Linda Herkenhoff
John Fogli

Applied Statistics for Business and Management using Microsoft Excel

Springer

Applied Statistics for Business and Management using Microsoft Excel

Linda Herkenhoff • John Fogli

Applied Statistics for Business and Management using Microsoft Excel

Springer

Linda Herkenhoff
Saint Mary's College
Orinda, California
USA

John Fogli
Sentenium Inc.
Pleasant Hill, California
USA

ISBN 978-1-4614-8422-6 ISBN 978-1-4614-8423-3 (eBook)
DOI 10.1007/978-1-4614-8423-3
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013946938

Printed on acid-free paper

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

Preface

Mathematical formulas and Greek letters seem to intimidate many people, so this book focuses on providing clear word descriptions and step-by-step Excel instructions, rather than including lots of x's and y's everywhere. We have specifically avoided including a lot of technical information. This book was written for those who want basic understanding of how to use statistics in the business world without all the details of statistical theory. It is designed to enhance a practical learning of statistics by nonstatisticians and to make the experience a little easier and maybe even fun.

This book has also been written with managers in mind who may need a quick refresher on how to complete Excel analysis without having to wade through pages of technical explanations. Those more detailed technical explanations can be found in traditional statistics textbooks. The key concepts in this book are more complex than the presentations in this book might suggest. We have aimed at concisely defining the concepts and providing simple descriptions and explanations on how to use them.

For instructors, this publication is designed as a companion guide to your core text. Students can practice their skills that have been introduced to them in the classroom. However, at the beginning of each chapter, the key concepts are summarized to remind them of what is important with the topic.

We believe that problems and case studies should involve actual data from the business community. Therefore, there is a continuing analysis throughout several of the chapters using a real business case from Infinity Auto Insurance. The data sets for the other practice problems in each chapter have been kept intentionally small so that the user can more easily maneuver within the data. The rigorous sample size rules have been put aside to accommodate this design goal. The same Excel procedures apply to all sizes of data sets, small or large.

This book was written using Excel 2010; the screenshots may appear slightly different if you are using an older version of Excel or if you are using a Mac. If you are comfortable in Excel, these differences should be minor.

We have included only those analytic Excel tools that are most commonly used in the workplace to keep things simple. This approach should work well for those having to run their own analysis, as well as for those needing to better understand statistical reports, so as to make informed business decisions.

No prior Excel experience is necessary, although it may prove helpful. Instructions are provided with the mindset that this is the first time you have used this particular Excel functionality. In other words, each chapter provides simple instructions for first-time users.

Some of the traditional statistical texts include Excel instructions. But these are typically embedded deep in the chapter as one of several different software instruction packets. This book focuses only on Excel and provides easy-to-find Excel instructions in each chapter.

In addition to offering this book in traditional textbook form, it is also available as an e-book in recognition of being environmentally responsible.

The authors bring a unique perspective as guides through the World of Statistics. Their combined skills and expertise in the workplace and the classroom equip them to understand the challenges students and managers face as they navigate the World of Statistics. Their active consulting and teaching lives keep them up to date on the common real-world issues in understanding and applying statistical concepts.

We have worked diligently to make sure that this book is accurate and error free. But if you happen to find any errors or have suggestions for improvements, please contact us at jfogli@hotmail.com or lherkenh@stmarys-ca.edu.

Many of you have survived statistics courses through pure memorization, but hopefully this book will encourage both logical reasoning and intuitive understanding. We believe that logical reasoning is fundamental to retaining analytical skills that can be practiced in the real world. Enjoy. . . .

Orinda, CA, USA Linda Herkenhoff
Pleasant Hill, CA, USA John Fogli

Acknowledgments

The foundation of this book was laid with numerous statistical discussions on the tennis court. Thanks to Dr. Linda Herkenhoff for convincing me to work with her on this idea that finally resulted into the first edition of our book. Thank you to our reviewers as well as several Diablo Valley College students in Pleasant Hill, California. A big thank you to Eva Meng from Sentenium, Inc., who spent hours helping with screenshots and editing this text. Finally, thanks to my parents and especially to my wife who stood beside me to take this book forward.

Pleasant Hill, CA, USA John Fogli

We would like to thank several of our colleagues who tirelessly reviewed each and every word and number in our book: Dr. Michael Wagner, Dr. Diana Borova, Margery Connor, and Eva Meng. Your fastidious reviews let us sleep at night. But additionally we had even more colleagues who generously reviewed various sections of the book, and to them also we owe our thanks: Dr. Peter Hosie, Dr. Madeline Wu, Dr. Jo Ann Heydenfeldt, Dr. Brian Sholey, and Dr. Rodolfo Martell.

We express appreciation to all of our past Statistics students who allowed us to better understand the challenges in bringing Statistics alive as a powerful business tool. This book reflects our many years of student input and observation both in the United States and abroad.

Special recognition is paid to Infinity Life Insurance Company for allowing us to use their data for our real-world case study.

Thanks to my husband Frederic and three sons Brett, Kyle and Eric for their patience and support throughout the writing process.

Orinda, CA, USA Linda Herkenhoff

We have done our best to make this book accessible online for comments and suggestions. In addition, we have made it convenient for students and professionals to download our data files available at http://www.sentenium.com/springer.html. These data files are used in various practice problems at the end of certain chapters. Our intention is to give you practice with data analysis and interpretation of results using Microsoft Excel.
We welcome comments about the book and suggestions for improvement.
Email:lherkenh@stmarys-ca.edu

Contents

Chapter 1
Data and Statistics

Case Study: Exploring New Marketing Strategies

INFINITY®
Auto Insurance

Infinity Property and Casualty Corporation (Infinity Auto Insurance) headquartered in Birmingham, Alabama provides personal automobile insurance with a concentration on nonstandard auto insurance. Nonstandard insurance serves individuals unable to obtain coverage through standard insurance companies, which can be due to a driving record with accidents and/or tickets, prior DUI, the driver's age, vehicle type, etc. Infinity Auto Insurance's products include personal automobile insurance for individuals, commercial vehicle insurance for businesses and classic collector insurance for individuals with classic and antique automobiles. Infinity Auto Insurance distributes its products primarily through independent agencies and brokers. Infinity Auto Insurance, a top-performing Infinity brand, provides nonstandard car insurance through more than 12,500 independent agents. Infinity Auto Insurance utilizes Internet-based software applications to provide many of its agents with real-time underwriting, claims and policy information. The Company is licensed to write insurance in all 50 states and the District of Columbia. The company traces its roots back to 1955 doing business as Dixie Insurance company, later known as The Infinity Group. In 1991, Pennsylvania Company (American Premier Underwriters) purchases The Infinity Group. American Premier Underwriters, Inc. and the American Financial Corporation merge and become the American Financial Group (AFG) in 1995.

(continued)

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_1,

Case Study: (continued)

The company was officially founded in 2002 when AFG transferred all common stock of Infinity and its sister companies to IPCC. In 2003, IPCC was listed on the NASDAQ as a registered public holding.

Food for Thought

Who are some of Infinity Auto Insurance's direct auto insurance competitors?
How large (annual revenue) is the Auto Insurance industry?
Discuss what type of customers Infinity Auto Insurance might target?
If you were a marketing consultant to Infinity Auto Insurance, what market areas would you suggest to target to access the customers identified above?

Possible Answers

Who are some of Infinity Auto Insurance's direct auto insurance competitors?
GEICO, Progressive, Mercury, AIG

How large (annual revenue) is the Auto Insurance industry?
Annual revenue for the Auto Insurance industry ranges in billions of dollars every year.

Discuss what type of customers Infinity Auto Insurance might target?
High risk auto insurance individuals

If you were a marketing consultant to Infinity Auto Insurance, what market areas would you suggest to target to access the customers identified above?
Large metropolitan areas or lower income areas

Key Concepts

Business statistics, Constant, Continuous data, Cross sectional data, Data, Descriptive statistics, Discrete data, Distribution, Frequency, Inferential statistics, Longitudinal data, Population, Probability, Qualitative data, Quantitative data, Random sample, Sample, Scales, Statistics, and Variables.

Discussion

This chapter will include some of the basic concepts and definitions necessary in understanding and applying statistical methods in business. One of the challenges in statistical analysis and subsequent report writing is the confusion over what words and statistical terms really mean. In statistics the mathematical meaning may not match up with the everyday usage of the same term, which can lead to confusion. For example in our daily communications the term "significant" means important i.e. Steve's attendance, as the current CEO, at the award ceremony was very significant to the press. In statistics when a statistic is "significant", it simply means that you are very sure that the statistic is reliable or real and not just due to chance. It doesn't mean the finding is important nor that it has any decision-making utility.

Statistics is often said to be more open to misuse, both deliberately and naively, than any other area in business. Part of that misuse may stem from inconsistency in statistical definitions. The next few pages will attempt to firmly establish the lexicon for moving into the next chapters, but each chapter will also provide content specific definitions. It is in your best interest to take the time to read thought these terms and not assume you have a complete and fully accurate understanding of these key concepts.

Business statistics: Business statistics include the areas of descriptive statistics, probability statistics, and inferential statistics applied to business.

Constant: A value that remains unchanged. For example let **C** be a constant, **C** = number of years of university education required for all entry level accountants in your company and in this case **C** = 4 years.

Cross-sectional data: Cross-sectional data is data that are collected from participants at one point in time (rule of thumb is usually within six or less months). Time is not considered one of the study variables in a cross-sectional research design. In a cross-sectional study, time is assumed to be a random effect that produces only variance, not bias. For example the number of IPOs in the San Francisco area.

Data: Data is specific values of the variables if interest that are collected, analyzed and summarized for presentation and interpretation.

Descriptive statistics: The statistics associated with organizing, summarizing and presenting data in a convenient and informative way. Both graphical and numerical techniques are employed.

Distribution: an arrangement of values of a variable showing their observed or theoretical frequency of occurrence.

Frequency: How often a value occurs.

Inferential statistics: The statistics associated with making an estimate, prediction, or decision from a small group to draw conclusions about a larger group. For example in a random sample we note that 40 % of the voters do not want an increase in business taxes. If the sample has been selected without bias, we can **infer** the population would also vote about 40 % against an increase in business taxes.

Longitudinal data or time series data: Data collected over several time periods. For example the number of IPOs in the San Jose area over the past 5 years. In this example the time period would be defined as 1 year and we are collecting data over 5 of these time periods, or years. Note the term longitudinal has nothing to do with geo-spatial data on the globe.

Population: All the existing individuals, items, or data about which you want to draw a conclusion. For example all of the MBA graduates in the USA, or all the copy machines owned by your company in the past 5 years.

Probability: The analysis associated with the chance of occurrence of one or more possible results of an unpredictable event. When potential students apply to an MBA school, they consider the chance or probability of being accepted.

Qualitative data: Categorical, non-numeric data such as gender, industry type. There is no mathematical relationship between the data values. The data labels can be numeric or non-numeric, and are descriptors of the data. For example the labels may be True/False or could be shown as 1/2. But there is no mathematical relationship between 1 and 2 in this case, they are just labels. In other words 2 is not twice as big as 1.

Quantitative data: Numerical data such as salaries, or sales. There is a mathematical relationship between the data values. The data labels must always be numeric. However quantitative data can be either discrete or continuous. **Continuous** quantitative data arise from a measurement process. Continuous data is information that can be measured on a continuum or scale. Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments, depending upon the precision of the measurement system. For example **time** may be considered as continuous quantitative data. **Discrete** quantitative data arise from a counting process. An example includes how many text messages you sent this past week.

Random sample: Data selected from a population in a way that ensures each data value has an equal opportunity of being selected.

Sample: A portion of a population selected for analysis.

Scales (measurement): A scale may be defined as any set of items which is progressively arranged according to value or magnitude into which an item can be placed according to its quantification. In other words a scale is a continuous spectrum or series of categories. The purpose of scaling is to represent, usually quantitatively, an item's, a person's or an event's place in the spectrum. The four types of scales are nominal, ordinal, interval and ratio.

- **Nominal scale**
  A scale in which the numbers or letters assigned to objects serve as labels for identification or classification. An example of this scale includes MBA schools.
- **Ordinal scale**
  A scale that arranges objects or alternatives according to their ranking. An example includes the Business Week ratings of MBA schools.

Scale Type	Numerical Operations	Descriptive Statistics
Nominal	Counting	Frequency in each category Percentage in each category Mode
Ordinal	Rank ordering	Median Range Percentile ranking
Interval	Arithmetic operations on intervals between numbers	Mean Standard deviation Variance
Ratio	Arithmetic operations on actual quantities	Geometric mean Coefficient of variation

Fig. 1.1 Descriptive statistics for types of scales

- **Interval scale**
  A scale that not only arranges objects according to their magnitudes, but also distinguishes this ordered arrangement in units or equal intervals. The interval between measurements is a meaningful value. It does not involve a true zero point. Standardized test scores in MBA entrance exams is an example of this scale.
- **Ratio scale**
  A scale having absolute rather than relative quantities and possessing an absolute zero, where there is an absence of a given attribute. The number of years an MBA program has been offered is measured on a ratio scale.

These scales and their associated descriptive statistics are summarized in Fig. 1.1. All statistics that are appropriate for lower-order scales (nominal is the lowest) are appropriate for higher-order scales (ratio is the highest) but not vice versa.

Statistics: Values derived from the data collected from a random sample.

Variables: Characteristics of interest that may change, unlike a constant, within the scope of a given problem. For example the variable $\mathbf{s}$ = executive salary, and the associated values include: $\mathbf{s_1} = \$200{,}000$, $\mathbf{s_2} = \$580{,}000$, $\mathbf{s_3} = \$610{,}000$.

Common Pitfalls

- Be careful of longitudinal data. Make sure the conditions for the data collection process have remained constant. Often environmental factors may change and thus bias the data.
- You cannot convert qualitative data to quantitative data by changing the labels to numeric format.
- Don’t show too many decimal places. You cannot make your answer more precise than the data that was used to calculate it.
- Make sure the scales have the correct units of measurement and variable name.

Final Thoughts and Activities

Practice Problems

1. Longitudinal data: Take a look at what has been happening with world record times in the men's mile run since 1900. Extrapolate forward 5 years from today, by simply continuing the same trend of the existing line. Are you surprised?
2. AVERT is an organization that is dedicated to eradicating AIDS in Africa. Explore their website and decide if the statistical data is helpful or is there just too much? Is the data presented in a useful way or is it somewhat misleading? Pay careful attention to the percentage data.

   ➢ Open www.avert.org
   ➢ Click on the tab **HIV and AIDS Topics**
   ➢ Click on **Statistics**
   ➢ Click on **Africa HIV and AIDS Statistics.** You have two different reports to view on South Africa and Sub-Saharan Africa

Discussion Boards

1. Often political goals can get in the way of accurate reporting of data. Discuss this and any other issues that might be obstacles in getting complete and accurate data on the unrest in Afghanistan.
2. This year's top mutual funds will likely slide in rank next year. Do you agree/disagree? Use statistical arguments to support your position.
3. Statistics can change your life... disagree or agree. Provide detailed examples to support your position.
4. Why was *Moneyball* such a popular book and movie?

Group Activity

Productivity Growth, defined as getting more output per hour worked by an employee, has been identified as a key goal in companies within the United States. Often companies provide managers and workers with rewards and special incentives for resulting gains in productivity.

Search the web to find supporting data on one of the positions (pro or con) in addressing the question: Are productivity gains always in the best interest of society?

Parting Thought

It has been proven that the celebration of birthdays is healthy. Statistics show that those people who celebrate the most birthdays are the oldest. . .

Problem Solutions

Problem #1: Longitudinal Data

Answer: (Answers may vary) The world record times have decreased in time since the 1900s. Runners have gotten faster.

The table of data for this problem was retrieved from Wikipedia on March 7, 2012 (http://en.wikipedia.org/wiki/Mile_run_world_record_progression#cite_note-iaaf-4)

Time	Athlete	Nationality	Date	Venue
4:14.4	John Paul Jones	United States	31 May 1913[5]	Allston, Mass.
4:12.6	Norman Taber	United States	16 July 1915[5]	Allston, Mass.
4:10.4	Paavo Nurmi	Finland	23 August 1923[5]	Stockholm
4:09.2	Jules Ladoumègue	France	4 October 1931[5]	Paris
4:07.6	Jack Lovelock	New Zealand	15 July 1933[5]	Princeton, N.J.
4:06.8	Glenn Cunningham	United States	16 June 1934[5]	Princeton, N.J.
4:06.4	Sydney Wooderson	United Kingdom	28 August 1937[5]	Motspur Park
4:06.2	Gunder Hägg	Sweden	1 July 1942[5]	Göteborg
4:06.2	Arne Andersson	Sweden	10 July 1942[5]	Stockholm
4:04.6	Gunder Hägg	Sweden	4 September 1942[5]	Stockholm
4:02.6	Arne Andersson	Sweden	1 July 1943[5]	Göteborg
4:01.6	Arne Andersson	Sweden	18 July 1944[5]	Malmö
4:01.4	Gunder Hägg	Sweden	17 July 1945[5]	Malmö
3:59.4	Roger Bannister	United Kingdom	6 May 1954[5]	Oxford
3:58.0	John Landy	Australia	21 June 1954[5]	Turku
3:57.2	Derek Ibbotson	United Kingdom	19 July 1957[5]	London
3:54.5	Herb Elliott	Australia	6 August 1958[5]	Santry, Dublin
3:54.4	Peter Snell	New Zealand	27 January 1962[5]	Wanganui
3:54.1	Peter Snell	New Zealand	17 November 1964[5]	Auckland
3:53.6	Michel Jazy	France	9 June 1965[5]	Rennes
3:51.3	Jim Ryun	United States	17 July 1966[5]	Berkeley, Cal.
3:51.1	Jim Ryun	United States	23 June 1967[5]	Bakersfield, Cal.
3:51.0	Filbert Bayi	Tanzania	17 May 1975[5]	Kingston
3:49.4	John Walker	New Zealand	12 August 1975[5]	Göteborg
3:49.0	Sebastian Coe	United Kingdom	17 July 1979[5]	Oslo
3:48.8	Steve Ovett	United Kingdom	1 July 1980[5]	Oslo
3:48.53	Sebastian Coe	United Kingdom	19 August 1981[5]	Zürich
3:48.40	Steve Ovett	United Kingdom	26 August 1981[5]	Koblenz
3:47.33	Sebastian Coe	United Kingdom	28 August 1981[5]	Bruxelles
3:46.32	Steve Cram	United Kingdom	27 July 1985[5]	Oslo
3:44.39	Noureddine Morceli	Algeria	5 September 1993[5]	Rieti
3:43.13	Hicham El Guerrouj	Morocco	7 July 1999[5]	Rome

[5] Referenced from “12th IAAF World Championships In Athletics: IAAF Statistics Handbook. Berlin 2009.” (PDF). Monte Carlo: IAAF Media & Public Relations Department. 2009. pp. Pages 546, 549–50. Retrieved August 4, 2009

Excel Output

Let's look at how often the record actually changes and by how much it changes. It seems that the changes as expected are very small. But even so, how fast can humans actually run… there must be some limit. The time between records seems to almost follow a cycle, where several records get set close together and then it takes a while for a new break through time. The last time the record was set was in 1999; based on the pattern, it seems that we are long overdue for a new record and should expect one in the next 5 years, but not by much.

Time	Difference in new record	Date	#years between records
04:14.4		31 May 1913	
04:12.6	00:01.8	16 July 1915	2
04:10.4	00:02.2	23 August 1923	8
04:09.2	00:01.2	4 October 1931	8
04:07.6	00:01.6	15 July 1933	2
04:06.8	00:00.8	16 June 1933	2
04:06.4	00:00.4	28 August 1937	3
04:06.2	00:00.2	1 July 1942	5
04:06.2	00:00.0	10 July 1942	0
04:04.6	00:01.6	4 September 1942	0
04:02.6	00:02.0	1 July 1943	1
04:01.6	00:01.0	18 July 1944	1
04:01.4	00:00.2	17 July 1945	1
03:59.4	00:02.0	6 May 1954	9
03:58.0	00:01.4	21 June 1954	10
03:57.2	00:00.8	19 July 1957	3
03:54.5	00:02.7	6 August 1958	1
03:54.4	00:00.1	27 January 1962	4
03:54.1	00:00.3	17 November 1964	2
03:53.6	00:00.5	9 June 1965	1
03:51.3	00:02.3	17 July 1966	1
03:51.1	00:00.2	23 June 1967	1
03:51.0	00:00.1	17 May 1975	8
03:49.4	00:01.6	12 August 1975	0
03:49.0	00:00.4	17 July 1979	4
03:48.8	00:00.2	1 July 1980	1
03:48.5	00:00.3	19 August 1981	1
03:48.4	00:00.1	26 August 1981	0
03:47.3	00:01.1	28 August 1981	0
03:46.3	00:01.0	27 July 1985	4
03:44.4	00:01.9	5 September 1993	8
03:43.1	00:01.3	7 July 1999	6

Problem #2: AVERT Case Study

Be careful when working with percentages. They can be misleading. It is important to have the actual sample size associated with the percentage. For example 50 % of

100 people is 50 people but 50 % of 2 people is only one person. Also be careful that you are comparing “apples with apples”. This organization tends to compare data from countries with different populations and demographics. Also comparing regional data with single country data can be misleading.

The first set of data that is presented in the case study is related to the impact on the health sector in the countries of sub-Saharan Africa. Consider that 40 % of midwives in Zambia were found to be HIV positive. This information, coupled with the decrease in health care workers, may lead you to believe that there is a lack of education for health care workers in how to care for AIDS patients. Perhaps more data on what the African government and AVERT are doing to promote education would be helpful.

The death toll in South African is also a great concern. The average life expectancy in the worst affected sub-Saharan countries has fallen by 20 years. The biggest increase in deaths among 20–49 year old adults accounts for 60 % of all deaths in sub-Saharan Africa. Perhaps an historical snapshot of the life expectancy of both men and women would be statistically helpful in analyzing the historical trend of life expectancy in these regions. Also, the case study refers to the increase in deaths, but does not mention the average (mean) or median age of death. This would provide a clearer picture on the impact that the epidemic is having on the population. Where does that 60 % fall within that age range?

More quantitative data in the area of economics should be provided. Although the case study states that increased ARV coverage by 50 % would reduce the negative effect on the economy by 17 %, AVERT does not mention the specific positive economic outcome as a result of ARV coverage. The study lacks statistical research in this area. As with life expectancy, some historical trend data should be provided as it relates to the impact of AIDS on the economy.

There is some key data that is missing from this case study, which may prevent potential donors from making a final and informed decision on whether to donate funds towards this charity. http://www.avert.org/aids-young-people.htm.

Chapter 2
Introduction to Excel and Basic Charts

Key Concepts

Area charts, Bar and Column charts, Bubble charts, Filtering, Formatting, Line charts, Pie charts, Pivot Tables, Sorting, Radar charts.

Discussion

In this chapter the key concepts are all about commands in Excel. The first step is to load the **Data Analysis Toolpak.** The standard data structure used throughout this book consists of rows of observations and columns of variables. Some of the basic functionality will be introduced to familiarize you in navigating through Excel spreadsheets.

In this chapter the emphasis is on the basic Excel procedures to allow you to successfully analyze and chart you data. These basic procedures will form the framework for future chapters that have more complicated Excel functionalities. The instructions are provided in a detailed step-by-step process so make sure you do not skip steps. One of the most common frustrations is Excel refusing to compute or complete your graph and usually it is because a vital step has been ignored by the user.

There are many versions of Excel for various types of computers under various operating systems. This book has been compiled using Excel 2010 running under the Microsoft® Windows operating system. However the instructions and flow remain very similar regardless of the Excel version you may be using.

These Excel functions and charts will provide a useful companion to your business applications in the workplace or to your business statistics textbook applications in the classroom.

