3	4	6	8	9	10.5	21
7	0	1	3	4	6	9	11	12	14	28
8	1	2	4	6	9	12	14	16	18	36
9	2	4	6	9	11	15	18	20	22.5	45
10	4	6	9	11	15	19	22	25	27.5	55
11	6	8	11	14	18	23	27	30	33	66
12	8	10	14	18	22	28	32	36	39	78
13	10	13	18	22	27	33	38	42	45.5	91
14	13	16	22	26	32	39	44	48	52.5	105
15	16	20	26	31	37	45	51	55	60	120
16	20	24	30	36	43	51	58	63	68	136
17	24	28	35	42	49	58	65	71	76.5	153
18	28	33	41	48	56	66	73	80	85.5	171

(*Continued*)

Table 3
(Continued)

	$W_{0.005}$	$W_{0.01}$	$W_{0.025}$	$W_{0.05}$	$W_{0.10}$	$W_{0.20}$	$W_{0.30}$	$W_{0.40}$	$W_{0.50}$	$\frac{n(n+1)}{2}$
19	33	38	47	54	63	74	82	89	95	190
20	38	44	53	61	70	83	91	98	105	210
21	44	50	59	68	78	91	100	108	115.5	231
22	49	56	67	76	87	100	110	119	126.5	253
23	55	63	74	84	95	110	120	130	138	276
24	62	70	82	92	105	120	131	141	150	300
25	69	77	90	101	114	131	143	153	162.5	325
26	76	85	99	111	125	142	155	165	175.5	351
27	84	94	108	120	135	154	167	178	189	378
28	92	102	117	131	146	166	180	192	203	406
29	101	111	127	141	158	178	193	206	217.5	435
30	110	121	138	152	170	191	207	220	232.5	465
31	119	131	148	164	182	205	221	235	248	496
32	129	141	160	176	195	219	236	250	264	528
33	139	152	171	188	208	233	251	266	280.5	561
34	149	163	183	201	222	248	266	282	297.5	595
35	160	175	196	214	236	263	283	299	315	630

36	172	187	209	228	251	279	299	317	333	666
37	184	199	222	242	266	295	316	335	351.5	703
38	196	212	236	257	282	312	334	353	370.5	741
39	208	225	250	272	298	329	352	372	390	780
40	221	239	265	287	314	347	371	391	410	820
41	235	253	280	303	331	365	390	411	430.5	861
42	248	267	295	320	349	384	409	431	451.5	903
43	263	282	311	337	366	403	429	452	473	946
44	277	297	328	354	385	422	450	473	495	990
45	292	313	344	372	403	442	471	495	517.5	1035
46	308	329	362	390	423	463	492	517	540.5	1081
47	324	346	379	408	442	484	514	540	564	1128
48	340	363	397	428	463	505	536	563	588	1176
49	357	381	416	447	483	527	559	587	612.5	1225
50	374	398	435	467	504	550	583	611	637.5	1275

SOURCE: Conover, W. J. (1999). *Practical Nonparametric Statistics.* 3rd ed. New York: Wiley, pp. 545–546.
Taken from Chernick and Friis (2003), Appendix C, pp. 366–367, with permission.

Table 4
χ^2 Distribution

df	*p*						
	0.99	0.95	0.90	0.10	0.05	0.01	0.001
1	0.0^3157	0.00393	0.0158	2.706	3.841	6.635	10.827
2	0.0201	0.103	0.211	4.605	5.991	9.210	13.815
3	0.115	0.352	0.584	6.251	7.815	11.345	16.266
4	0.297	0.711	1.064	7.779	9.488	13.277	18.467
5	0.554	1.145	1.610	9.236	11.070	15.086	20.515
6	0.872	1.635	2.204	10.645	12.592	16.812	22.457
7	1.239	2.167	2.833	12.017	14.067	18.475	24.322
8	1.646	2.733	3.490	13.362	15.507	20.090	26.125
9	2.088	3.325	4.168	14.684	16.919	21.666	27.877
10	2.558	3.940	4.865	15.987	18.307	23.209	29.588
11	3.053	4.575	5.578	17.275	19.675	24.725	31.264
12	3.571	5.226	6.304	18.549	21.026	26.217	32.909
13	4.107	5.892	7.042	19.812	22.362	27.688	34.528
14	4.660	6.571	7.790	21.064	23.685	29.141	36.123
15	5.229	7.261	8.547	22.307	24.996	30.578	37.697
16	5.812	7.962	9.312	23.542	26.296	32.000	39.252
17	6.408	8.672	10.085	24.769	27.587	33.409	40.790
18	7.015	9.390	10.865	25.989	28.869	34.805	42.312
19	7.633	10.117	11.651	27.204	30.144	36.191	43.820
20	8.260	10.851	12.443	28.412	31.410	37.566	45.315
21	8.897	11.591	13.240	29.615	32.671	38.932	46.797
22	9.542	12.338	14.041	30.813	33.924	40.289	48.268
23	10.196	13.091	14.848	32.007	35.172	41.638	49.728
24	10.856	13.848	15.659	33.196	36.415	42.980	51.179
25	11.524	14.611	16.473	34.382	37.652	44.314	52.620
26	12.198	15.379	17.292	35.563	38.885	45.642	54.052
27	12.879	16.151	18.114	36.741	40.113	46.963	55.476
28	13.565	16.928	18.939	37.916	41.337	48.278	56.893
29	14.256	17.708	19.768	39.087	42.557	49.588	58.302
30	14.953	18.493	20.599	40.256	43.773	50.892	59.703