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_2,

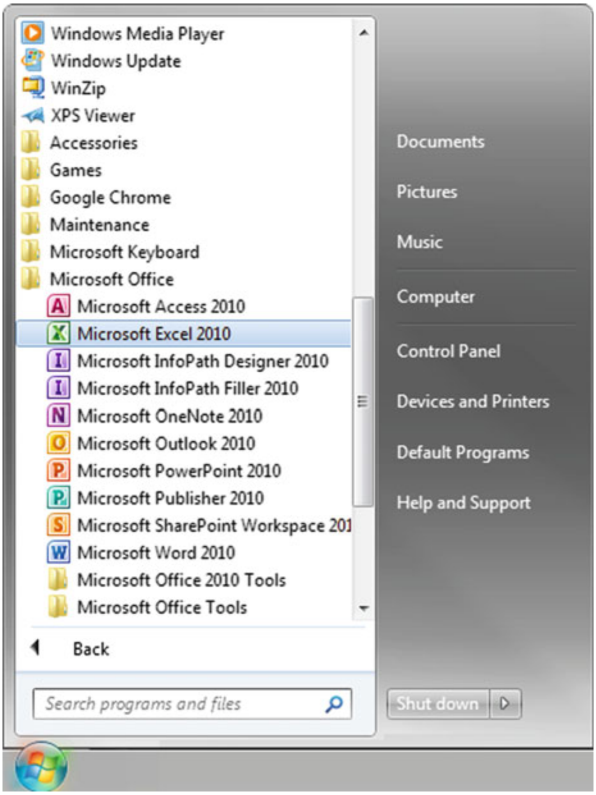
Basic Concepts

Start Up

➢ Click on the **Start button** in the lower left corner of your Windows desktop. The Windows Start menu will open

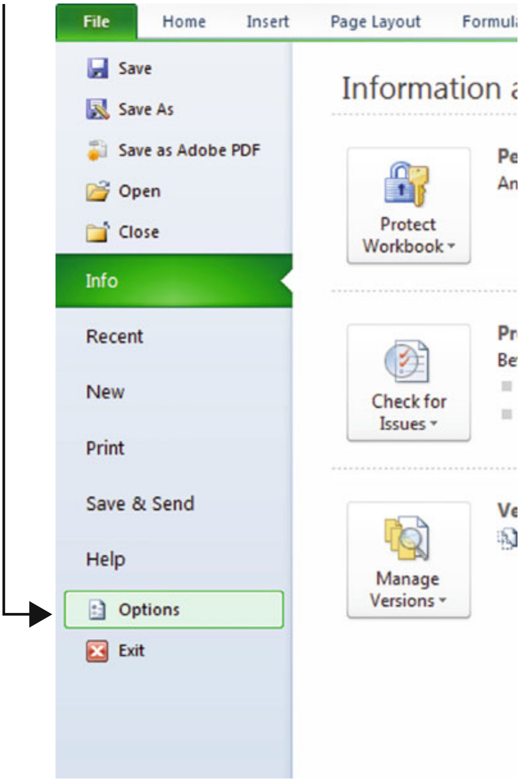
➢ Click on **All Programs** and the Programs menu will appear to the right of the Windows Start menu

➢ Open the Microsoft Office folder and select **Microsoft Excel 2010** (or whatever version you are using)

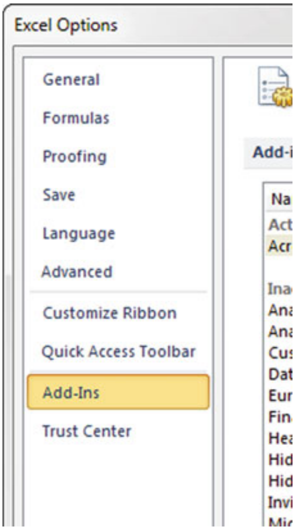


Adding Data Analysis Toolpak

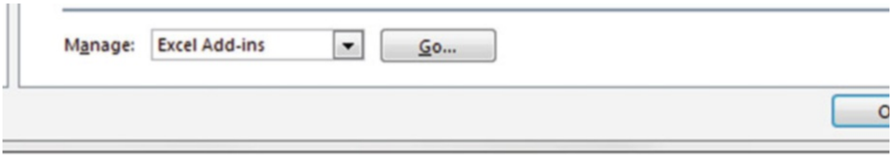
➢ Click the **File** tab on the upper left hand of the Excel window
➢ Select **Options** near the bottom of the command menu



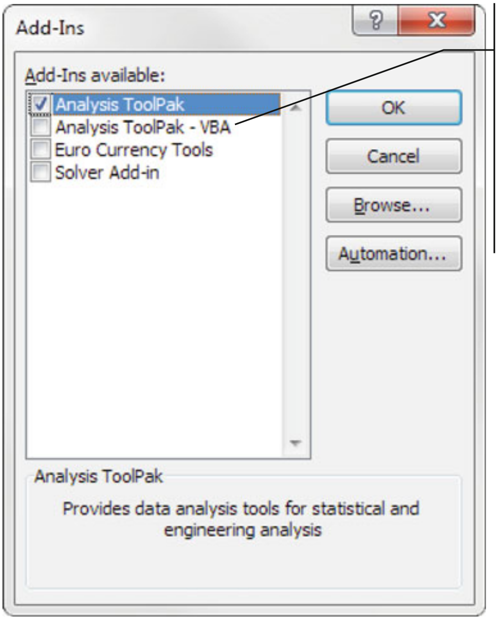
➢ Select **Add-Ins** on the left hand side



➢ Select **Excel Add-ins**, under "Manage:"


Manage: Excel Add-ins

➢ Click **Go**. . .
➢ Check **Analysis ToolPak**



Note: Even though there are two (2) **Analysis ToolPak**s, you only need to select the first one. You will not need to install the **Analysis ToolPak - VBA** because that ToolPak uses a programing language called Visual Basic, which will not be covered in this book.

➢ Click **OK**

Once the Analysis ToolPak is installed, it will launch automatically upon opening Excel

◈ Sometimes the Add-in may seem to vanish. This tool is available to you to re-load as necessary each time you launch Excel.

Excel Elements

Element	Purpose
Title bar	Displays the name of the application and the current Excel document
Menu bar	Contains Excel menus with commands for Excel tasks
Toolbar	Contains buttons that allow one click access to Excel commands and features
Formula bar	Displays the formula or value entered into the currently selected cell
Worksheet	Displays the contents of a chart or Excel spreadsheet results
Cells	Stores individual text or numeric entries
Column headings	Organizes cells into lettered columns
Row headings	Organizes cells into numeric rows
Scroll bar	Used to view cell entries that lie outside the main Excel window
Name box	Displays the names or references of the currently selected object or cell
Task pane	Displays commonly used commands
Sheet tabs	Click to display individual worksheets

Entering Formulas

➢ Click inside the formula bar

➢ Type in an "=" sign and type in your formula or click on the ***fx*** symbol to choose from the drop down menu

  ◇ To make the formula easier to use make sure you input the **cell location** where your variable is located **not the actual value** of the variable.

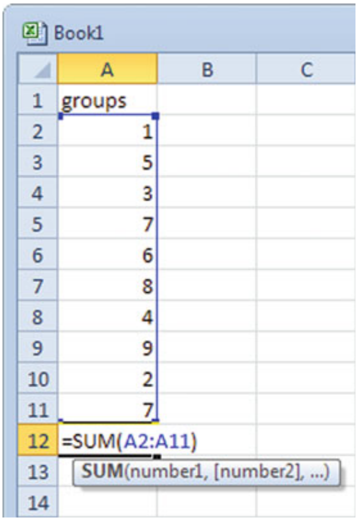
When you input an Excel function the following symbols can be used:

Operator	Description
+	Addition
–	Subtraction
/	Division
*	Multiplication
^	Exponentiation

Cell References

- For a list of data values, use a colon. If you copy and paste a formula in a row or column, the cell reference in the new formulas will shift along with the cell. In this example we are asking Excel to sum up all of the data from cell A2 to cell A11

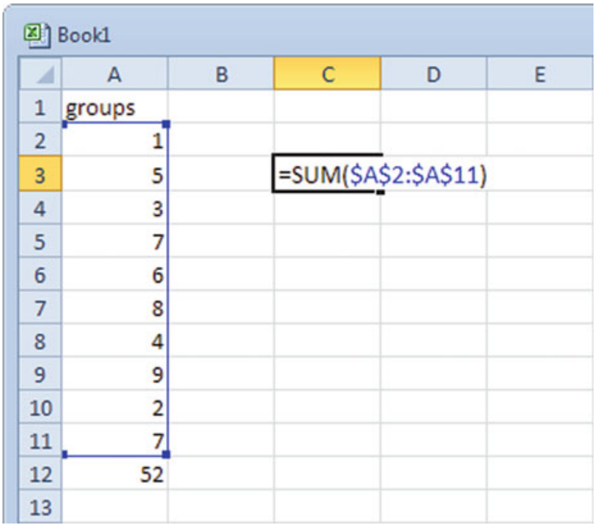
  Example: SUM(A2:A11)



	A	B	C
1	groups		
2	1		
3	5		
4	3		
5	7		
6	6		
7	8		
8	4		
9	9		
10	2		
11	7		
12	=SUM(A2:A11)		
13	SUM(number1, [number2], ...)		
14			

- For a fixed reference, use dollar signs. If you copied this formula into other cells, it would still point to cells A2:A11 and would not be shifted to reflect a new cell location. Note Excel will recognize the following formats as being the same: $A$2, $A2, A$2

  Example SUM($A$2:$A$11)



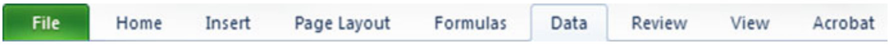
Book1

	A	B	C	D	E
1	groups				
2	1				
3	5		=SUM($A$2:$A$11)		
4	3				
5	7				
6	6				
7	8				
8	4				
9	9				
10	2				
11	7				
12	52				
13					

Sorting Data

To sort a column of data:

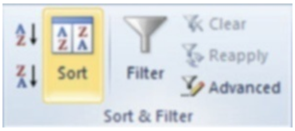
➢ Click on the **Data** tab



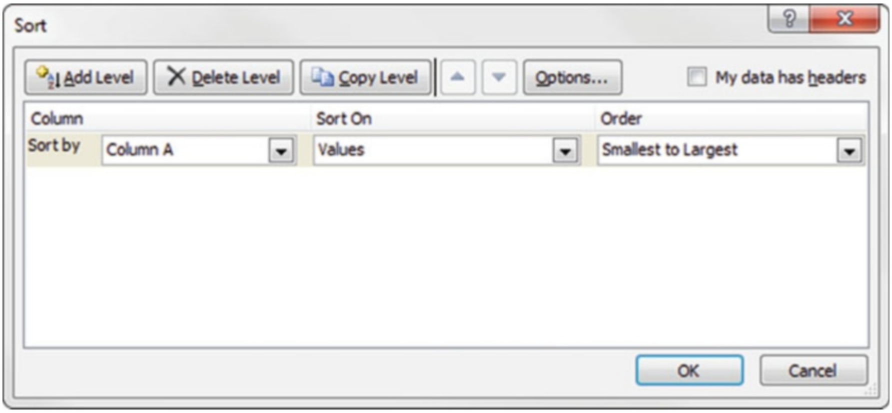
➢ Highlight the previous data set above or create your own data set

◇ If you want to preserve the original data set/order, you need to copy and paste it in another column; otherwise Excel will replace the unordered column with the ordered data.

➢ Choose the **Sort** function



- ➢ Select the type of sorting
- ➢ You can sort by more than one level

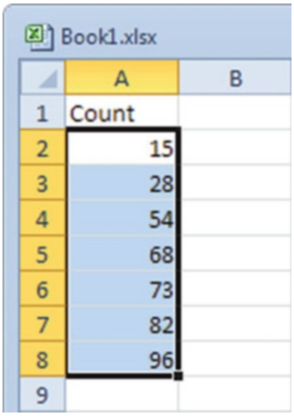


□ My data has headers

Column	Sort On	Order
Sort by Column A	Values	Smallest to Largest

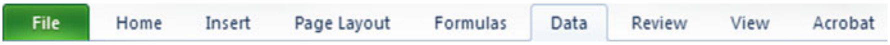
Filtering Data

- ➢ Highlight the data set above or create your own data set that you want to filter
  - ◇ If you want to preserve the original data set, you need to copy and paste it in another column. Excel will replace the unordered column with the ordered data.

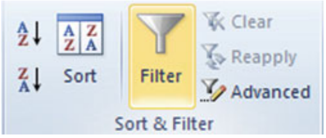

Book1.xlsx

	A	B
1	Count	
2	15	
3	28	
4	54	
5	68	
6	73	
7	82	
8	96	
9		

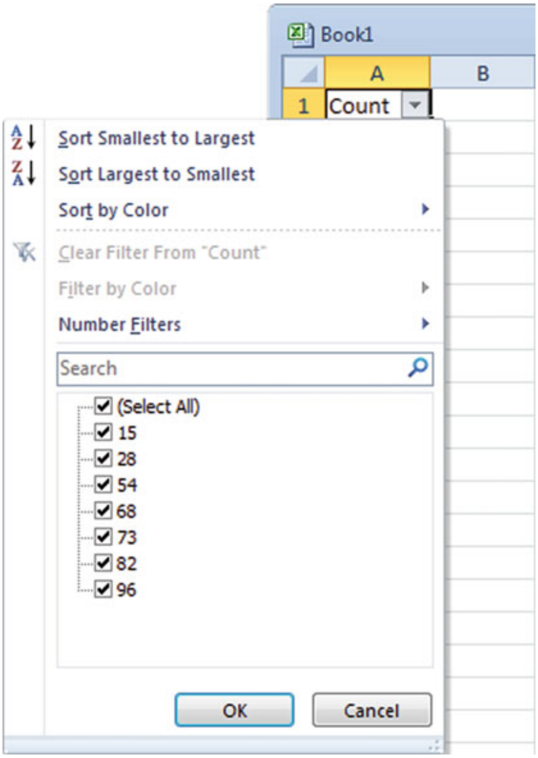
- ➢ Select the first row with the data label
- ➢ Click on the **Data** tab



➢ Click on **Filter** function



➢ Click on the drop down arrow on the first cell



☑ (Select All)
☑ 15
☑ 28
☑ 54
☑ 68
☑ 73
☑ 82
☑ 96

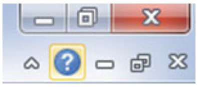
You can apply a sorting function or a filter(s).

Examples

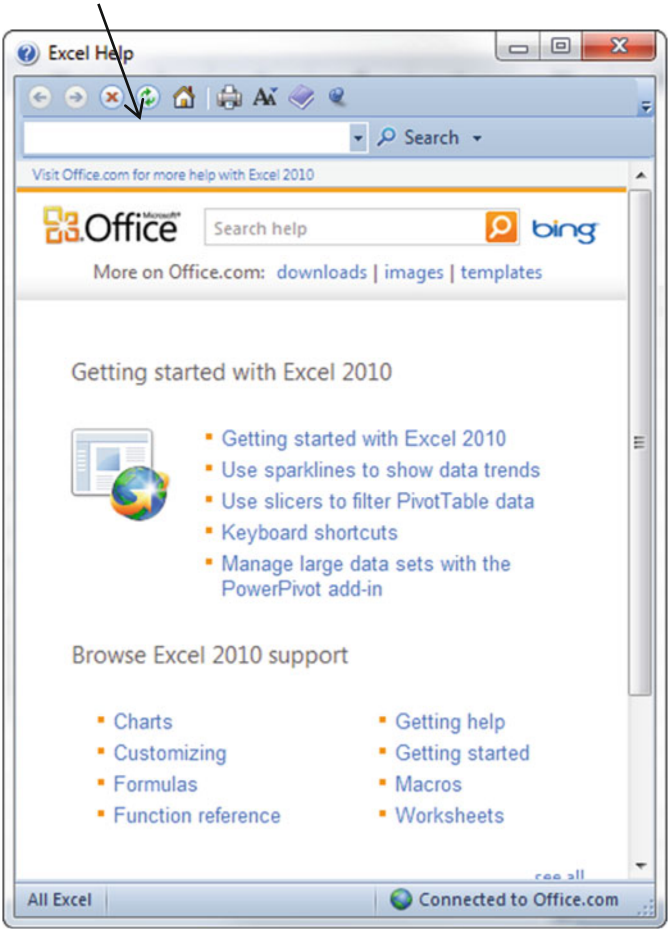
Choose all values greater than 30. Excel will list the values 54, 68, 73, 82, and 96. Choose all values $< =75$ and $> =25$. Excel will list the values 28, 54, 68, and 73.

Getting Excel Help

➢ Click on the question mark on the far right of the upper toolbar



➢ You can type in the concept or the actual question



Excel Help

Search

Visit Office.com for more help with Excel 2010

Office

Search help

bing

More on Office.com: downloads | images | templates

Getting started with Excel 2010

- Getting started with Excel 2010
- Use sparklines to show data trends
- Use slicers to filter PivotTable data
- Keyboard shortcuts
- Manage large data sets with the PowerPivot add-in

Browse Excel 2010 support

- Charts
- Customizing
- Formulas
- Function reference
- Getting help
- Getting started
- Macros
- Worksheets

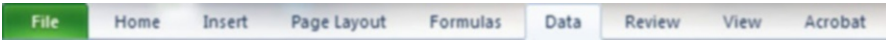
see all

All Excel

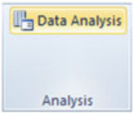
Connected to Office.com

Statistical Tools

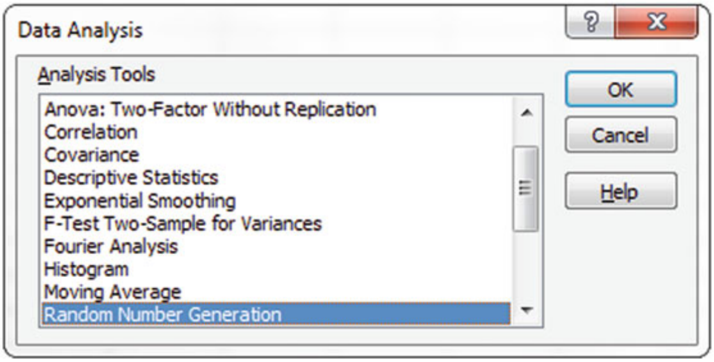
- ➢ Click on the **Data** tab



- ➢ Click in the **Data Analysis** function



- ➢ Highlight the statistical function you need. If the function is not displayed in the list box, click on the vertical scroll bar on the right of the box or scroll to it



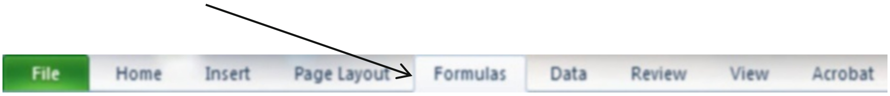
Data Analysis

Analysis Tools

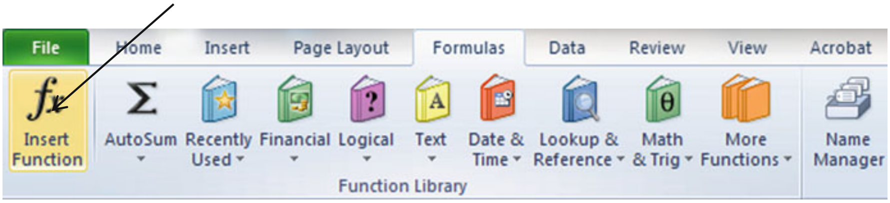
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
Moving Average
Random Number Generation

Predefined or Built-In Formulas

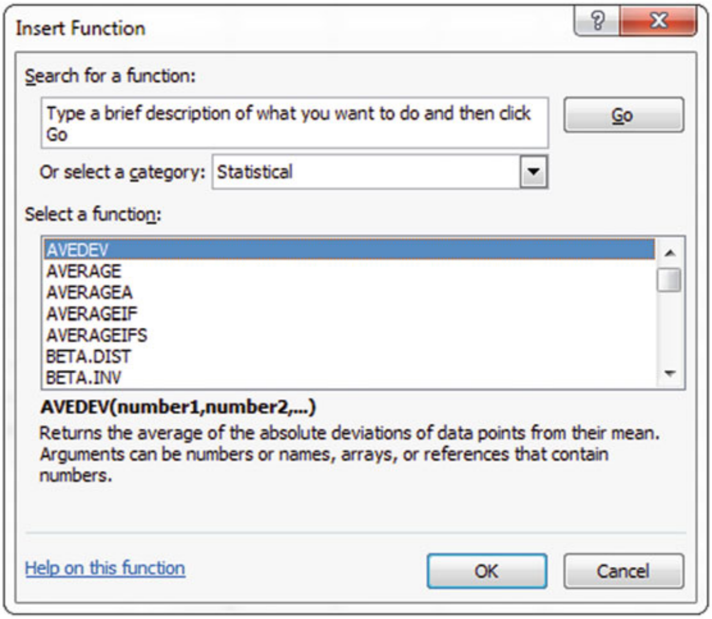
- ➢ Click on the **Formulas** tab



- ➢ Click on the **fx** or **Insert Function** button



- ➢ Click on **Statistical** from the drop down menu to indicate the category
- ➢ Highlight the function you want or type in a description of what you want the function to do



Search for a function:

Type a brief description of what you want to do and then click Go

Or select a category: Statistical

Select a function:

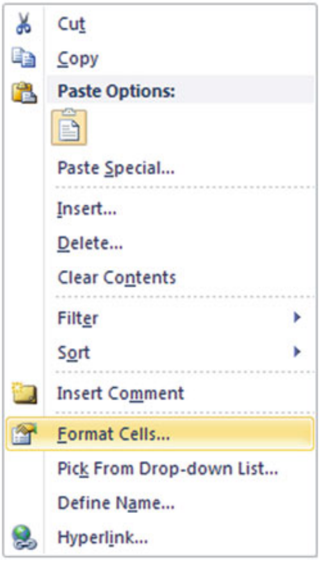
AVEDEV
AVERAGE
AVERAGEA
AVERAGEIF
AVERAGEIFS
BETA.DIST
BETA.INV

AVEDEV(number1,number2,...)

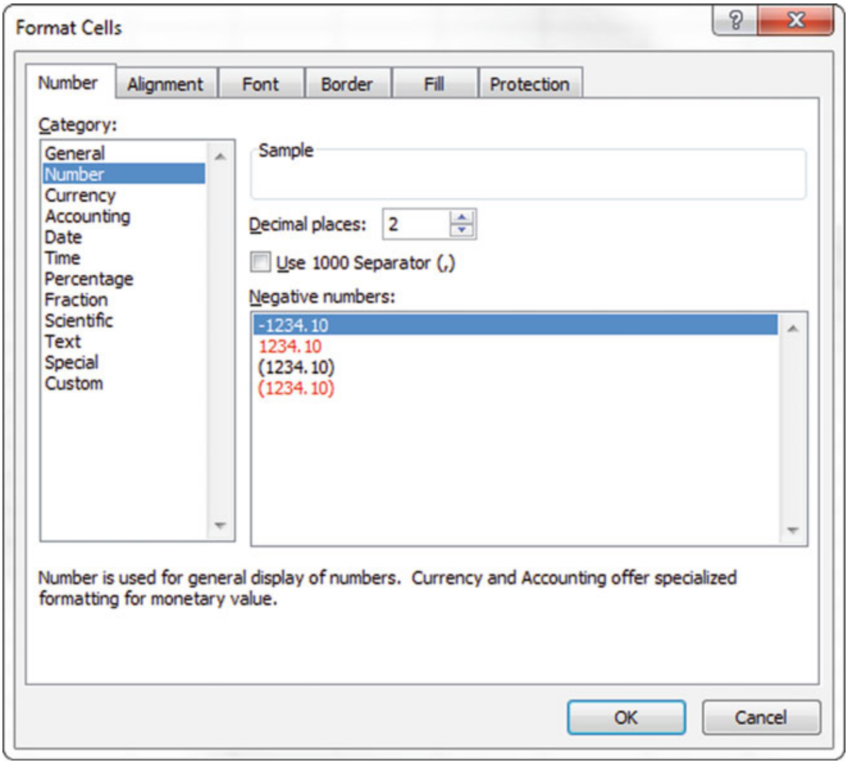
Returns the average of the absolute deviations of data points from their mean. Arguments can be numbers or names, arrays, or references that contain numbers.

Formatting Data

- ➢ Right click on the cells you wish to format
- ➢ Select the **Format Cells** option



- ➢ Select **Number** under Category



Category: Number

Decimal places: 2

□ Use 1000 Separator (,)

Negative numbers: -1234.10

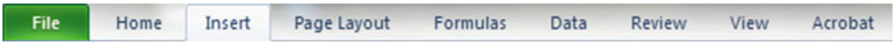
Number is used for general display of numbers. Currency and Accounting offer specialized formatting for monetary value.

- ➢ Choose the correct number of decimal places, usually 2 is sufficient

Chart Wizard

To access the chart wizard:

➢ Click on the **Insert** tab



➢ Select the type of chart you want from the “Charts” section

➢ You can select any of the following chart types listed. See Fig. 2.1

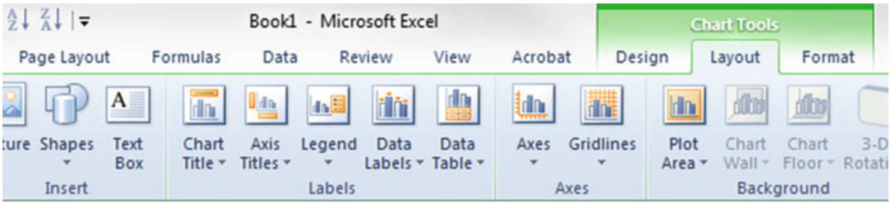
Name	Icon	Description
Column	Column	Shows how data changes over time or between categories. Values are displayed vertically, categories horizontally.
Line	Line	Shows trends in data spaced at equal intervals. Can also be used to compare values between groups.
Pie	Pie	Shows the proportional size of items that make up the whole. The chart is limited to one data series.
Bar	Bar	Shows how data changes over time or between categories. Values are displayed horizontally, categories vertically.
Area	Area	Displays the magnitude of change over time or between categories. You can also display the sum of group values, showing the relationship of each part to the whole.
Scatter	Scatter	Displays the relationship between numeric values in several data series. This chart is commonly referred to as a scatter plot.
Other Charts	Other Charts ▾	Displays a sub-set of charts that are less commonly used.
Stock	Stock	Displays stock market data, including opening, closing, low and high daily values.
Surface	Surface	Shows the value of a data series in relation to the combination of the values of two other data series.
Doughnut	Doughnut	Shows the proportional size of items relative to the whole. It can display more than one data series at a time.
Bubble	Bubble	This is a type of scatterplot but the size of the bubbles is proportional to the value of the third data series.
Radar	Radar	Shows values from different categories radiating from a center point. Lines connect the values within each data series.

Fig. 2.1 Chart types

Formatting of Graphs

To change graph titles or scales:

- ➢ Click on the graph
- ➢ Click on the **Layout** tab (not **Page Layout**, it should be bracketed off in green as "Chart Tools")



- ➢ Click on **Chart Title** function or **Data Labels** function or other aspects of the graph you wish to format

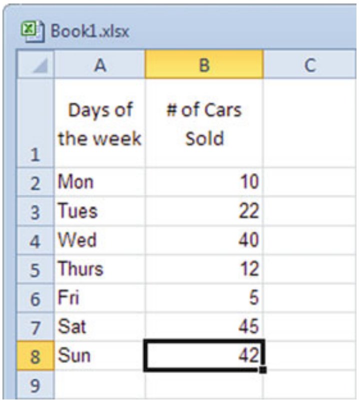
Bar and Column Charts

A bar chart or bar graph is a chart with rectangular bars with lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

Example Problem

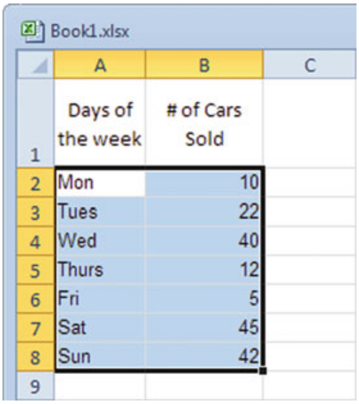
You own a small used car lot and want to do some analysis with car sales and the days of the week that the cars were sold.

- ➢ Enter the text and numbers in columns A and B



	A	B	C
1	Days of the week	# of Cars Sold	
2	Mon	10	
3	Tues	22	
4	Wed	40	
5	Thurs	12	
6	Fri	5	
7	Sat	45	
8	Sun	42	
9			

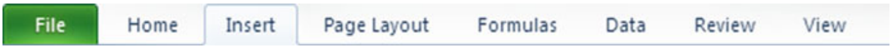
➢ Highlight the 2 columns of data



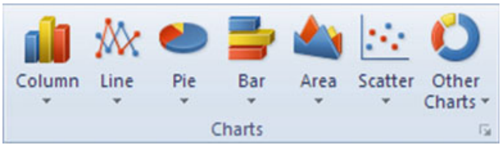
Book1.xlsx

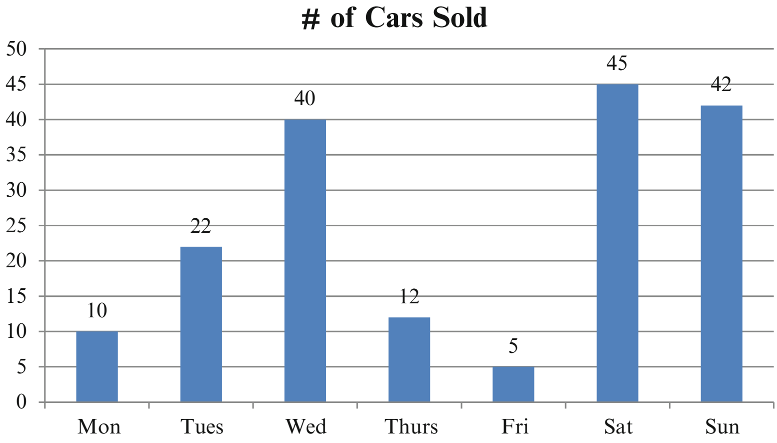
	A	B	C
1	Days of the week	# of Cars Sold	
2	Mon	10	
3	Tues	22	
4	Wed	40	
5	Thurs	12	
6	Fri	5	
7	Sat	45	
8	Sun	42	
9			

➢ Select the **Insert** tab



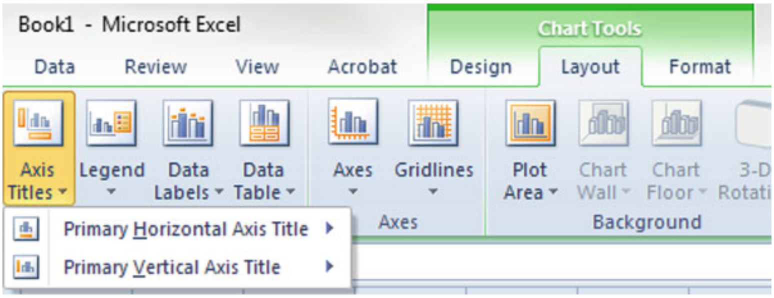
➢ Select the **Column** chart function

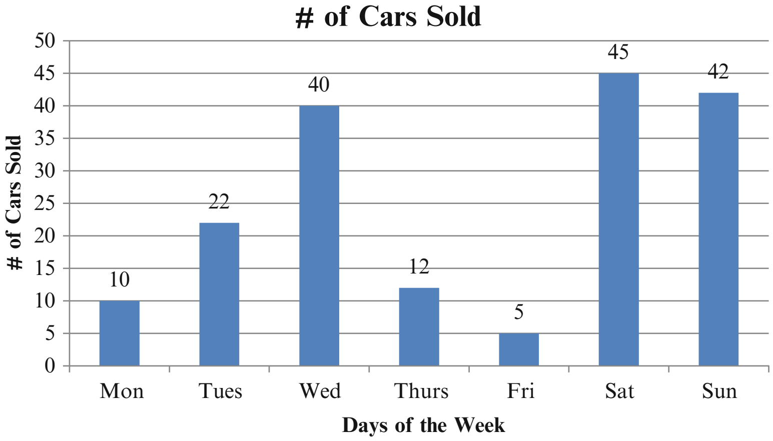




➢ Click anywhere on the chart
➢ Click the **Layout** tab, under Chart Tools
➢ Click the **Data Labels** function to add data labels

➢ Click the **Axis Titles** function to add the appropriate vertical and horizontal labels





Pie Charts

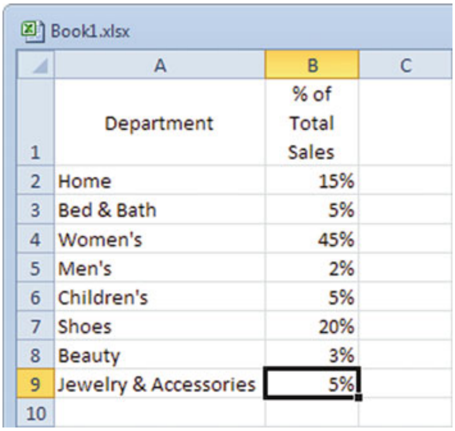
In the pie chart the sum of the percentages on the chart should add to 100 % if you have used all of the input data correctly. In this example the input data is given as percent of total sales. In the next chapter you will learn how to calculate the percentages from the raw data.

Example Problem

As the store manager you wish to complete some sales analysis on each of the departments in your store. The sales data is "% of total sales" by "department."

➢ Enter the text and numbers in columns A and B

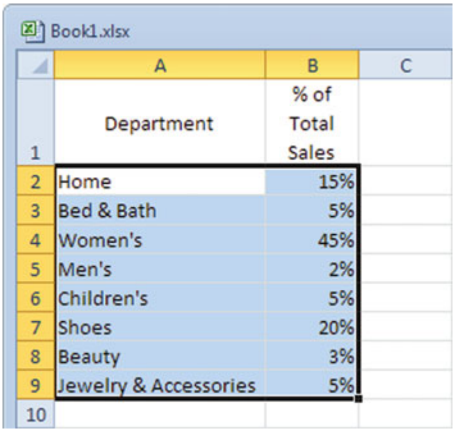
- ◊ Excel will assume the first column is labels so be careful how you enter the data.



Book1.xlsx

	A	B	C
1	Department	% of Total Sales	
2	Home	15%	
3	Bed & Bath	5%	
4	Women's	45%	
5	Men's	2%	
6	Children's	5%	
7	Shoes	20%	
8	Beauty	3%	
9	Jewelry & Accessories	5%	
10			

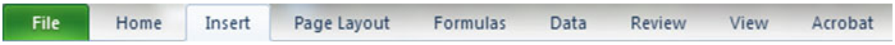
➢ Highlight the 2 columns of data



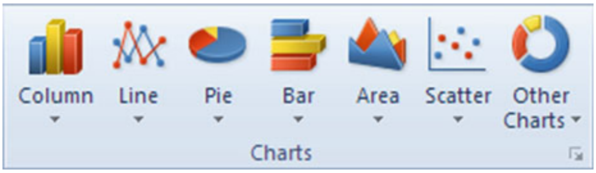
Book1.xlsx

	A	B	C
1	Department	% of Total Sales	
2	Home	15%	
3	Bed & Bath	5%	
4	Women's	45%	
5	Men's	2%	
6	Children's	5%	
7	Shoes	20%	
8	Beauty	3%	
9	Jewelry & Accessories	5%	
10			

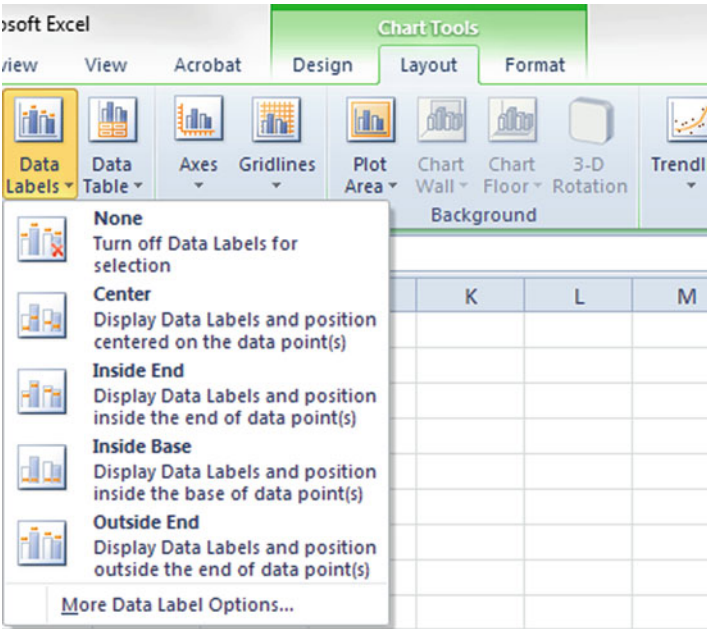
➢ Select the **Insert** tab



➢ Select the **Pie** chart function



- ➢ Choose type of pie chart you want
- ➢ Click anywhere on the chart
- ➢ Click the **Layout** tab, under Chart Tools
- ➢ Click the **Data Labels** function, to add data labels



None
Turn off Data Labels for selection

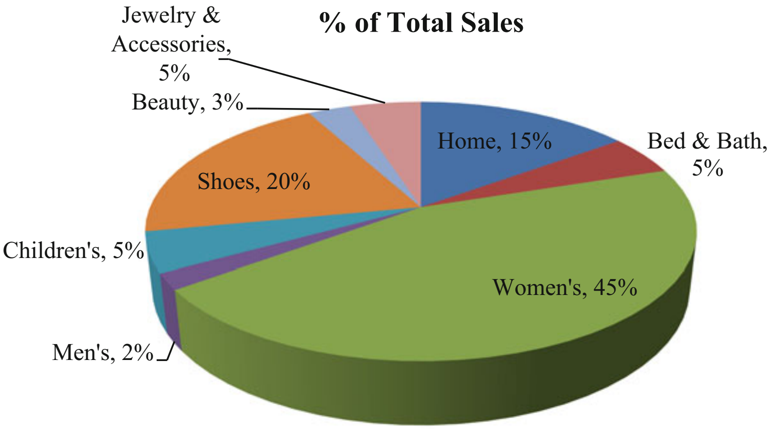
Center
Display Data Labels and position centered on the data point(s)

Inside End
Display Data Labels and position inside the end of data point(s)

Inside Base
Display Data Labels and position inside the base of data point(s)

Outside End
Display Data Labels and position outside the end of data point(s)

More Data Label Options...



- ◇ Remember if you want percentages on the chart you need to format the data cells as percentage. You can select the output as percentages by checking that box.

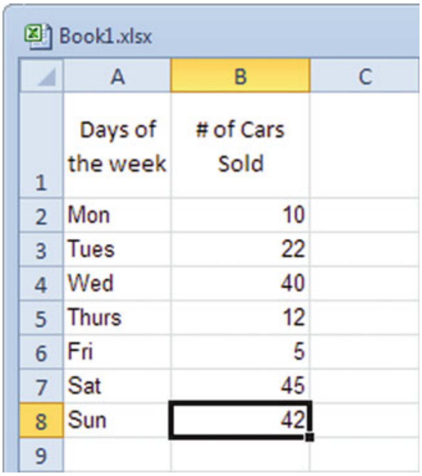
Line Charts and Area Charts

A line chart is a type of chart that displays information as a series of data points connected by straight line segments. An area chart is based on the line chart. The area between the axis and the data line are commonly filled in with colors, textures and hatchings.

Example Problem

You own a small used car lot and want to do some analysis with car sales and the days of the week that the cars were sold.

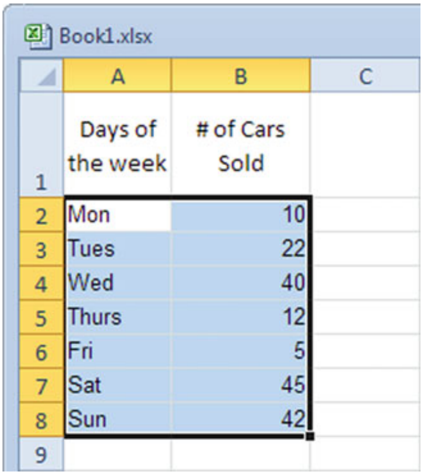
➢ Enter the text and numbers in columns A and B



Book1.xlsx

	A	B	C
1	Days of the week	# of Cars Sold	
2	Mon	10	
3	Tues	22	
4	Wed	40	
5	Thurs	12	
6	Fri	5	
7	Sat	45	
8	Sun	42	
9			

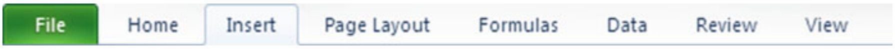
➢ Highlight the 2 columns of data



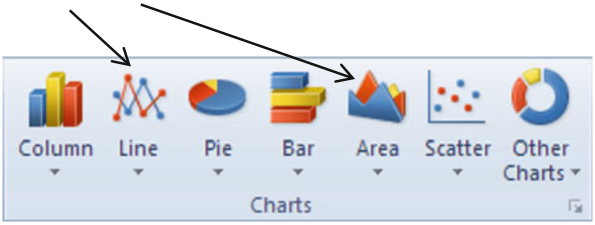
Book1.xlsx

	A	B	C
1	Days of the week	# of Cars Sold	
2	Mon	10	
3	Tues	22	
4	Wed	40	
5	Thurs	12	
6	Fri	5	
7	Sat	45	
8	Sun	42	
9			

- ➢ Select the **Insert** tab

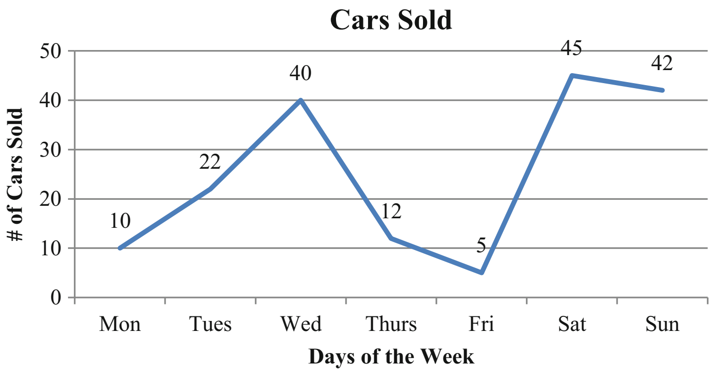


- ➢ Select **Line** or **Area** function



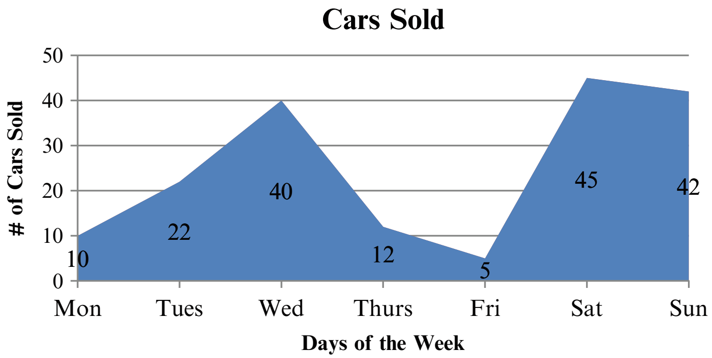
- ➢ Select the type of Line or Area chart desired

Line Graph Example



Area Chart Example

The area chart just fills in the area below the line.



Other Charts

Bubble Chart

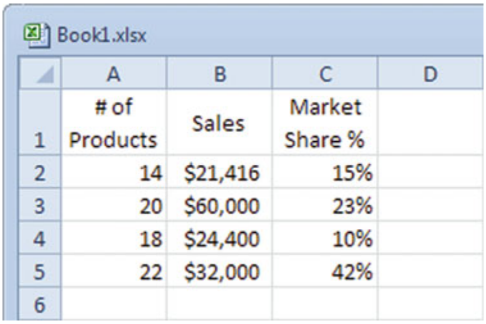
A bubble chart plots 3 dimensional data as a 2 dimensional graph using the size of the bubble to depict the third dimension.

◇ You need to have three arrays of data: x, y, size of bubble.

Example Problem

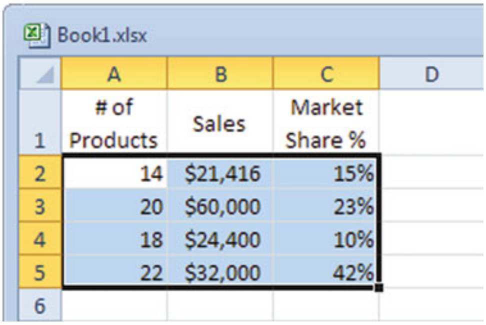
The analysis you wish to complete involves reviewing the product sales and market share by location. The data that is available to you includes the number of different products sold in each of your four geographic areas of operation, the sales generated by those products and the percent of your total market each location represents.