SOURCE: Adapted from Table IV of R. A. Fisher and F. Yates (1974). *Statistical Tables for Biological, Agricultural, and Medical Research.* 6th ed. London: Longman Group, Ltd. (Previously published by Oliver & Boyd, Ltd., Edinburgh). Used with permission of the authors and publishers.

Taken from Chernick and Friis (2003), Appendix D, pp. 368–369, with permission.

Table 5
The Standard Normal Distribution

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

(*Continued*)

Table 3
(Continued)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.5	0.4332	0.4345	0.4350	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

SOURCE: Public domain.
Taken from Chernick and Friis (2003), Appendix E, p. 370, with permission.

Table 6
Percentage Points, Student's *t*-Distribution

F / n	0.90	0.95	0.975	0.99	0.995
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	2.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

SOURCE: Beyer, William H., ed. (1966). *Handbook of Tables for Probability and Statistics.* Cleveland, Ohio: The Chemical Rubber Co., p. 226.
Taken from Chernick and Friis (2003), Appendix F, pp. 371–372, with permission.

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Agresti, A. (2002). *Categorical Data Analysis*. 2nd ed. New York: Wiley.

Altman, D. G. (1991). *Practical Statistics for Medical Research*. London: Chapman & Hall.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics*. New York: Wiley.

Berkson, J. and Gage, R. P. (1952). Survival curves for cancer patients following treatment. *Journal of the American Statistical Association*, **47**, 501–515.

Best, J. (2001). *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians and Activists*. Berkeley, CA: University of California Press.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester: Wiley.

Broemeling, L. D. (2007). *Bayesian Biostatistics and Diagnostic Medicine*. Boca Raton, FL: Chapman & Hall/CRC.

Campbell, S. K. (1974). *Flaws and Fallacies in Statistical Thinking*. Englewood Cliffs, NJ: Prentice Hall.

Campbell, M. J. and Machin, D. (1999). *Medical Statistics: A Commonsense Approach*. 3rd ed. Chichester: Wiley.

Chernick, M. R. (2007). *Bootstrap Methods: A Guide for Practitioners and Researchers*. 2nd ed. Hoboken, NJ: Wiley.

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

Chernick, M. R. and Friis, R. H. (2003). *Introductory Biostatistics for the Health Sciences; Modern Applications Including Bootstrap*. Hoboken, NJ: Wiley.

Chernick, M. R., Poulsen, E. G., and Wang, Y. (2002). Effects of bias adjustment on actuarial survival curves. *Drug Information Journal*, **36**, 595–609.

Cleves, M., Gould, W., Gutierrez, R., and Marchenko, Y. (2008). *An Introduction to Survival Analysis Using Stata*. 2nd ed. College Station, TX: Stata Press.

Conover, W. J. (1999). *Practical Nonparametric Statistics*. 3rd ed. New York: Wiley.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. New York: Chapman & Hall.

Cutler, S. J. and Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival. *Journal of Chronic Diseases*, **8**, 699–712.

Dorey, F. J. and Korn, E. L. (1987). Effective sample sizes for confidence intervals for survival probabilities. *Statistics in Medicine*, **6**, 679–687.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. 3rd ed. New York: Wiley.

Efron, B. and Tibshirani, R. J. (1993). *Introduction to the Bootstrap*. New York: Chapman & Hall.

Egger, M., Smith, G. D., and Altman, D., eds. (2001). *Systematic Reviews in Health Care: Meta-Analysis in Context*. 2nd ed. London: BMJ Publishing Group.

Fisher, R. A. (1935). *Design of Experiments*. London: Oliver and Boyd.

Friis, R. H. and Sellers, T. A. (1999). *Epidemiology for Public Health Practice*. 2nd ed. Gaithersburg, MD: Aspen.

Gallant, A. R. (1987). *Nonlinear Statistical Methods*. New York: Wiley.

Gönen, M. (2007). *Analyzing Receiver Operating Characteristic Curves with SAS*. Cary, NC: SAS Press.

Gonick, L. and Smith, W. (1993). *The Cartoon Guide to Statistics*. New York: Harper Perennial.