➢ Enter the text and numbers in columns A, B, and C



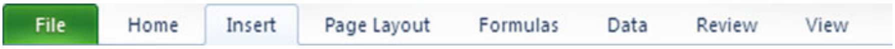
	A	B	C	D
1	# of Products	Sales	Market Share %	
2	14	$21,416	15%	
3	20	$60,000	23%	
4	18	$24,400	10%	
5	22	$32,000	42%	
6				

➢ Highlight the three columns of data

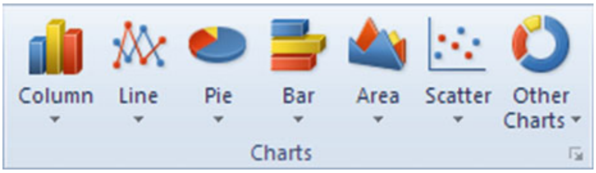


	A	B	C	D
1	# of Products	Sales	Market Share %	
2	14	$21,416	15%	
3	20	$60,000	23%	
4	18	$24,400	10%	
5	22	$32,000	42%	
6				

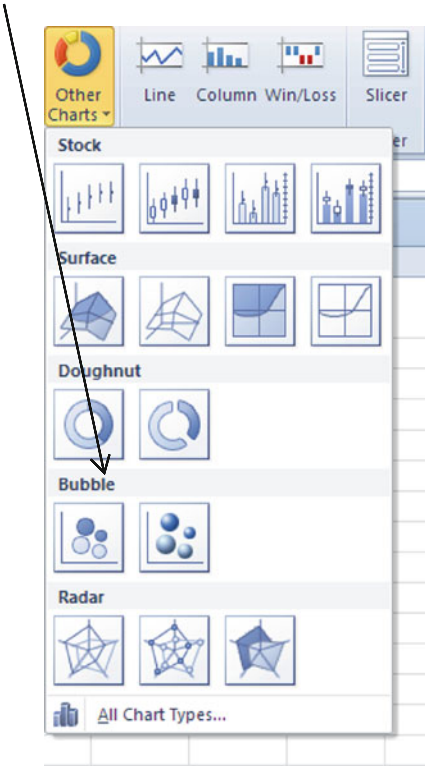
➢ Select the **Insert** tab



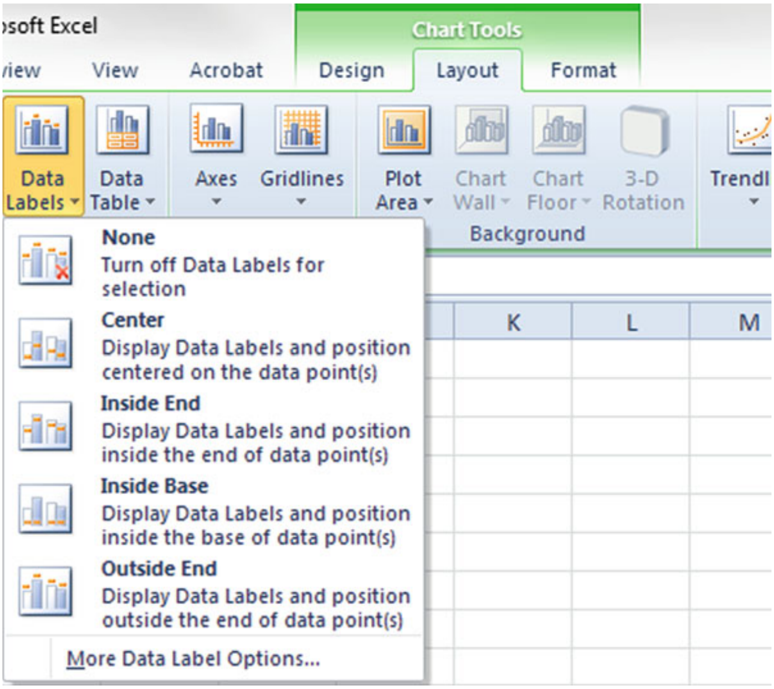
- ➢ Select the **Other Charts** chart function



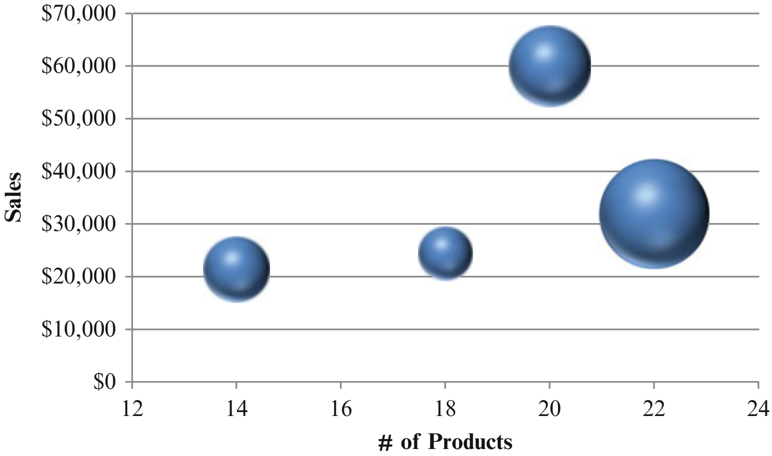
- ➢ Choose a **Bubble** chart



- ➢ Click anywhere on the chart
- ➢ Click the **Layout** tab, under Chart Tools
- ➢ Click the **Data Labels** function to add data labels



›soft Excel
Chart Tools
/iew
View
Acrobat
Design
Layout
Format
Data Labels
Data Table
Axes
Gridlines
Plot Area
Chart Wall
Chart Floor
3-D Rotation
Trendl
Background
None
Turn off Data Labels for selection
Center
Display Data Labels and position centered on the data point(s)
Inside End
Display Data Labels and position inside the end of data point(s)
Inside Base
Display Data Labels and position inside the base of data point(s)
Outside End
Display Data Labels and position outside the end of data point(s)
More Data Label Options...
K
L
M



The size of the bubbles is in proportion to the percent of the market share that each store holds. The bubble chart is included in this book as an example of the different charts that can be created using Excel; however, this specific chart is difficult to use and may not have many applications.

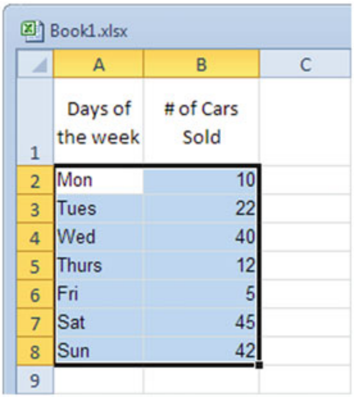
Radar Chart

A radar chart plots multivariate data in the form of a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point. It is also known as a spider chart, web chart or star chart. It consists of a sequence of equi-angular spokes, called radii, with each spoke representing one of the variables. The length of a spoke is proportional to the size of the variable for the data point relative to the maximum size of the variable across all data points. A line is drawn connecting the data values for each spoke. This gives the plot a star-like appearance.

Example Problem

You own a small used car lot and want to do some analysis with car sales and the day of the week that the car was sold.

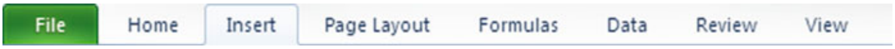
➢ Enter the text and numbers in columns A and B



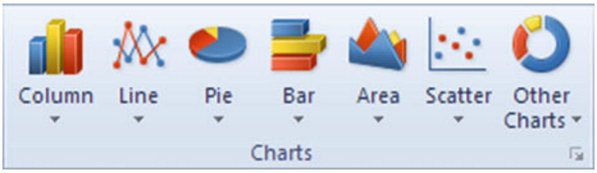
Book1.xlsx

	A	B	C
1	Days of the week	# of Cars Sold	
2	Mon	10	
3	Tues	22	
4	Wed	40	
5	Thurs	12	
6	Fri	5	
7	Sat	45	
8	Sun	42	
9			

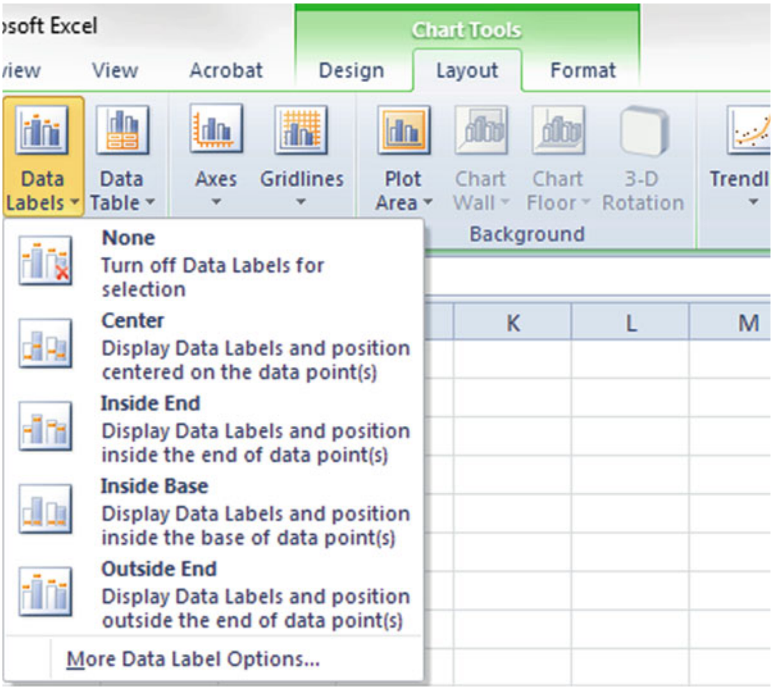
➢ Select the **Insert** tab



➢ Select the **Other Charts** chart function



➢ Choose a **Radar** chart
➢ Click anywhere on the chart
➢ Click the **Layout** tab, under Chart Tools
➢ Click the **Data Labels** function, to add data labels



None
Turn off Data Labels for selection

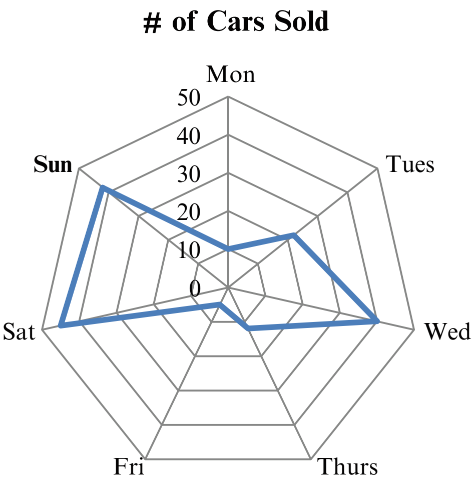
Center
Display Data Labels and position centered on the data point(s)

Inside End
Display Data Labels and position inside the end of data point(s)

Inside Base
Display Data Labels and position inside the base of data point(s)

Outside End
Display Data Labels and position outside the end of data point(s)

More Data Label Options...



The intercepts (the point where the blue line intersects the radial black lines) in the radar chart indicate how many cars were sold on that specific day of the week.

PivotTables (Aka Crosstabs)

PivotTables (as they are known in Excel) or crosstabs allow you to show how two different variables are related to each other in a table format.

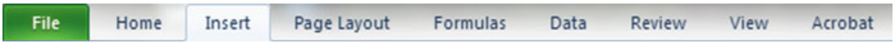
Example Problem

Your survey company has been asked to collect data on the importance of name brands to consumers with special attention to the age of the respondents. The consumers in your sample indicate their importance rating using a scale from 1 to 5, with 5 being the highest possible score and 1 being the lowest possible score. Input the following sample data you collected:

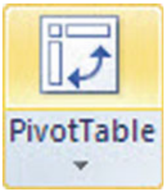
Age	Importance of name brands
18–21	5
18–21	5
18–21	5
18–21	5
18–21	4
18–21	4
18–21	4
18–21	4
18–21	3
18–21	3
18–21	2
18–21	1
22–25	1
22–25	2
22–25	3
22–25	3
22–25	4
22–25	4
22–25	2

To insert a pivot table:

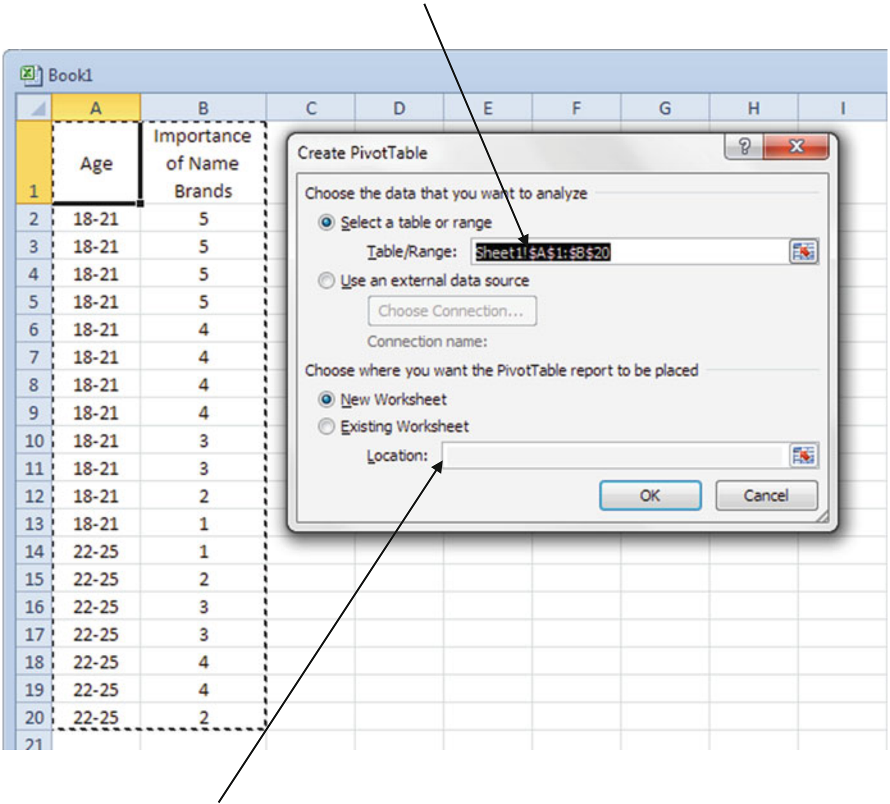
- ➢ Click on the **Insert** tab



- ➢ Select the **PivotTable** function



- ➢ Select the appropriate data for **Table/Range**



	A	B
1	Age	Importance of Name Brands
2	18-21	5
3	18-21	5
4	18-21	5
5	18-21	5
6	18-21	4
7	18-21	4
8	18-21	4
9	18-21	4
10	18-21	3
11	18-21	3
12	18-21	2
13	18-21	1
14	22-25	1
15	22-25	2
16	22-25	3
17	22-25	3
18	22-25	4
19	22-25	4
20	22-25	2
21		

Choose the data that you want to analyze

● Select a table or range

Table/Range: Sheet1!$A$1:$B$20

○ Use an external data source

Connection name:

Choose where you want the PivotTable report to be placed

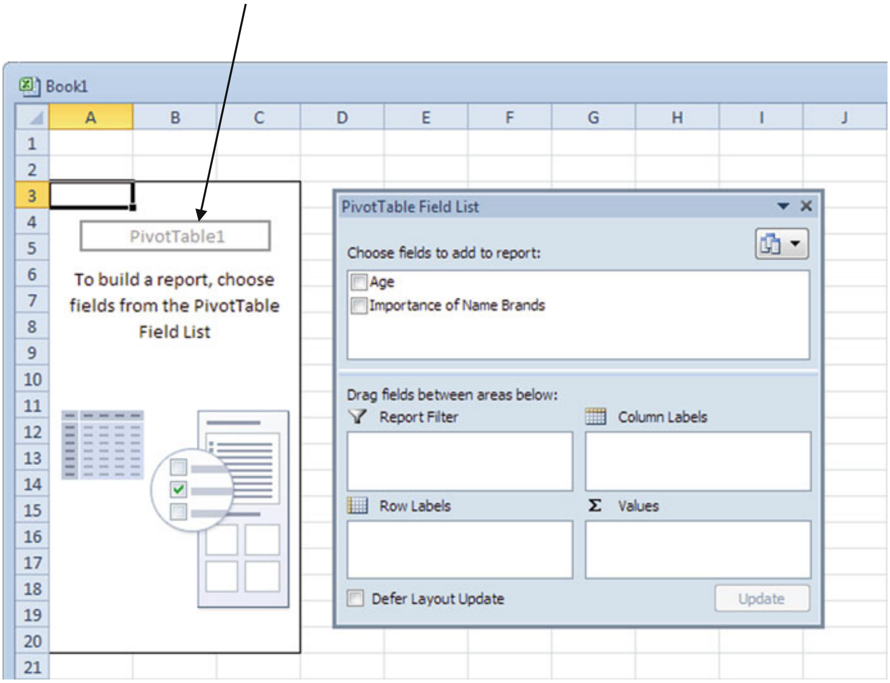
● New Worksheet

○ Existing Worksheet

Location:

- ➢ Choose where you want the PivotTable report to be placed
- ➢ Click **OK**
- ➢ The PivotTable Field List will appear

➢ Click on the **PivotTable1** box



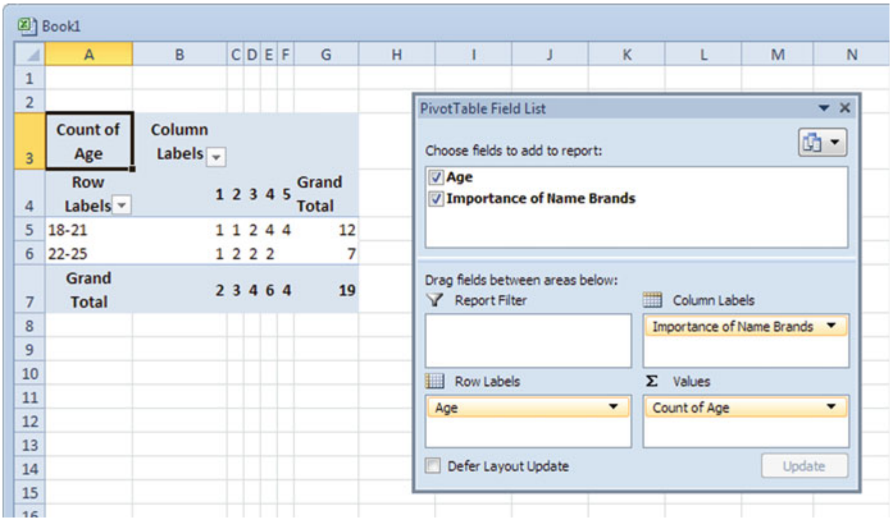
To build a report, choose fields from the PivotTable Field List

Choose fields to add to report:

□ Age
□ Importance of Name Brands

□ Defer Layout Update

➢ Set up your table



Count of Age	Column Labels					
Row Labels	1	2	3	4	5	Grand Total
18-21	1	1	2	4	4	12
22-25	1	2	2	2		7
Grand Total	2	3	4	6	4	19

Choose fields to add to report:

☑ **Age**
☑ **Importance of Name Brands**

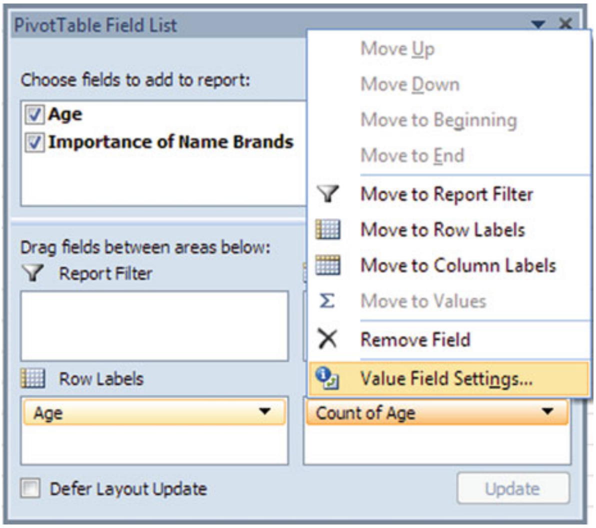
Column Labels: Importance of Name Brands

Row Labels: Age

Σ Values: Count of Age

□ Defer Layout Update

➢ Click the **Values Data** or **Value Field Settings** to edit the table with the appropriate data



Choose fields to add to report:

☑ **Age**

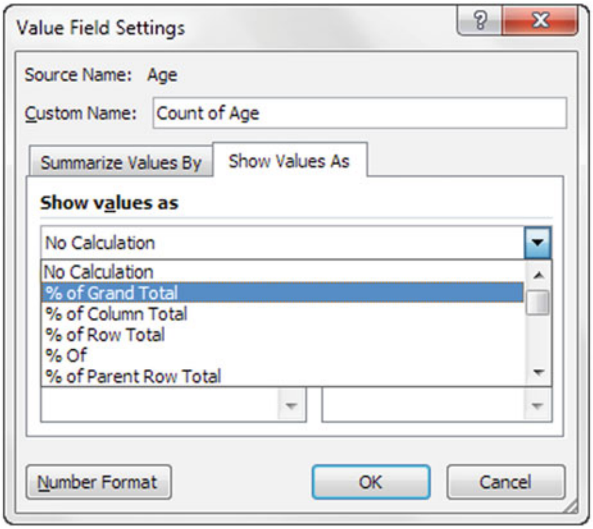
☑ **Importance of Name Brands**

Row Labels: Age

Count of Age

□ Defer Layout Update

➢ Click the **Show values as** tab



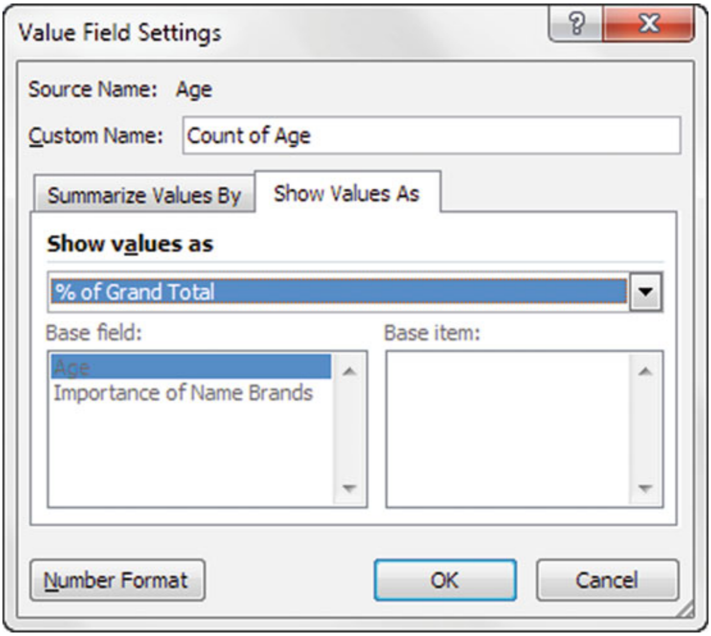
Source Name: Age

Custom Name: Count of Age

Show values as

No Calculation

➢ Select the appropriate calculation



Source Name: Age

Custom Name: Count of Age

Show values as

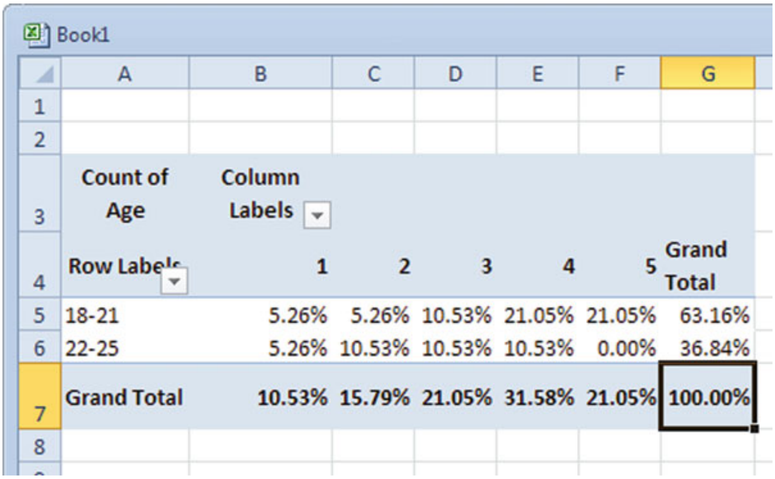
% of Grand Total

Base field:

Age
Importance of Name Brands

Base item:

➢ Click **OK**



Count of Age	Column Labels					
Row Labels	1	2	3	4	5	Grand Total
18-21	5.26%	5.26%	10.53%	21.05%	21.05%	63.16%
22-25	5.26%	10.53%	10.53%	10.53%	0.00%	36.84%
Grand Total	**10.53%**	**15.79%**	**21.05%**	**31.58%**	**21.05%**	**100.00%**

The frequency data for each importance rating value is given within each of the two defined age categories. It seems that the younger age group contains a larger sample, 63 % versus 37 %. The younger age group indicated higher ratings in the 4 and 5 categories of importance. This means younger consumer behaviors may be influenced more by brand names, but we would need much more data to be certain.

Excel

Common Pitfalls

- When doing a pivot table, place the variables in the appropriate boxes for Report Filter, Column Labels, Row Labels, and Values. If you just check the variables, they will end up in Row Labels and Values by default.
- Bubble Charts are difficult to use unless you have 3 interconnected variables. Use caution when deciding to use a bubble chart.
- Some charts will not display the complete variable labels if the labels are long and exceed a certain character limit. If you want to create a chart that will fit a longer variable label, consider using a Column Chart and sizing the graph so that you can expose more text, or changing the text size.
- Don't forget to include the "=" sign when you input a formula.
- Include labels and units for all axes on your charts. By default, Excel will not label your axes.

Final Thoughts and Activities

Practice Problems

1. Prepare a pie chart showing annual sales as a proportion of total sales. There are several options for developing different pie charts from the data.

Region	Annual sales ($)
North	$18,000
South	$52,000
East	$26,000
West	$30,000

2. Prepare a bar chart to show the number of mistakes made in a year by employees with different levels of expertise. There are several options for developing different bar charts from the data.

Expertise level	# of mistakes
Novice	36
Apprentice	20
Tradesmen	9
Master tradesman	3

3. Prepare a line chart to show dollars spent on Valentine's Day flowers over time.

Year	Sale of Valentine's Day flowers
2011	$15,730
2010	$14,160
2009	$14,750
2008	$17,200
2007	$16,890

Discussion Boards

1. Bubble charts and Radar charts are helpful in displaying data. Why are they not more common?
2. Pie charts can be a powerful tool for managers to use on a daily basis but can be easily misused. Explain.

Group Activity

Applied statistical research is an important support activity for the marketing department. There are consulting companies that provide this type of data. What types of statistical analysis do these companies advertise on their websites. What are the challenges in getting managers to accept this data?

Parting Thought

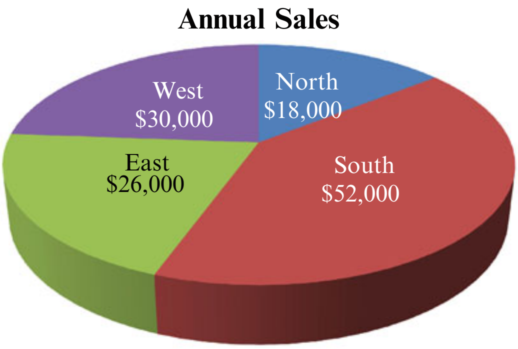
Facts are stubborn things, but statistics are more pliable.

Problem Solutions

1. Prepare a pie chart showing annual sales as a proportion of total sales.

Region	Annual sales ($)
North	$18,000
South	$52,000
East	$26,000
West	$30,000

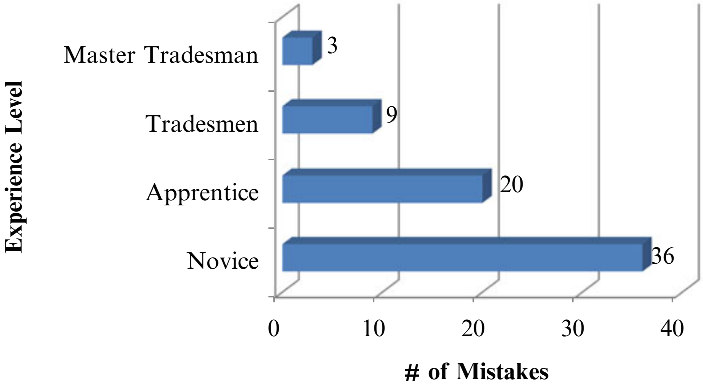
Answer:



2. Prepare a bar chart to show the number of mistakes made in a year by employees with different levels of expertise

Expertise level	# of mistakes
Novice	36
Apprentice	20
Tradesmen	9
Master tradesman	3

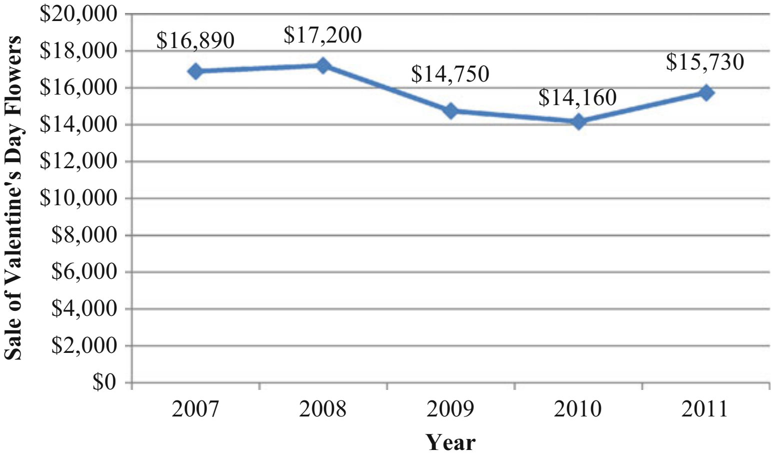
Answer:



3. Prepare a line chart to show dollars spent on Valentine's Day flowers over time.

Year	Sale of Valentine's Day flowers
2011	$15,730
2010	$14,160
2009	$14,750
2008	$17,200
2007	$16,890

Answer:


Sale of Valentine's Day Flowers
$20,000
$18,000
$16,000
$14,000
$12,000
$10,000
$8,000
$6,000
$4,000
$2,000
$0
$16,890
$17,200
$14,750
$14,160
$15,730
2007
2008
2009
2010
2011
Year

Chapter 3
Summarizing Data: Descriptive Statistics and Histograms

Key Concepts

Central tendency, Dispersion, Histogram, Kurtosis, Mean, Median, Mode, Range, Sample variance, Skewness, Standard deviation, and Standard error.

Discussion

This chapter examines the analysis tools used for summarizing single-variable (univariate) data and contains a detailed discussion of **histograms**. The concepts of central tendency and dispersion are explored. Excel output associated with the generation of the histogram provides all of the key descriptive statistics associated with a univariate data set. These are discussed in detail. The access to Excel statistical functions is explained with applied examples.

This chapter concludes with an important section on the pitfalls when working with histograms and the associated basic statistics.

Central Tendency has to do with how the distribution of data clusters around the middle. We typically refer to this position using three important statistics: **mean, median,** or **mode**.

- **Mean**: The mean of a data set refers to the arithmetic average of all the data values. To calculate the mean, sum all of the data and divide the sum by the number of data values.
- **Median**: The median of a data set refers to the value in the middle when the data values are arranged in ascending order. Half of the values are greater than this value and half are less than this value. The median may or may not correspond to a value in your data set; it is just a location.

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_3,

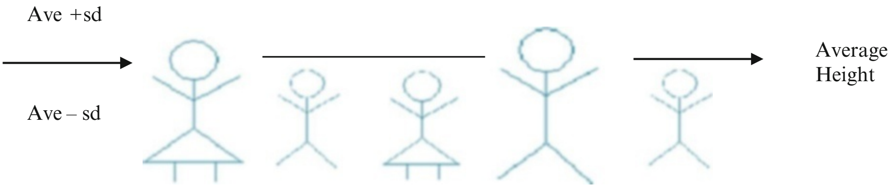
- **Mode**: The mode of a data set refers to the value that occurs with the greatest frequency. If all of the data values occur with the same frequency there will not be a mode.

Dispersion refers to how spread out the data is as a distribution. We use **range** and **standard deviation** as a means of describing dispersion.

By combining the information about the **central tendency** and **dispersion**, you get a good idea of how the data is distributed even without the aid of graphs.

- **Range**: The range of a data set refers to the difference between the largest and smallest data values. The range is the simplest measure of dispersion. The range can be described by stating the smallest and the largest data values (i.e., $25–$100), or by stating the actual number of units between these end points (i.e., $75).
- **Standard Deviation (σ, SD, or sd)**: The standard deviation refers to the most commonly used statistic to measure the variation in any data set. This is a group statistic that summarizes the overall variation of your data. In other words, it indicates how much your data bounces around the average value. Excel provides this statistic in the same units as the original values. Figure 3.1 depicts the concept of standard deviation and provides the formula for a sample.

The concepts of a constant mean and varying spread (dispersion) are shown in Fig. 3.2. Both distributions have the same center value (the average or mean) but the



$$sd = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Fig. 3.1 Standard deviation

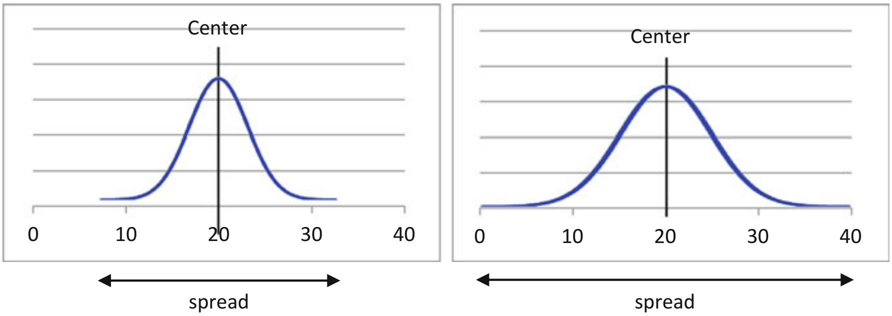


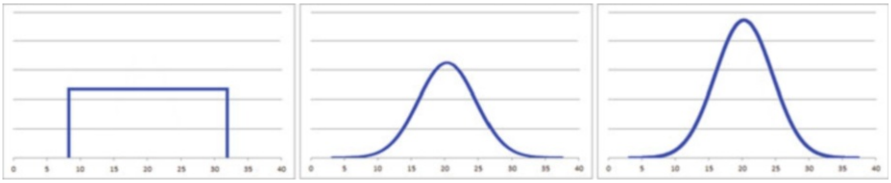
Fig. 3.2 Center and spread

second distribution is more spread out (greater dispersion). The more spread out a distribution is, the larger the standard deviation and range.

Other less used statistics that are provided in the Excel Data Analysis output include:

Kurtosis: Kurtosis refers to how peaked a distribution is or conversely how flat it is. If there are more data values in the tails, than what you expect from a normal distribution, the kurtosis is positive. Conversely if there are less data values in the tails, than you would expect in a normal distribution, the kurtosis is negative.

- Excel cannot calculate this statistic unless you have at least four data values.

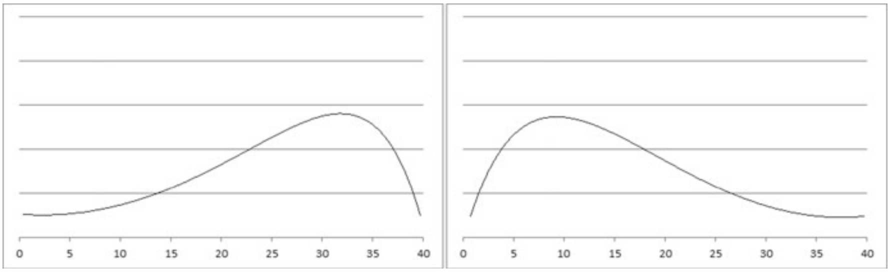


Positive Normal Negative

Fig. 3.3 Kurtosis

Skewness: Skewness describes the lack of symmetry in the distribution. If the data is pulled to the right creating a right hand tail, it is a positive skew. Positively skewed data has a mean greater than the median. Conversely if the data is pulled to the left creating a left hand tail, it is a negative skew and the mean is less than the median.

◇ Excel cannot calculate this if you input less than three data values.



Negative Skew Positive Skew

Fig. 3.4 Skewness

Standard error: This is a measure of the uncertainty about the mean. This becomes an important statistic when working with inference analysis, which is discussed in Chap. 7.

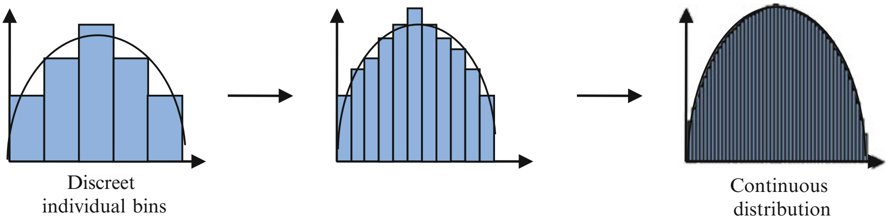


Fig. 3.5 Histogram

Sample variance: This is just the standard deviation squared.

Symbols

Σ = Greek capital letter Sigma which means "take the sum of" or "sum together" everything that follows the sign

$\bar{x}$ or **μ** = average, mean

N or **n** = number of data values in your data set; to be accurate, we use capital **N** for a population (all the possible data) and lower case **n** for a sample (some of the data)

σ or **SD** or **sd** = standard deviation; to be accurate, we use capital **SD** or **σ** for the standard deviation of a population (all possible data), and lower case **sd** for a sample (some of the data)

The Histogram

Histograms are used for charting frequency data. These graphs or charts use individual categories, bins, or classes to count how many times a variable occurs. We can also think of the histogram as more of a continuous or connected graph. See Fig. 3.5.

We see from Fig. 3.5 that if we use an infinite or unlimited number of narrower and narrower bins we would eventually approximate a continuous distribution.

Excel

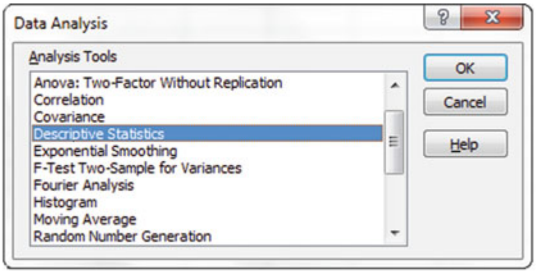
Descriptive Statistics

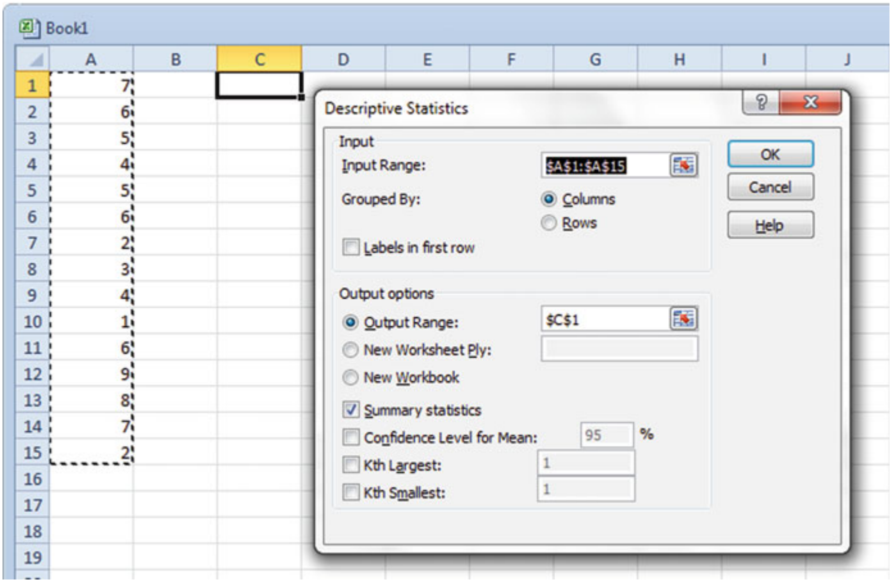
➢ Input the following data into Column A cells A1 through A15: 7, 6, 5, 4, 5, 6, 2, 3, 4, 1, 6, 9, 8, 7, 2

➢ From the **Data** tab, choose the **Data Analysis** function.

➢ Select the **Descriptive Statistics** option, under Analysis Tools

➢ Click **OK**



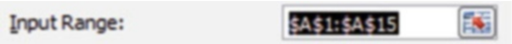


	A
1	7
2	6
3	5
4	4
5	5
6	6
7	2
8	3
9	4
10	1
11	6
12	9
13	8
14	7
15	2
16	
17	
18	
19	

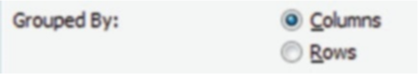
Input
Input Range: $A$1:$A$15
Grouped By: ● Columns ○ Rows
□ Labels in first row

Output options
● Output Range: $C$1
○ New Worksheet Ply:
○ New Workbook
☑ Summary statistics
□ Confidence Level for Mean: 95 %
□ Kth Largest: 1
□ Kth Smallest: 1

➢ Enter the range of data in **Input Range**:


Input Range: $A$1:$A$15

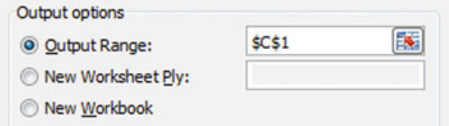
➢ Choose the option that fits your data in **Grouped By:** [columns or rows]


Grouped By: ● Columns ○ Rows

➢ Select **Output Range:**

➢ Click inside the text box.

- Important to do this **first** or you will create problems in your **Input Range**. Then select the output range.



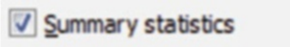
Output options

● Output Range: $C$1

○ New Worksheet Ply:

○ New Workbook

➢ Click **Summary Statistics**



☑ Summary statistics

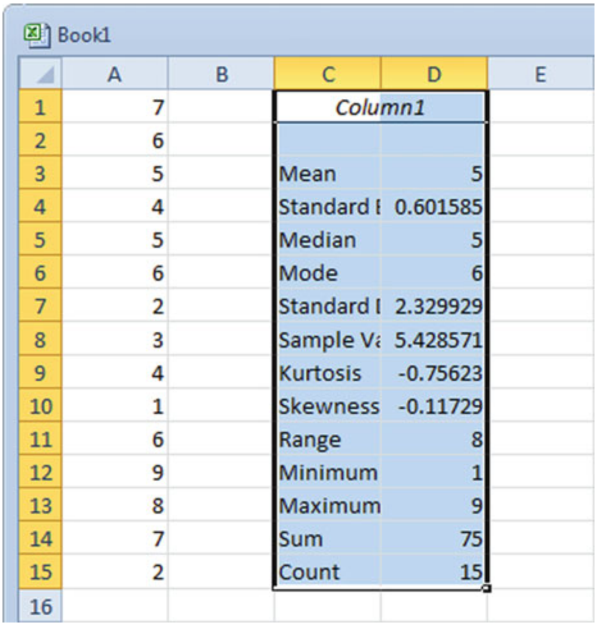
➢ **Confidence Level for the Mean**: Excel default is 95 % which should be acceptable for all of your work. However you can input any value greater than 0 and less than 100. ***You can leave this unchecked.***

➢ **Kth largest***:* If you want the third largest value in the data set you input 3. The default is 1. Values can range from 1 up to the number of data points you have in the distribution. ***You can leave this unchecked.***

➢ **Kth smallest**: If you want the second smallest value in the data set you input 2. The default is 1. Values can range from 1 up to the number of data points you have in the distribution. ***You can leave this unchecked.***

➢ Click **OK**

Descriptive Output Results



Book1

	A	B	C	D	E
1	7		*Column1*		
2	6				
3	5		Mean	5	
4	4		Standard I	0.601585	
5	5		Median	5	
6	6		Mode	6	
7	2		Standard I	2.329929	
8	3		Sample Va	5.428571	
9	4		Kurtosis	-0.75623	
10	1		Skewness	-0.11729	
11	6		Range	8	
12	9		Minimum	1	
13	8		Maximum	9	
14	7		Sum	75	
15	2		Count	15	
16					

The most important output value is the **Count**; this is your check that all of the data has actually been input and used in the calculations.

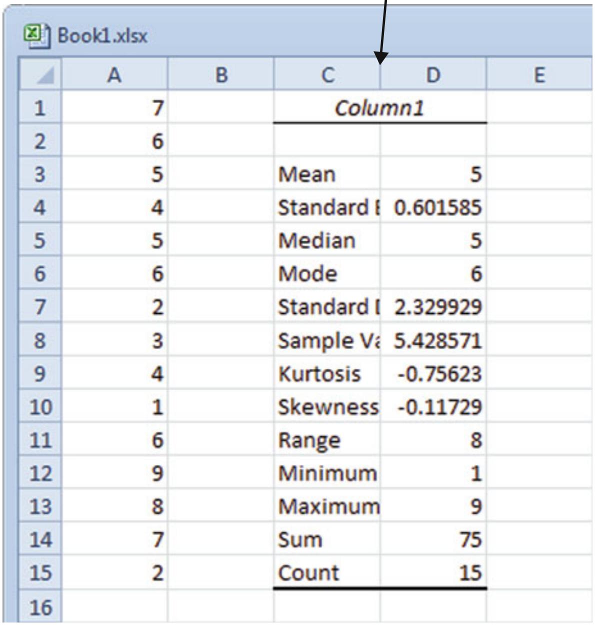
But when it is difficult to read the names of the statistics, we need to adjust the column width.

Changing the Width of the Column Output

Let’s make the column width large enough to show all the words

➢ Place your cursor on the line at the far right of the top cell in the column you want to make wider. In this case we want Column C to be wider.

Before:

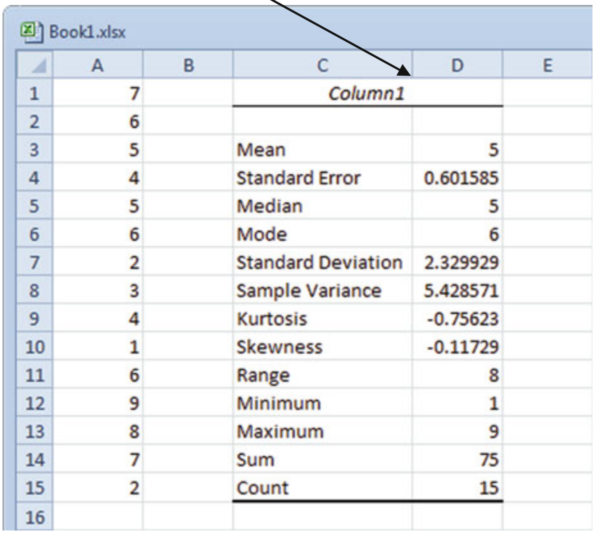


Book1.xlsx

	A	B	C	D	E
1	7		*Column1*		
2	6				
3	5		Mean	5	
4	4		Standard E	0.601585	
5	5		Median	5	
6	6		Mode	6	
7	2		Standard D	2.329929	
8	3		Sample Va	5.428571	
9	4		Kurtosis	-0.75623	
10	1		Skewness	-0.11729	
11	6		Range	8	
12	9		Minimum	1	
13	8		Maximum	9	
14	7		Sum	75	
15	2		Count	15	
16					

➢ Drag this border line to the width you want for the column.

After:

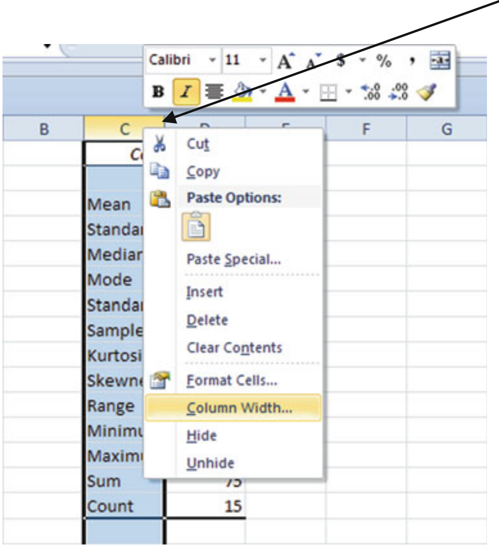


Book1.xlsx

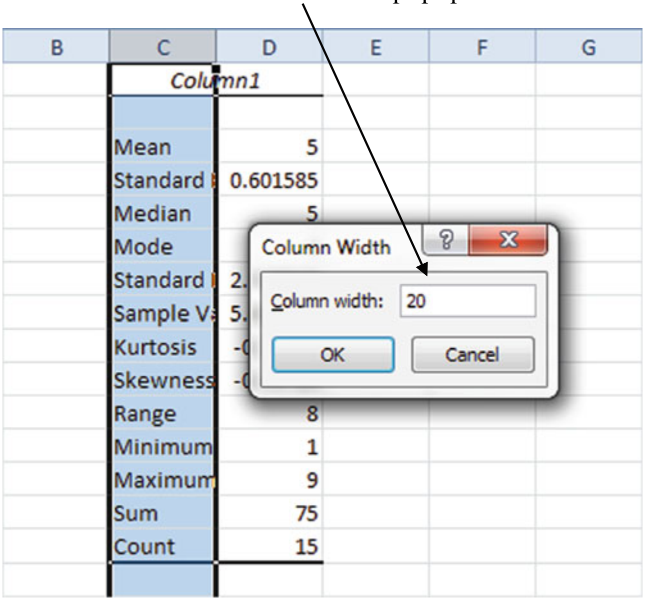
	A	B	C	D	E
1	7		*Column1*		
2	6				
3	5		Mean	5	
4	4		Standard Error	0.601585	
5	5		Median	5	
6	6		Mode	6	
7	2		Standard Deviation	2.329929	
8	3		Sample Variance	5.428571	
9	4		Kurtosis	-0.75623	
10	1		Skewness	-0.11729	
11	6		Range	8	
12	9		Minimum	1	
13	8		Maximum	9	
14	7		Sum	75	
15	2		Count	15	
16					

OR:

➢ Right click the top cell of the column you want to make wider. Select **Column Width.** You can make the column as wide as you need.