Good, P. I. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer-Verlag.

Hand, D. J. (2008). *Statistics*. New York: Sterling.

Hand, D. J. and Crowder, M. J. (1996). *Practical Longitudinal Data Analysis*. London: Chapman & Hall Ltd.

Hardin, J. W. and Hilbe, J. M. (2003). *Generalized Estimating Equations*. 2nd ed. London: Chapman & Hall/CRC Press.

Hartung, J., Knapp, G., and Sinha, B. K. (2008). *Statistical Meta-Analysis with Applications*. Hoboken, NJ: Wiley.

Hedges, L. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

Hernán, M., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, **11**, 561–570.

Hernán, M. A., Cole, S., Margolick, J., Cohen, M., and Robins, J. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety* (published online 19 Jan 2005).

Higgins, J. P. T. and Green, S., eds. (2008). *Cochrane Handbook of Systematic Reviews of Interventions*. Chichester: Wiley.

Hilbe, J. M. (2009). *Logistic Regression Models*. Boca Raton, FL: Chapman & Hall/CRC Taylor and Francis.

Hosmer, D. W. and Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: Wiley.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd ed. New York: Wiley.

Hosmer, D. W., Lemeshow, S., and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. 2nd ed. Hoboken, NJ: Wiley.

Huber, P. (1981). *Robust Statistics*. New York: Wiley.

Huff, D. (1954). *How to Lie with Statistics*. New York: W. W. Norton and Company.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag.

Irwin, J. O. (1935). Tests of significance for the differences between percentages based on small numbers. *Metron*, **12**, 83–94.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd ed. New York: Wiley.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**, 81–93.

Klein, J. P. and Moeschberger, M. L. (2003, paperback 2010). *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer-Verlag.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*. Boca Raton, FL: Chapman & Hall/CRC.

Kuzma, J. (1998). *Basic Statistics for the Health Sciences*. Mountain View, CA: Mayfield Publishing Company.

Lachin, J. M. (2000). *Biostatistical Methods: The Assessment of Relative Risks*. New York: Wiley.

Lachin, J. M. (2011). *Biostatistical Methods: The Assessment of Relative Risks*. 2nd ed. Hoboken, NJ: Wiley.

Lee, E. T. (1992). *Statistical Methods for Survival Data Analysis*. 2nd ed. New York: Wiley.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis of Missing Data*. 2nd ed. New York: Wiley.

Lloyd, C. J. (1999). *Statistical Analysis of Categorical Data*. New York: Wiley.

Maller, R. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. New York: Wiley.

Mantel, N. (1966). *Evaluation of su*rvival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, **50**, 163–170.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Hoboken, NJ: Wiley.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

O'Quigley, J. (2008). *Proportional Hazards Regression*. New York: Springer Science+Business Media, LLC.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, U.K.: Cambridge University Press.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge, U.K.: Cambridge University Press.

Pepe, M. S. (2004). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press.

Riffenburgh, R. H. (1999). *Statistics in Medicine*. San Diego, CA: Academic Press.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. *Statistical Models in Epidemiology: The Environment and Clinical Trials*. IMA Volume **116**. M. E. Halloran and D. Berry, eds. New York: Springer-Verlag, pp. 95–134.

Rothstein, H. R., Sutton, A. J., and Borenstein, M., eds. (2005). *Publication Bias in Meta-Analysis*. Chichester: Wiley.

Rubin, D. B. (2006). *Matched Sampling by Causal Effects*. Cambridge, U.K.: Cambridge University Press.

Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York: W. H. Freeman and Company.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, **15**, 72–101.

Stangl, D. and Berry, D., eds. (2000). *Meta-Analysis in Medicine and Health Policy*. New York: Dekker.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer Science+Business Media, LLC.

Upton, G. and Cook, I. (2002). *Oxford Dictionary of Statistics*. Oxford: Oxford University Press.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer-Verlag.

van der Laan, M. J. and Robins, J. M. (2010). *Unified Methods for Censored Longitudinal Data and Causality*. 2nd ed. New York: Springer-Verlag.

Verbeke, G. and Molenberghs, G., eds. (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach*. New York: Springer-Verlag.

Walker, G. A. and Shostak, J. (2010). *Common Statistical Methods for Clinical Research with SAS® Examples*. 3rd ed. Cary, NC: SAS Press.

Whitehead, A. (2002). *Meta-Analysis of Controlled Clinical Trials*. Chichester: Wiley.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society*, Suppl. 1, 217–235.

Zhou, X.-H., McClish, D. K., and Obuchowski, N. A. (2002). *Statistical Methods in Diagnostic Medicine*. Hoboken, NJ: Wiley.

Author Index

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.

Subject Index

The Essentials of Biostatistics for Physicians, Nurses, and Clinicians,
First Edition. Michael R. Chernick.