➢ Type in the desired column width in the popup box.



Column1	
Mean	5
Standard E	0.601585
Median	5
Mode	
Standard D	2.
Sample Va	5.
Kurtosis	-0
Skewness	-0
Range	8
Minimum	1
Maximum	9
Sum	75
Count	15

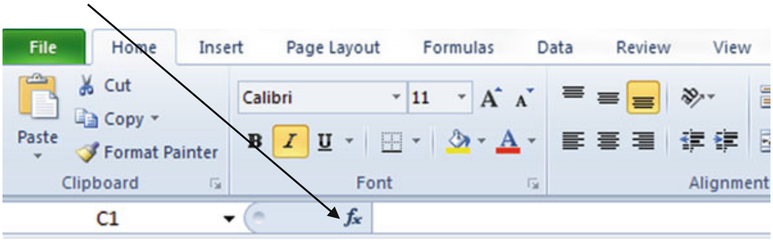
Column width: 20

You can also auto format the column width by double clicking the line on the right side of the column label, which is above the first cell, of the column you want to make wider.

Using Excel Functions

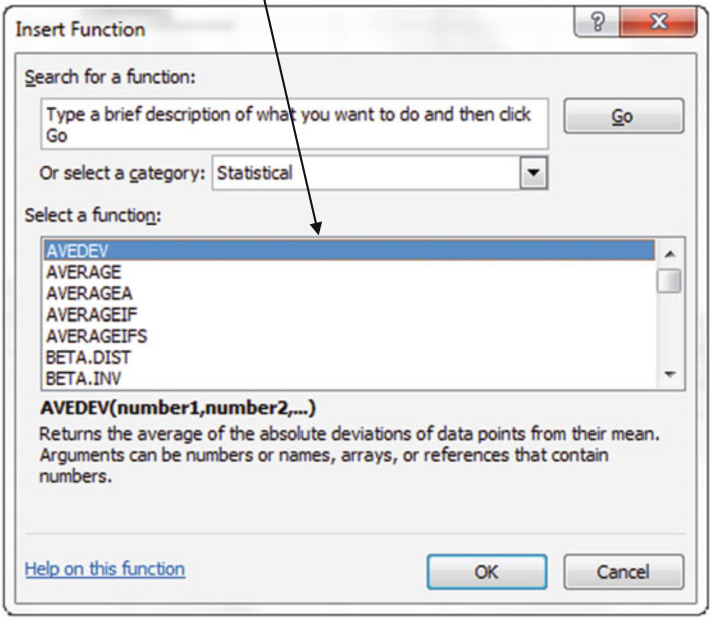
We can also calculate the descriptive statistics one by one using the separate functions in Excel.

➢ Click the ***fx*** symbol in the white function bar



➢ Input the description of what you want to do, if you do not know the name of the statistical function, **OR** click on the next box and select **Statistical**, if you do know the name of the function

➢ Select the function you need



Type a brief description of what you want to do and then click Go

Or select a category: Statistical

AVEDEV(number1,number2,...)

Returns the average of the absolute deviations of data points from their mean. Arguments can be numbers or names, arrays, or references that contain numbers.

➢ Click **OK**

Statistic	Function name	Calculates
Mean	AVERAGE	Arithmetic average of the data set
Median	QUARTILE QUARTILE.INC QUARTILE.EXC (You need to input 2 for the quartile value)	Median of the data set
Mode	MODE MODE.MULT MODE.SNGL	Mode of the data set
Standard Deviation	STDEV STDEV.P STDEV.S STDEVA STDEVPA STDEVP	Standard deviation of the data set (both population standard deviation and sample standard deviation)
Range	MIN MAX	Range of the variables
Count	COUNT	Total count
Skewness	SKEW	Calculates how skewed the data is
Kurtosis	KURT	Calculates the kurtosis of the data
Kth largest	LARGE	Displays the "K" largest number in your data set
Kth smallest	SMALL	Displays the "K" smallest number in your data set

Fig. 3.6 Functions

Be careful to choose the exact function as listed in Fig. 3.6. Excel has many functions that are only different by one letter and each of those functions will deliver a different result.

Histograms

Excel provides a table with the frequency data as well as the actual histogram graph. Remember a histogram does not have spaces between the bars because both the x and y axes are scales and show quantitative data.

Example Problem

You have collected some proficiency scores for the administrative assistants in your department. The proficiency scale ranges from 0 to 1,000, with 0 being the lowest possible score and 1,000 being the highest possible score. You decide that the best way to get a sense of the data is to graph the data in a histogram. Here are the steps you would need to take.

Input the following data in column B

Data
99
150
300
780
350
500
510
360
200
175
196
450

Setting Up the Bin Ranges

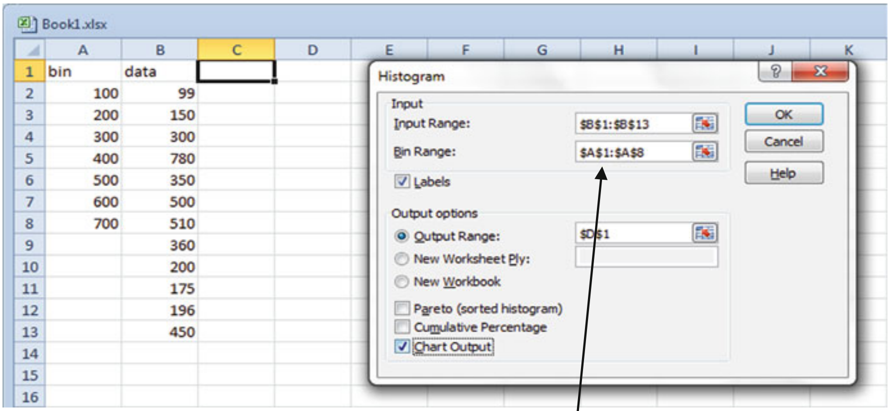
Before you create the histogram you need to identify your bin/category/class ranges. As a guiding principle, charts have somewhere between 5 and 15 bins. Too few bins or too many bins confuse the story.

Arbitrarily choose a certain number of bins. Let's choose seven bins. To get the width of those bins determine the range of the data and divide by the number of bins you want. In this example the range goes from \$99 to \$780 (Fig. 3.7).

$$\begin{gathered}\text{7 bins}\\ \text{Range } \$99\text{–}\$780\\ \text{Bin width} = (\$780-\$99)/7 = \$97.28\end{gathered}$$

Because using bin widths of \$97.28 would be confusing, we can round up or down to choose an easier bin width, such as \$100.

In a new column, type in the largest number in each of the bins to indicate the end point of each bin. In Fig. 3.7 all of the data $\leq$ 100 will be reported in that first bin. All of the data greater than 100 and $\leq$ 200 will be reported in that second bin. Remember these are the end points of the bins you input to Excel.



	A	B	C	D
1	bin	data		
2	100	99		
3	200	150		
4	300	300		
5	400	780		
6	500	350		
7	600	500		
8	700	510		
9		360		
10		200		
11		175		
12		196		
13		450		
14				
15				
16				

Histogram

Input

Input Range: $B$1:$B$13

Bin Range: $A$1:$A$8

☑ Labels

Output options

● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

□ Pareto (sorted histogram)

□ Cumulative Percentage

☑ Chart Output

Fig. 3.7 Input Bin ranges

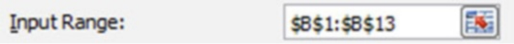
Keep track of where you have entered your bin ranges as you will need this as an input field for the histogram.

Creating the Histogram Chart

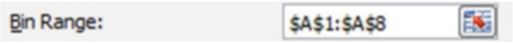
Now you are ready to create a histogram.

- ➢ Under the **Data** tab click on the **Data Analysis** function on the far right
- ➢ Select **Histogram** from the list of options under Analysis Tools and click **OK**

- ➢ Highlight the **Input Range** for the data

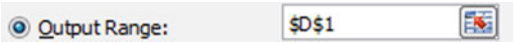

Input Range: $B$1:$B$13

- Highlight the range of cells where you have typed the end points of the bins



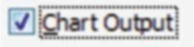
Bin Range: $A$1:$A$8

- Click on the **Output Range:** button and then click inside the box. Now highlight a cell where you want the output to start



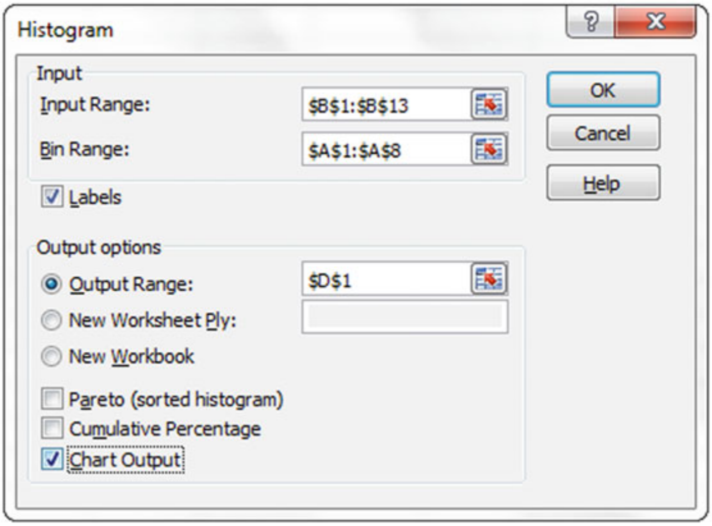
● Output Range: $D$1

- Check **Chart Output** or you won't get a graph



☑ Chart Output

- Click **OK**



Histogram

Input

Input Range: $B$1:$B$13

Bin Range: $A$1:$A$8

☑ Labels

Output options

● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

□ Pareto (sorted histogram)

□ Cumulative Percentage

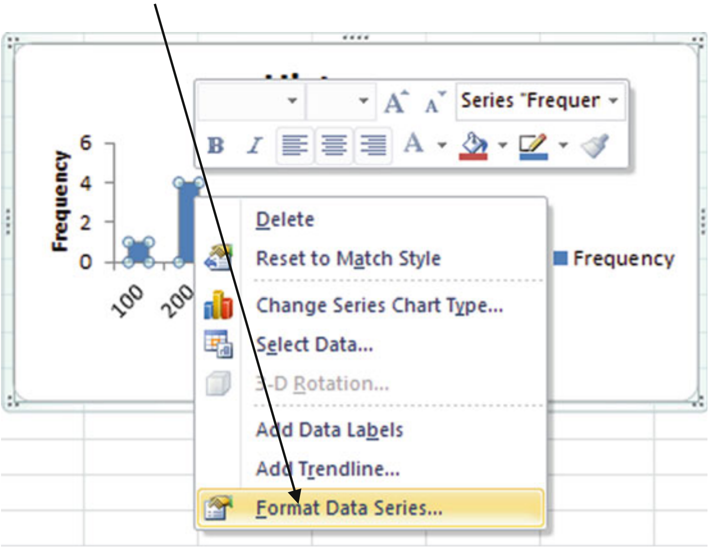
☑ Chart Output

Histogram Clean Up

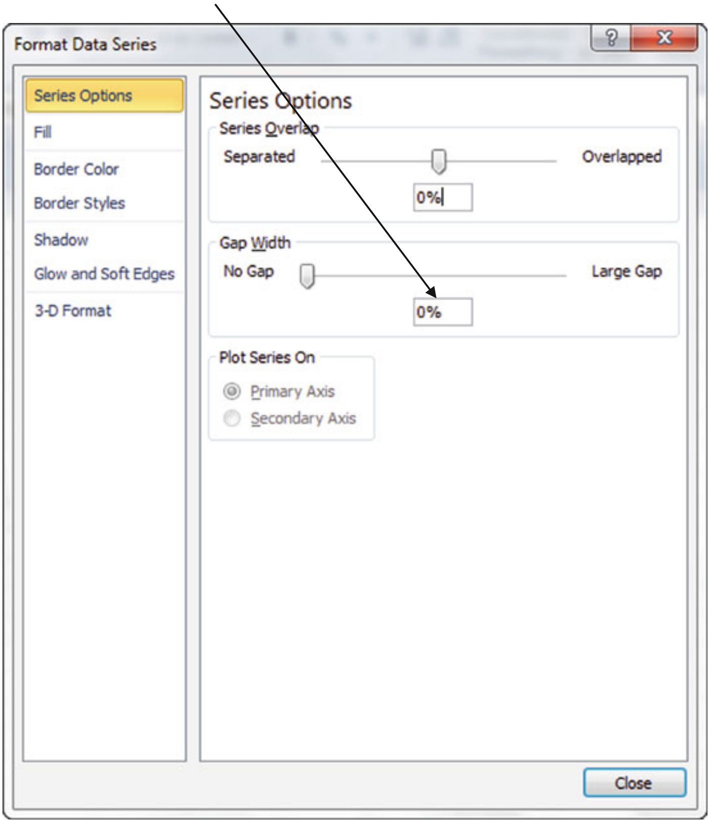
The actual histogram needs some cleaning up.

Closing Gaps Between Bars

- ➢ Right Click in any column on the graph
- ➢ Click the **Format Data Series** option from the menu



- ➢ Click on **Series Options**
- ➢ Go to **Gap Width**. Set to 0 % and click **Close**



Series Overlap: 0%

Gap Width: 0%

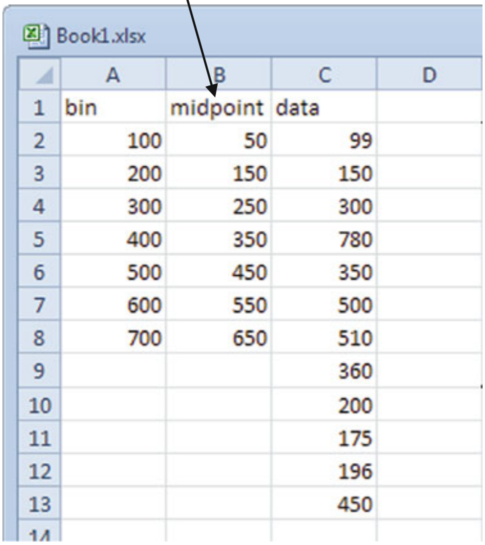
Plot Series On: ● Primary Axis ○ Secondary Axis

Change Labels on x-axis

- ➢ Change the x-axis values from ***upper limits*** of the bins to ***midpoints*** of the bins

➢ Create the column of midpoint labels you want added to your graph

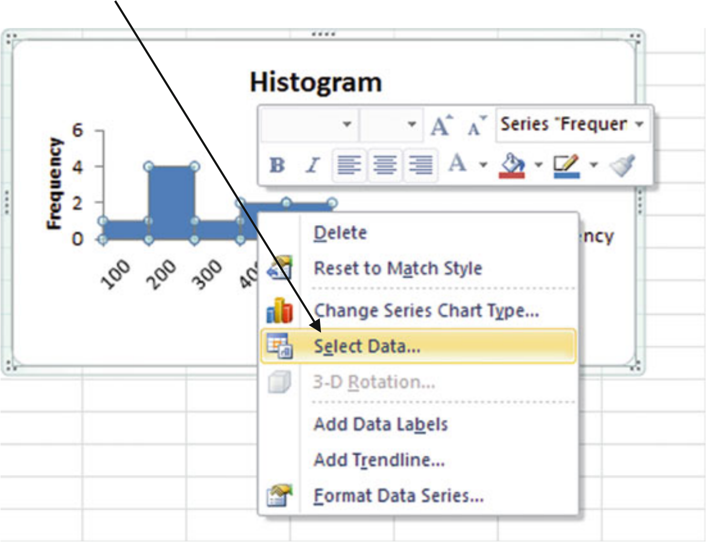
◈ **NOTE**: The data labels on the x-axis indicate the upper limits of the bins. In order, to change this to the midpoints, first you need to create a column of data with the mid-points of the bins.



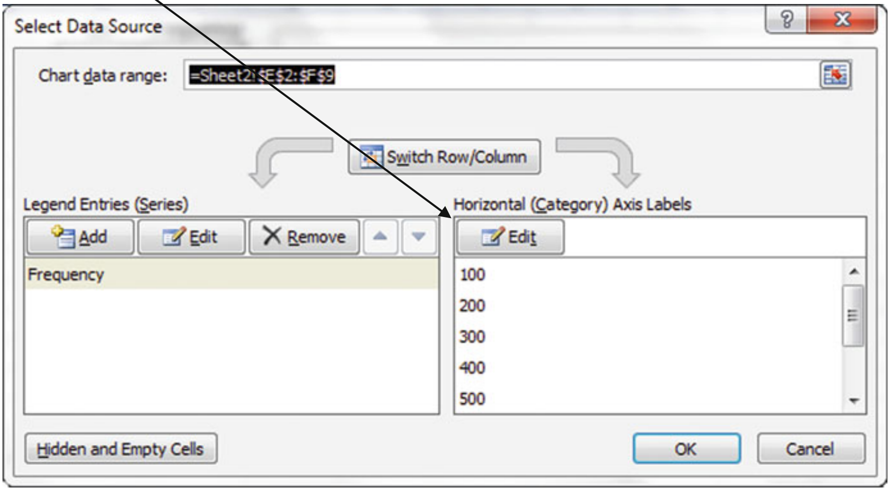
Book1.xlsx

	A	B	C	D
1	bin	midpoint	data	
2	100	50	99	
3	200	150	150	
4	300	250	300	
5	400	350	780	
6	500	450	350	
7	600	550	500	
8	700	650	510	
9			360	
10			200	
11			175	
12			196	
13			450	

➢ Right click on any column on the graph
➢ Choose the **Select Data** option



➢ Under **Horizontal Axis Labels** click the **Edit** option



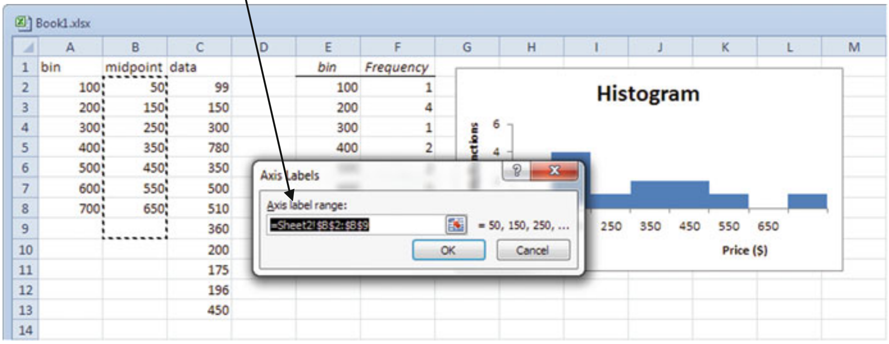
Select Data Source

Chart data range: =Sheet2!$E$2:$F$9

Legend Entries (Series): Frequency

Horizontal (Category) Axis Labels: 100, 200, 300, 400, 500

➢ Click inside the **Axis label range** box. Highlight the new column of midpoint labels



Book1.xlsx

	A	B	C	D	E	F
1	bin	midpoint	data		bin	Frequency
2	100	50	99		100	1
3	200	150	150		200	4
4	300	250	300		300	1
5	400	350	780		400	2
6	500	450	350			
7	600	550	500			
8	700	650	510			
9			360			
10			200			
11			175			
12			196			
13			450			
14						

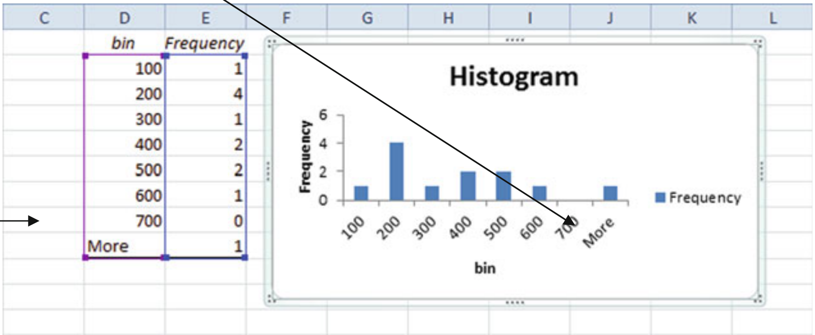
Axis Labels

Axis label range: =Sheet2!$B$2:$B$9 = 50, 150, 250, ...

➢ Click **OK**

Removing More from the Chart and Labeling the Last Column

The word **More** may appear on your graph.



bin	Frequency
100	1
200	4
300	1
400	2
500	2
600	1
700	0
More	1

Fig. 3.8 End-point Bin values

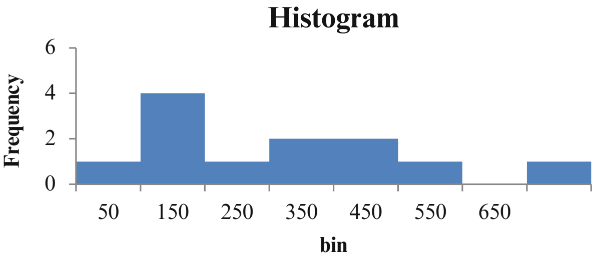
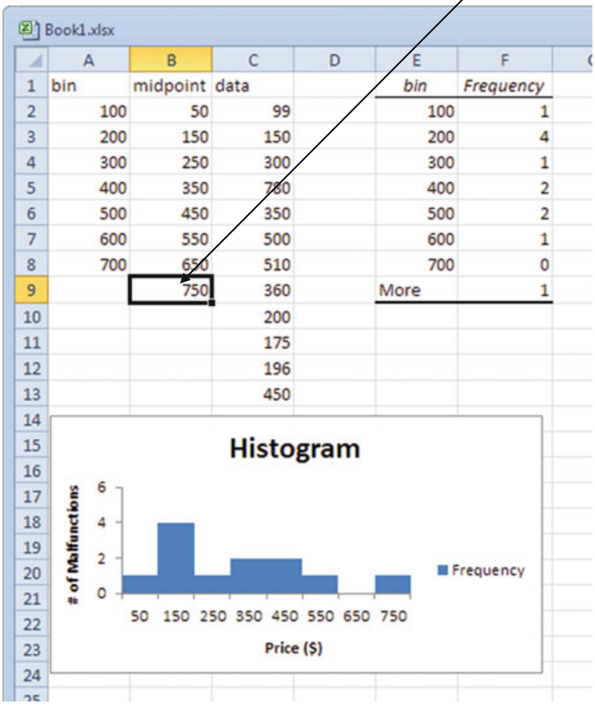


Fig. 3.9 Mid point Bin values

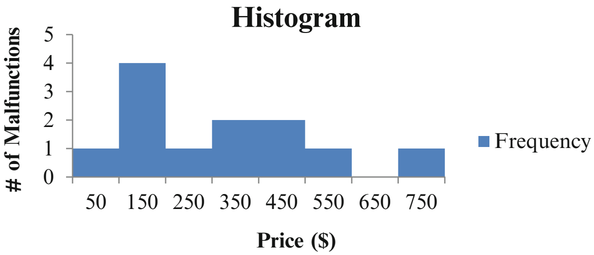
The last column may not be assigned a numeric label. This is because there are data values greater than the last bin value you originally input. In Fig. 3.8 the last bin end value originally selected was 700. In Fig. 3.9 this was changed to a mid-point bin value of 650. But as you can see from the data there is one value greater than 700. Excel automatically names this last bin **More** and will not assign a numeric label for it on the x-axis.

You can go back and add more bins to ensure all of your data is being included on your graph with the proper labels. In this case we added **750**.



	A	B	C	D	E	F
1	bin	midpoint	data		*bin*	*Frequency*
2	100	50	99		100	1
3	200	150	150		200	4
4	300	250	300		300	1
5	400	350	780		400	2
6	500	450	350		500	2
7	600	550	500		600	1
8	700	650	510		700	0
9		750	360		More	1
10			200			
11			175			
12			196			
13			450			

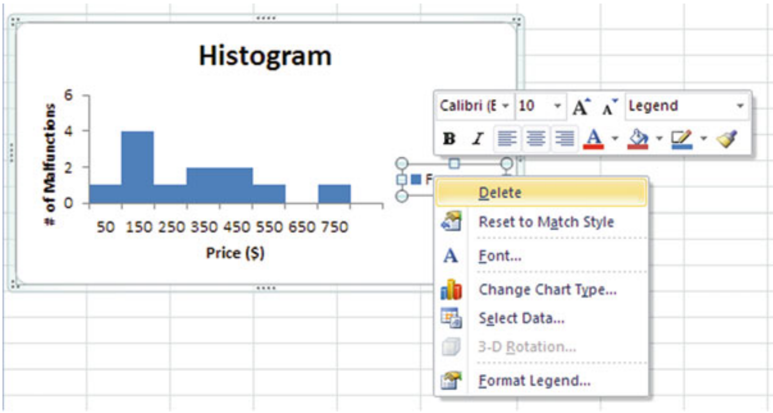
Once you have added the extra bins, edit the bin values which will remove the word “More” from the graph.



◇ Changing the bin values will only change the labels on your x axis. The frequency will still count the data using the initial points you entered, not the midpoints.

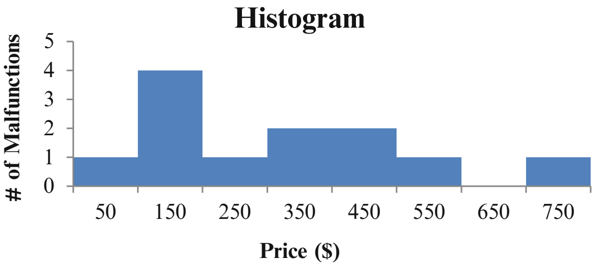
Remove Legend

➢ Right click on the legend and click the **Delete** option



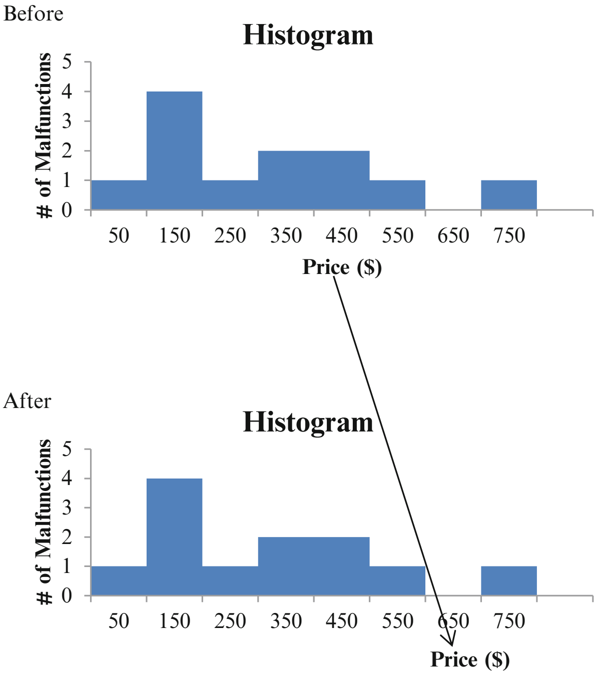
Axes Labels

➢ To change the axes labels click the frequency label (y) or bin label (x) and type the appropriate titles



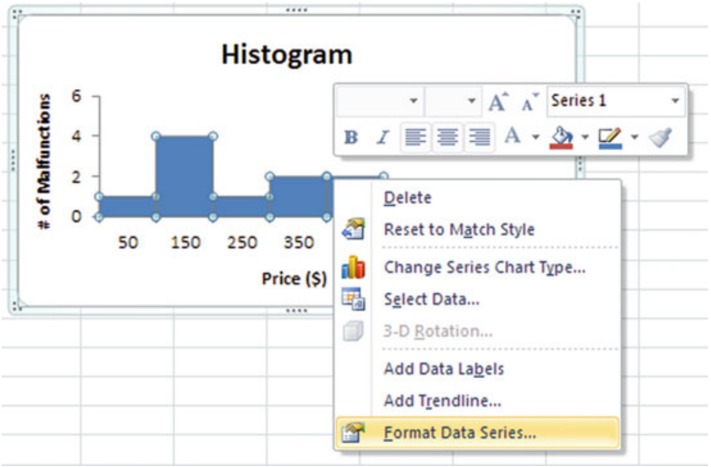
Moving Axes Labels

Left click on the axis you want to move; drag and drop



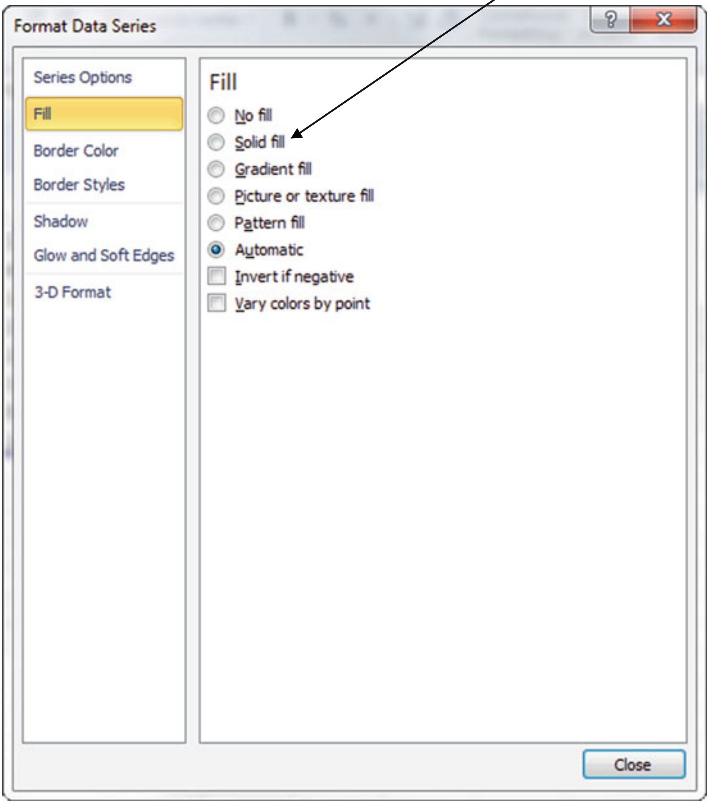
Changing the Bar Color

Right click on any column. Choose the **Format Data Series** option



Select **Fill**.

- ➢ Change from **Automatic** to the fill you prefer for the bins; if you click on **Solid Fill** you can choose from a variety of solid colors by clicking on the color
- ➢ Click **Close**, when you are finished

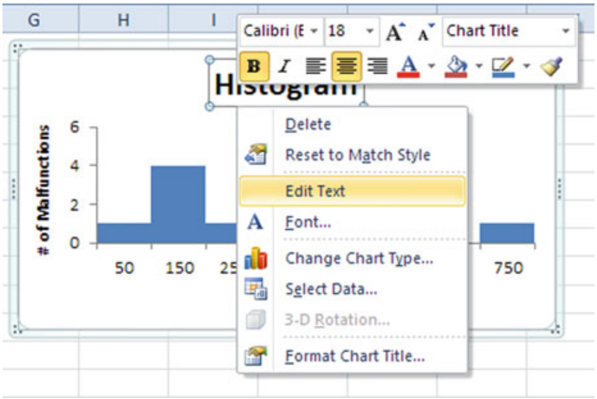


Fill

○ No fill
○ Solid fill
○ Gradient fill
○ Picture or texture fill
○ Pattern fill
● Automatic
□ Invert if negative
□ Vary colors by point

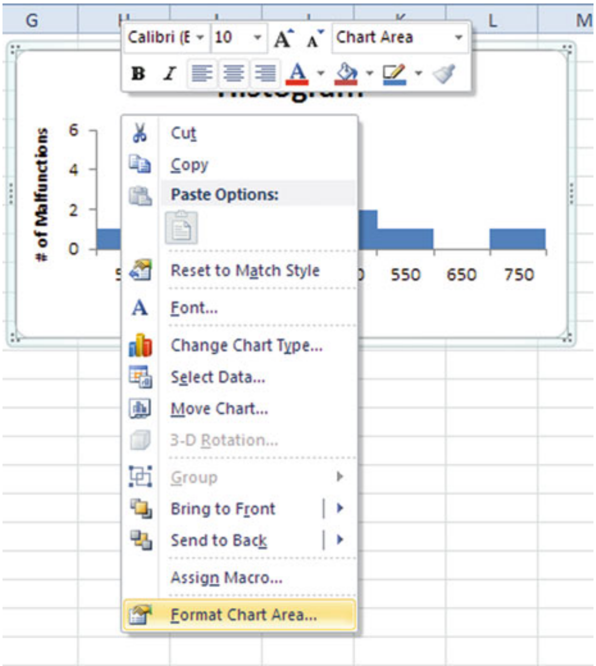
Changing Chart Title

➢ Right click on the chart title. Select the **Edit Text** option; you can retype the title as you choose but make sure to click outside the text box when done typing

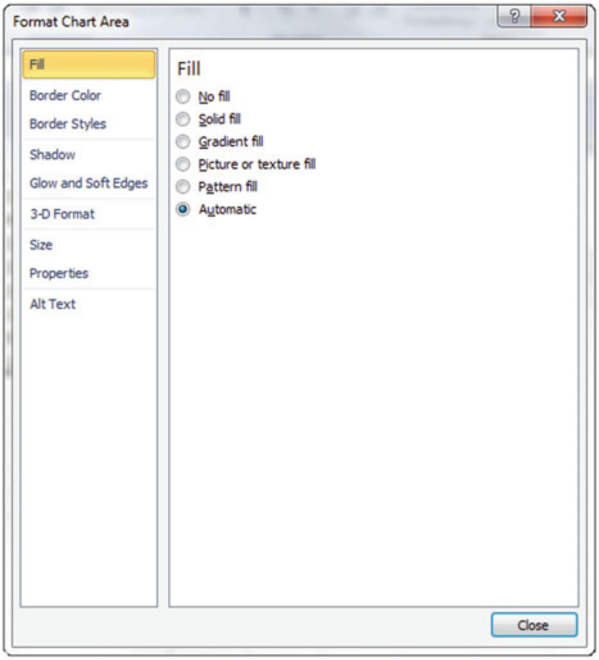


Changing Chart Background Fill

➢ Right click on the chart background. Click the **Format Chart Area** option



- ➢ Select **Fill** from the left hand menu
- ➢ Select the type of fill and color options you desire



Format Chart Area

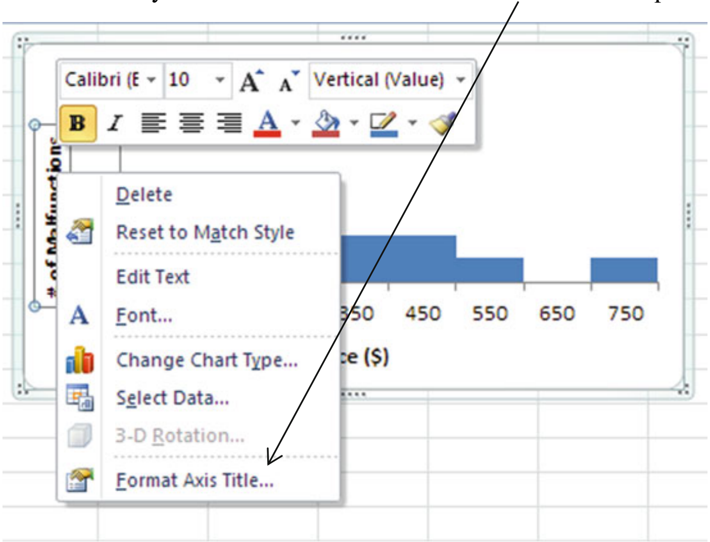
Fill

○ No fill
○ Solid fill
○ Gradient fill
○ Picture or texture fill
○ Pattern fill
● Automatic

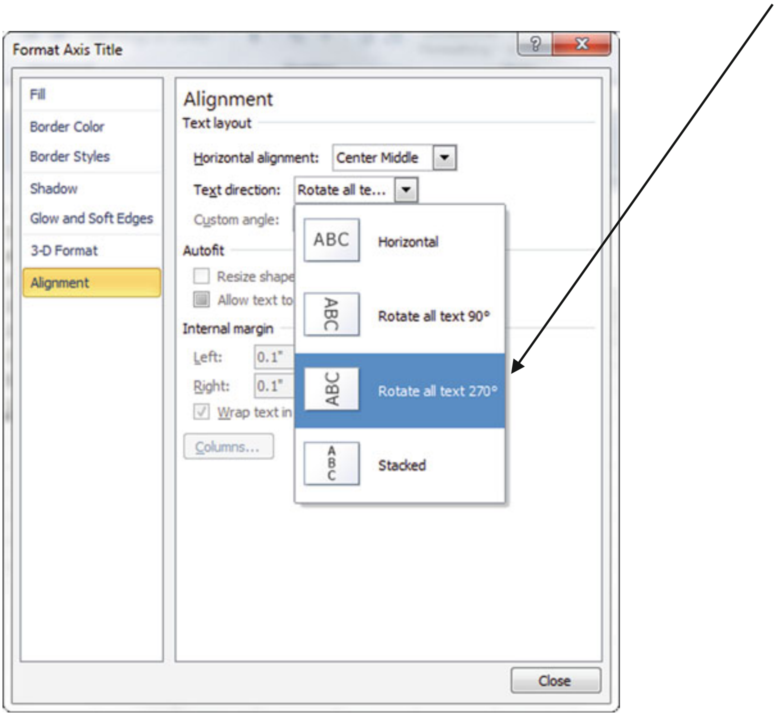
- ➢ Click **Close** when you are finished

Rotating the y-axis Label from Vertical to Horizontal

- ➢ Right click on the y axis label. Click on the **Format Axis Title** option



➢ Select **Alignment** from the left hand menu**;** you can change the alignment and direction



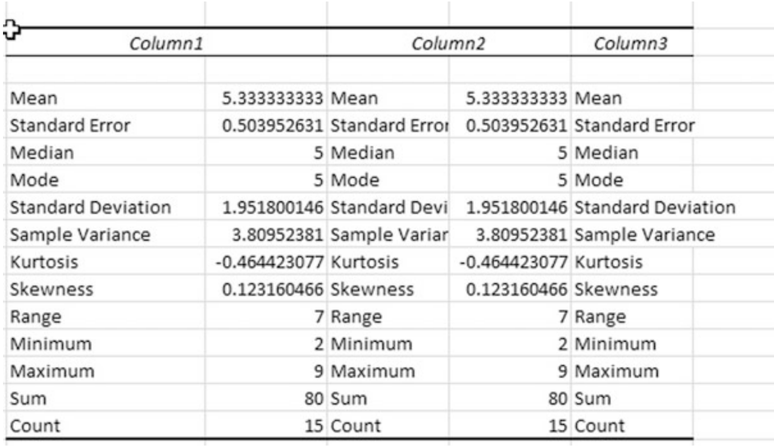
Horizontal alignment: Center Middle

Text direction: Rotate all te...

Common Pitfalls

- Be very careful when calculating the mode. When there is more than one value with the same frequency (more than one mode) in a data set, Excel will report the value that appears first in your data set. If each data value appears with the same frequency Excel will report **#N/A**.
  - {1,1,1,2,2,2,3,3,4} Excel will report that the mode is 1
  - {1,2,3,3,3,4,5,6,7} Excel will report that the mode is 3
  - {1, 2, 3, 4, 5, 6} Excel will report #NA
- The formula used in Excel for STDEV.S to calculate the standard deviation uses a denominator of (n-1), whereas STDEV.P uses the denominator of (N). The P stands for population and the S stands for sample. When you calculate the standard deviation by hand you typically use (N) in the denominator of a population and (n-1) for a sample. These two methods will yield slightly different answers, so be careful when deciding which formula to use.

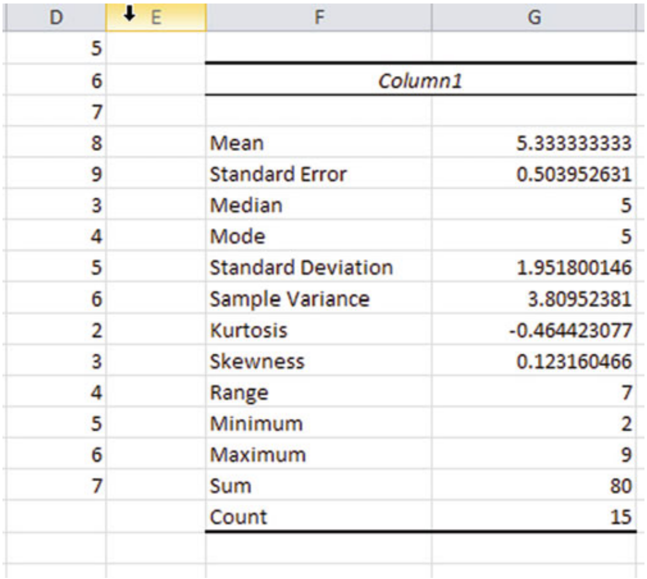
- Label all axes with variable names and units. Excel will not label the axes by default.
- Excel reports the maximum value for each interval as the bin value not the midpoint.
- Excel will include a bin labeled **More**. Rename the **More** bin as a numerical label to avoid confusion.
- Excel labels all frequency graphs as histograms. Remember bar graphs can have spaces between bins. Histograms have no spaces between bins. You need to close the gaps in the histogram to zero.
- Be careful to tell Excel if you are including labels as the first set of cells in the input. Always check the count (n) to make sure you have all of your data is included in the Excel calculations.
- If you expand your output area Excel will keep repeating the results to fill up the area. In Fig. 3.10 the data labeled as Column 1 is the same output as Column 2. There is no need to show duplicate data. You can avoid this by keeping the output area large enough for just one set of output data. To expand the columns you need to right click on the top of the column and change the width. If you simply drag the lowered left corner of the output box you will generate repeated data.



Column1		*Column2*		*Column3*
Mean	5.333333333	Mean	5.333333333	Mean
Standard Error	0.503952631	Standard Erroı	0.503952631	Standard Error
Median	5	Median	5	Median
Mode	5	Mode	5	Mode
Standard Deviation	1.951800146	Standard Devi	1.951800146	Standard Deviation
Sample Variance	3.80952381	Sample Variar	3.80952381	Sample Variance
Kurtosis	-0.464423077	Kurtosis	-0.464423077	Kurtosis
Skewness	0.123160466	Skewness	0.123160466	Skewness
Range	7	Range	7	Range
Minimum	2	Minimum	2	Minimum
Maximum	9	Maximum	9	Maximum
Sum	80	Sum	80	Sum
Count	15	Count	15	Count

Fig. 3.10 Output area. Output from dragging the lower right hand corner of the output area

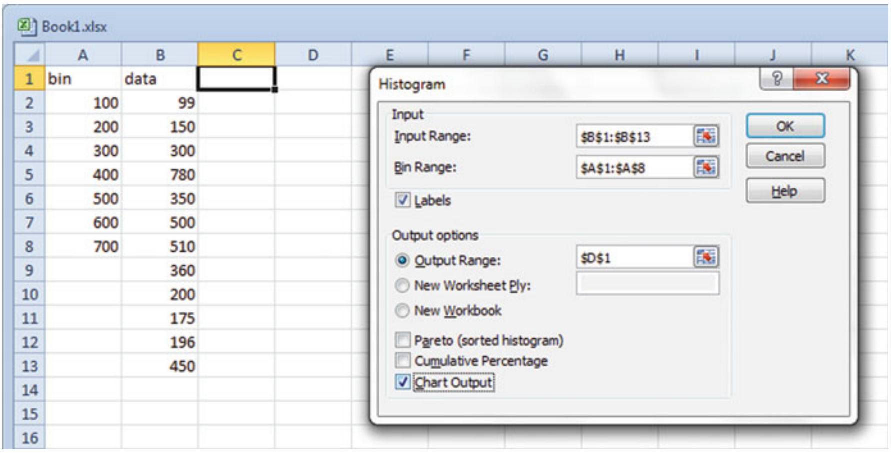
- Too many decimal places. Excel will calculate as many decimal places as you indicate in the output format. If no decimal limit is set, Excel will fill the cell and show nine decimal places. In Fig. 3.11 the input data does not have any decimal places and yet the Excel output is reported in nine decimal places. This obviously does not make any sense; Excel is reporting a greater number of significant figures than we had recorded in the data. You cannot magically create a higher level of accuracy and precision in your data than exists in the data you are using for the analysis. The rule of thumb is to allow one more decimal place in the output than exists in the input data.



D	E	F	G
5			
6		*Column1*	
7			
8		Mean	5.333333333
9		Standard Error	0.503952631
3		Median	5
4		Mode	5
5		Standard Deviation	1.951800146
6		Sample Variance	3.80952381
2		Kurtosis	-0.464423077
3		Skewness	0.123160466
4		Range	7
5		Minimum	2
6		Maximum	9
7		Sum	80
		Count	15

Fig. 3.11 Decimal places

◇ If you check **Labels** in the histogram input window, you need to include a label in your end point bin data; otherwise, Excel will assume the first bin end point is a label and will not use this value in the calculations.



Book1.xlsx

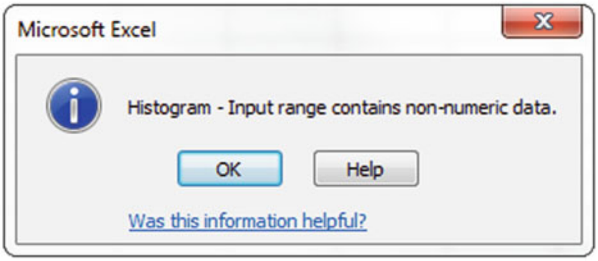
	A	B	C
1	bin	data	
2	100	99	
3	200	150	
4	300	300	
5	400	780	
6	500	350	
7	600	500	
8	700	510	
9		360	
10		200	
11		175	
12		196	
13		450	
14			
15			
16			

Histogram

Input
Input Range: $B$1:$B$13
Bin Range: $A$1:$A$8
☑ Labels

Output options
● Output Range: $D$1
○ New Worksheet Ply:
○ New Workbook
□ Pareto (sorted histogram)
□ Cumulative Percentage
☑ Chart Output

◇ If you don't check the Labels box, and you include labels in your input range, Excel will create a warning message.



Histogram - Input range contains non-numeric data.

Histogram Checklist.

Do not include a legend unless it is really necessary.
Label both axes with units and variable names.
Make the graph easy to read; clean it up.
If you will be copying the graph in black and white make sure you can still differentiate data without the color. You can use different patterns rather than different colors.

Final Thoughts and Activities

Practice Problems

1. The data in the **Beverage** file represents the amount of fluid in a sample of fifty 2-l bottles. Using the data from the file, complete the following exercises.
   (a) Construct a frequency distribution.
   (b) From the output, what is the mean?
   (c) From the output, what is the standard deviation?
2. The data in the **Electric** file represents the cost of electricity in June for a random sample of fifty one-bedroom apartments in a large city in the southern portion of the United States. Using the data from the file complete the following exercises.
   (a) Construct a frequency distribution.
   (b) Construct a histogram.
   (c) Around what amount does the monthly electricity cost seem to center?

Discussion Boards

1. Today no one wants to be considered average but in 1835 the average man was a symbol of an egalitarian society. According to Quetelet, a French statistician in the nineteenth century, "If an individual at any given epoch in society possessed all the qualities of the average man, he would represent all that is good, great or beautiful." How would you define the average employee in your current department?
2. Are TV commercials getting longer, or does it just seem that way? In 1990, 6 % of commercials were a maximum of 1 min long. Recently, out of a sample of 50 commercials only 16 were 1 min or shorter.
3. Class size Paradox. There are two ways to obtain a mean class size and you need to choose one for your training department's report. In one approach you take the number of employees in all 25 seminars offered this year and average those results. The second approach is on an employee by employee basis, compile a list of class sizes for all seminars he or she attended during the year and average those results. The results for the two approaches will be different. Discuss.

Group Activity

1. The Central Intelligence Agency has specialists who analyze the frequencies of letters of the alphabet in an attempt to decipher intercepted messages. For example in Standard English text the letter "r" is used at the rate of 7.7 %. Choose another letter and using the Internet determine its usage rate. In an intercepted message sent to Iraq, a page of 2,000 characters is found to have the letter "f" occurring 42 times. Discuss if this is unusual.
2. What is meant by "Six degrees of separation" and "The Small World Problem?"
3. The Cost of Laughing Index (CLI) is developed using the same approach as is used to develop the Consumer Price Index. While standard scores and percentiles allow us to compare different values, they do not take into account time. Index numbers such as the CLI allow you to compare the current value to a value at some base time period. Describe the CLI methodology and what it is this year in the United States.

Parting Thought

Figures don't lie, but liars can figure.

Problem Solutions

1. The data in the **Beverage** file represents the amount of fluid in a sample of fifty 2-l bottles. Using the data from the file, complete the following exercises.

   (a) Construct a frequency distribution.

Bin	Frequency
2.25	1
2.28	1
2.31	6
2.34	11
2.37	15
2.40	9
2.43	5
More	2

Amount	
Mean	2.35672
Standard error	0.006302254
Median	2.36
Mode	2.368
Standard deviation	0.044563662
Sample variance	0.00198592
Kurtosis	0.080400572
Skewness	0.008832396
Range	0.215
Minimum	2.25
Maximum	2.465
Sum	117.836
Count	50

   (b) From the output, what is the mean?

   Answer: Mean: 2.36 bottles but we can't have .36 bottles so just 3 bottles

   (c) From the output, what is the standard deviation?

   Answer: sd = 0.04 bottles

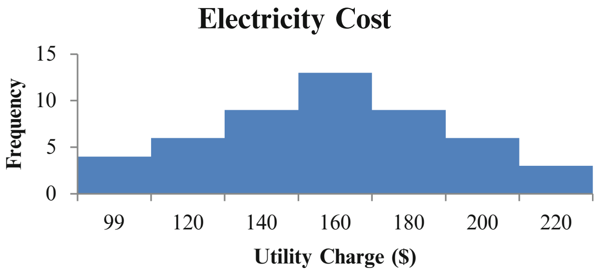
2. The data in the **Electric** file represents the cost of electricity in June for a random sample of fifty one-bedroom apartments in a large city in the southern portion of the United States. Using the data from the file complete the following exercises.

(a) Construct a frequency distribution.

Bin	Frequency
85	1
103.71	3
122.43	7
141.14	8
159.86	12
178.57	10
197.29	5
More	4

Utility charge	
Mean	150.06
Standard error	4.481837269
Median	151.5
Mode	133
Standard deviation	31.69137525
Sample variance	1004.343265
Kurtosis	−0.544163238
Skewness	0.015845641
Range	131
Minimum	85
Maximum	216
Sum	7,503
Count	50

(b) Construct a histogram.



(c) Around what amount does the monthly electricity cost seem to center around?

Answer: The majority of utility charges are clustered around $160.

Chapter 4
Normal Distributions

Key Concepts

Chebyshev approximation, Normal distribution, Percentiles, Standard deviation, Standard units, and z-score.

Discussion

This chapter will discuss one of the most common distributions in the business world, the “normal curve”. This is another example where the everyday usage of the word means something quite different from the statistical definition. In statistics the term “normal” refers to a specific mathematically defined curve that looks bell-shaped. An entire chapter is dedicated to this topic because of the normal distribution’s usefulness in many different applications throughout the business world. The area under the curve is often described as the probability and will be discussed in more detail in Chap. 8.

As a basis for comparison we provide a brief description of a non-normal technique for analysis, the Chebyshev approximation.

Normal Distribution: This is a bell-shaped or Gaussian distribution (Fig. 4.1).

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_4,

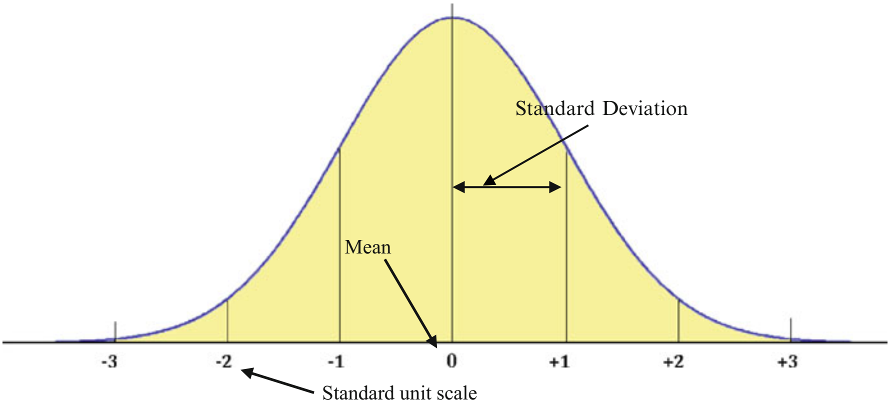


Fig. 4.1 Normal curve

The normal distribution is often described as a central region with a left hand tail and right hand tail. The boundary between the central region and the tails is usually defined by the particular problem you are trying to solve.

The **mean** and the **standard deviation** of the distribution describe the bell curve.

Mean (x̄): The arithmetic average value of the x variable. This pins down the center of the bell shape. When using the standard scale unit, the mean is always zero.

Standard Deviation (σ, sd, SD): The sd defines the width of the bell shape for a sample.

$$sd = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

where:

Σ is the capital Greek letter sigma, which indicates summing up all variables that follow it. The subscript and superscript indicate you should add the first value through the "nth" value. Often the short hand version just shows the summing sign without the sub/superscripts.

$\boldsymbol{x_i}$ is the variable you are summing up the values for and the subscript is just a place holder. So if the subscript was "1", it means the first value of x.

$\bar{\boldsymbol{x}}$ is the average or mean value of x

n is the number of values you are summing together, more commonly referred to as the sample size

Standard (std.) Units: Several features of this distribution are very important.

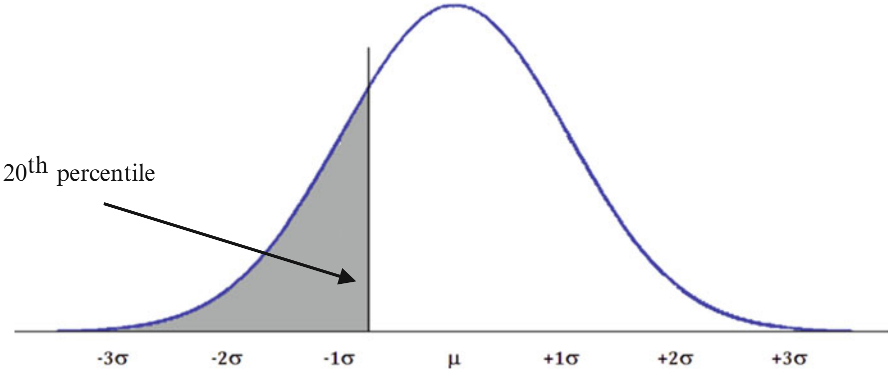
1. The graph is symmetric about 0. The part of the curve on the left side of 0 is the mirror image of the right side of the curve.
2. The total area under the curve is 100 %.

3. The area under the curve between −1 and +1 std. units is about 68 %.
4. The area under the curve between −2 and +2 std. units is about 95 %.
5. The area under the curve between −3 and +3 std. units is about 99 %.

An individual value is converted to standard units by seeing how many standard deviations it is above or below the average. Remember the standard deviation is a group statistic that allows you to convert individual values in the data set to standard units.

z-score: The z-score or standard score is found by converting a value from its measured units to standard units. In other words a z-score is the number of standard deviations a given value of x is above or below the mean. In addition, z-scores provide the key to using a table of standard normal probabilities to perform calculations related to normal distributions.

Percentile: The value of a variable, below which a certain percent of observations fall. For example, the 20th percentile is the value (or score) which indicates that 20 % of the observations are found below (or less than) that value.



Chebyshev approximation: This approximation can be used with any distribution. Not all distributions are normally distributed, or you simply may not know what the distribution looks like. To err on the side of conservative estimates, it is better to use the **Chebyshev approximation** in these cases. Because of its general application across all distributions, the answers are not as precise. By this, we mean that the answer you calculate will indicate the MINIMUM amount of data, in other words it defines the lowest value of how much data you should expect for your range of interest. So for example, if you want to know how much data falls between ±2 standard deviations in your data set, the solution will indicate **at least** 75 % of the data will be located between these values, regardless of the distribution. For non-normal problems, we use the symbol ***k*** for the standard deviation. Thanks to Chebyshev, we know that for any non-normal distribution:

> At least $(1-1/k^2)$ of the items in any data set will be within k standard deviations of the mean, where k is any value greater than 1.

- At least 75 % of the items must be within $k = \pm 2$ standard deviations of the mean.
- At least 89 % of the items must be within $k = \pm 3$ standard deviations of the mean.
- At least 94 % of the items must be within $k = \pm\ 4$ standard deviations of the mean.

Excel

When we have normally distributed data, the normal curve allows us to calculate how much data falls between any two points, and how much data exists in one or both tails. We can also easily calculate specific percentiles. But remember, we need to work in **standard units**. When you have all of the raw data, you can use Excel to calculate the standard deviation and the mean. However if you don't have the raw data you need to be given the mean and the standard deviation to use these functions.

The functions we will use in this section apply to both cases where you do not have the raw data and also to those cases where you do have all of the raw data available in Excel:

- **NORM.DIST** calculates the probability or distribution of data to the left of your value and requires the mean and standard deviation statistics. In other words, this calculates the percentile. The data is input using the following format.

  *NORM.DIST (the **x value** in measured units, the **mean** in measured units, **standard deviation** in measured units, **1**)*

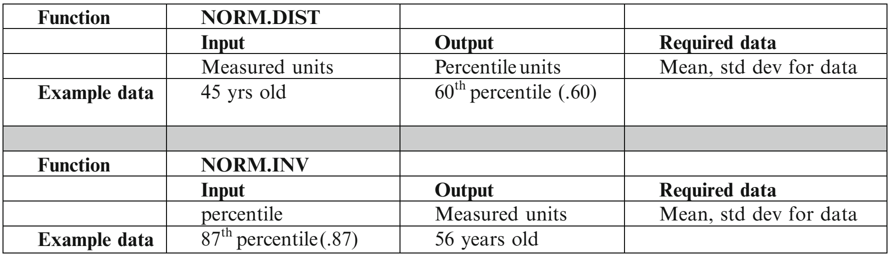
  The argument 1 (TRUE) tells Excel to compute the normal cumulative distribution. If the last argument is 0 (FALSE), Excel returns the actual value of the normal random variable.

  ◈ The answer is given as a decimal so you will need to multiply by 100 to get the percentage answer.

- **NORM.INV** will convert the percentile value to measured units. The data is input using the following format.

  *NORM.INV (the **percentile** to be converted to measured units and expressed as a decimal value, the **mean** in original measured units, **standard deviation** in original measured units)*

These Excel functions are summarized by way of examples in Fig. 4.2. In these examples we assume a dataset which has a normal distribution of employee ages.



Function	**NORM.DIST**		
	Input	**Output**	**Required data**
	Measured units	Percentile units	Mean, std dev for data
Example data	45 yrs old	60th percentile (.60)	
Function	**NORM.INV**		
	Input	**Output**	**Required data**
	percentile	Measured units	Mean, std dev for data
Example data	87th percentile(.87)	56 years old	

Fig. 4.2 Sample input/output

◇ Note that other Normal functions exist in Excel (i.e. NORM.S.DIST, NORM.S.INV) in which all input/output are in standard units. However, normal distribution analysis can be adequately handled using the NORM.DIST and NORM.INV functions.

The problem types using these Excel functions include:

1. **Percentile Calculation Problems (NORM.DIST) including creating area graphs**
2. **Converting Percentiles to Measured Units (NORM.INV)**
3. **Converting Measured units to z-scores (STANDARDIZE)**
4. **Rank and Percentile (RANK, PERCENTILE)**
5. **Non-normal Distributions**

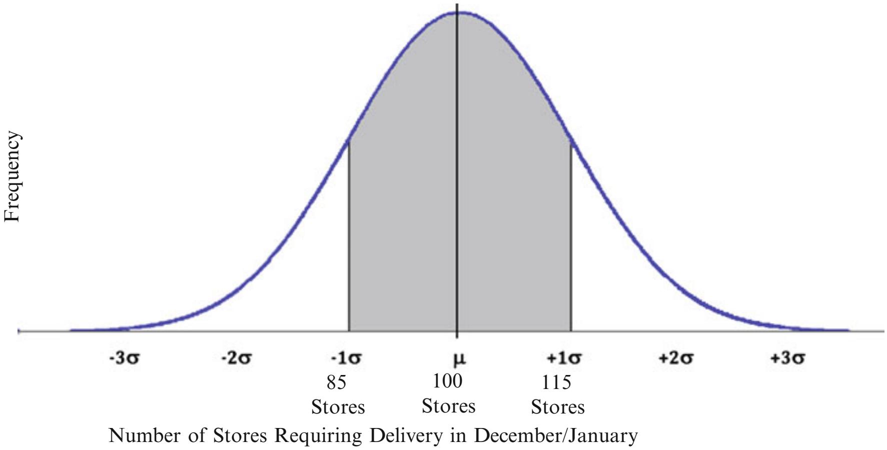
1. Percentile Calculation Problems (NORM.DIST)

There are several distribution (area) types that use this function. These include:

(a) Calculating the area to the left of a value
(b) Calculating the area between two values
(c) Calculating the area to the right of a value
(d) Creating a normal distribution or area graph

Example Problem

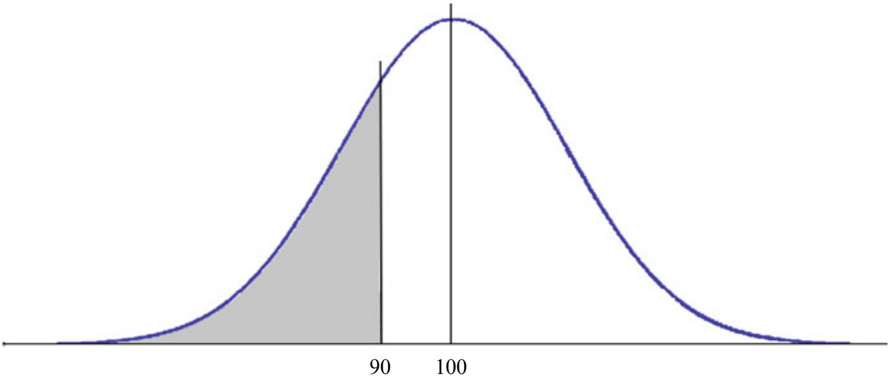
Your transportation company provides delivery service 7 days a week to stores selling synthetic fire logs. You need to determine how many trucks and drivers to have on hand to deliver the product. The data you have collected from last year indicate the number of stores requiring deliveries on a daily basis during the months of December and January. This data seems to follow a normal distribution. The average number of stores requiring deliveries on any given day during this time is 100. The standard deviation is 15 stores.



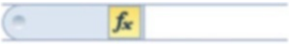
◈ Note to answer these questions you don't actually need a data set as long as you know the data follows a normal distribution. Very powerful and helpful analysis tool.

(a) Calculating the area to the left of a value

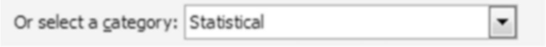
What % of the Time Did You Deliver to Less Than (<) 90 Stores During Last December and January? In Other Words How Much Data Is in the Left Hand Tail?



- ➢ Click on an empty cell where you want the answer to be output
- ➢ Click on the *fx* button

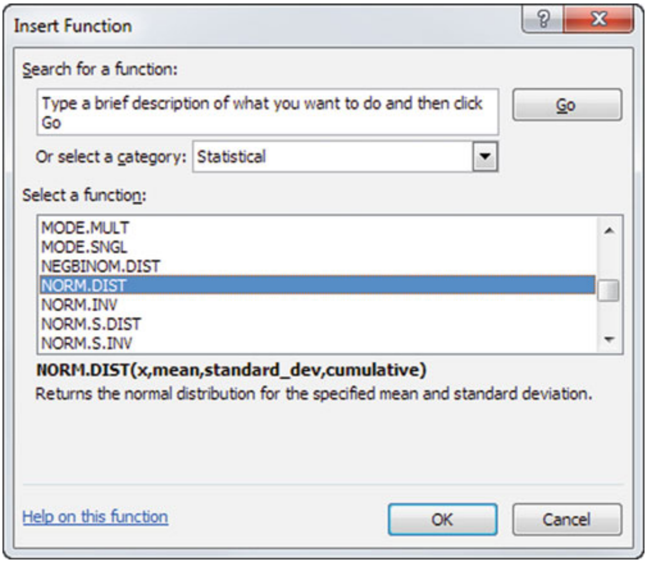


- ➢ Select the **Statistical** category



Or select a category: Statistical

- ➢ Highlight the **NORM.DIST** function from the list or just type in the **NORM.DIST** function in the function bar



Insert Function

Search for a function:

Type a brief description of what you want to do and then click Go

Or select a category: Statistical

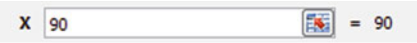
Select a function:

MODE.MULT
MODE.SNGL
NEGBINOM.DIST
NORM.DIST
NORM.INV
NORM.S.DIST
NORM.S.INV

NORM.DIST(x,mean,standard_dev,cumulative)

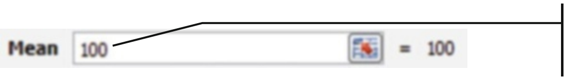
Returns the normal distribution for the specified mean and standard deviation.

- ➢ Click **OK**
- ➢ Insert the value of X in original measured units. This is the upper limit of the area for which you want to calculate the distribution of data. In other words, this is the value you want to convert to a percentile



X 90 = 90

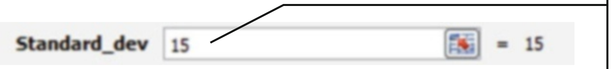
- ➢ Insert the **Mean** value



Mean 100 = 100

Note: You must already have the *Mean* value.

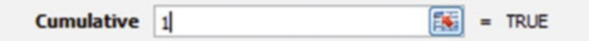
➢ Insert the **Standard_dev** value



Standard_dev 15 = 15

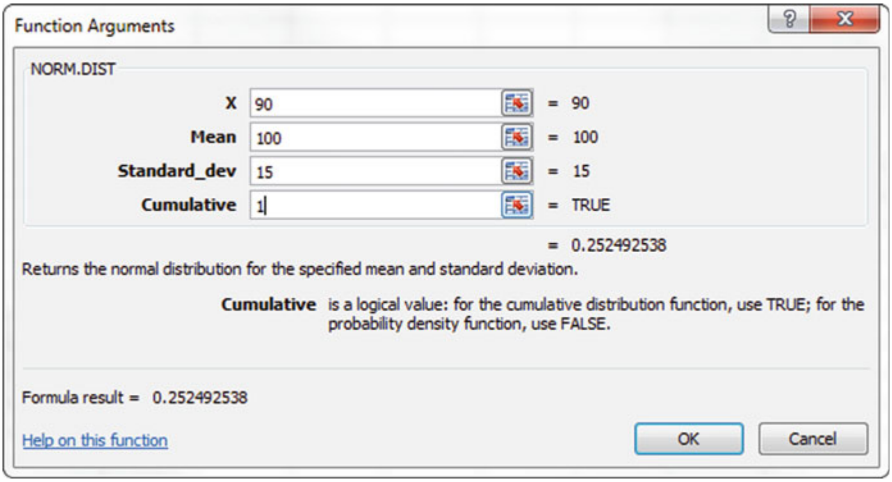
Note: You must already have the *Standard_dev* value.

➢ Input the value of "**1**" or the word "**TRUE**" for normal cumulative calculations such as this one. Since this is a percentile calculation we select **1** as the last value in the function input.



Cumulative 1 = TRUE

➢ Click **OK**



Function Arguments

NORM.DIST

X 90 = 90

Mean 100 = 100

Standard_dev 15 = 15

Cumulative 1 = TRUE

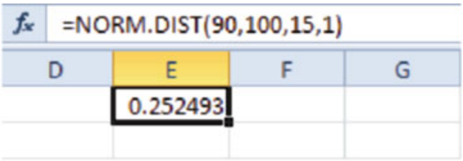
= 0.252492538

Returns the normal distribution for the specified mean and standard deviation.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability density function, use FALSE.

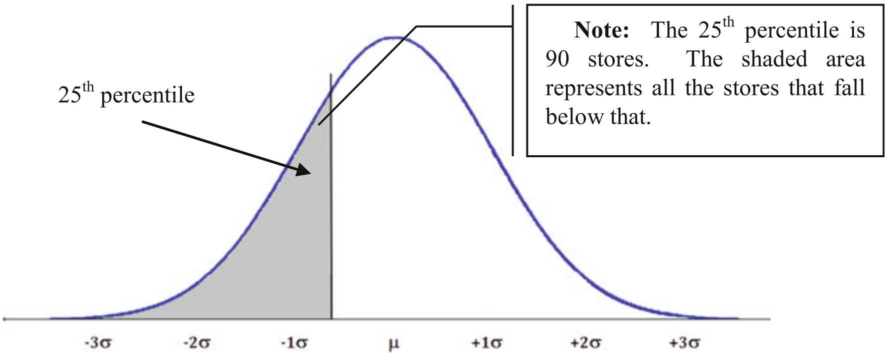
Formula result = 0.252492538

Based on this data, the answer is given as .252493 which translates to about 25 % of the time you delivered fire logs to less than 90 stores.



=NORM.DIST(90,100,15,1)

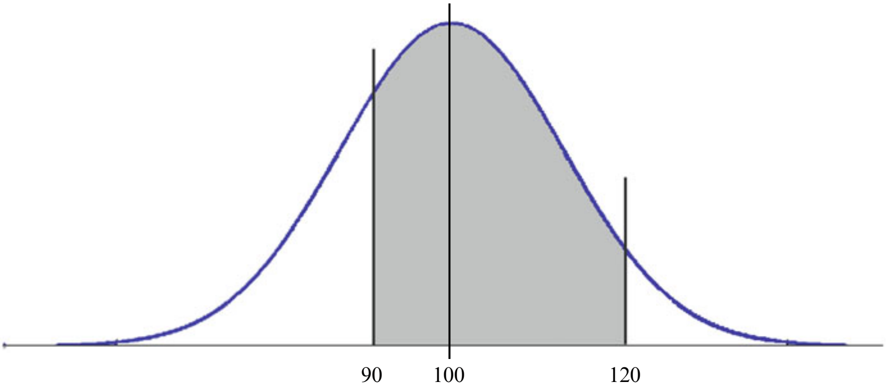
D	E	F	G
	0.252493		



Note: The 25th percentile is 90 stores. The shaded area represents all the stores that fall below that.

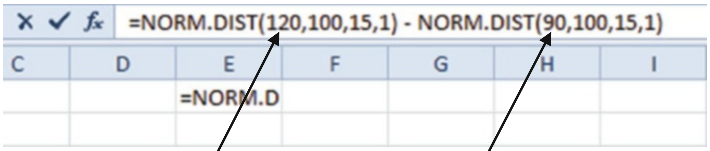
(b) Calculating the area between 2 values

What % of the Time Did You Deliver Fire Logs to Between 90 and 120 Stores During Last December and January?



- ➢ Click on the empty cell where you want the answer
- ➢ Type "=" inside the ***fx*** box (this prepares Excel to accept a function format)
- ➢ Type in the formula NORM.DIST (limit value, mean, sd, 1)

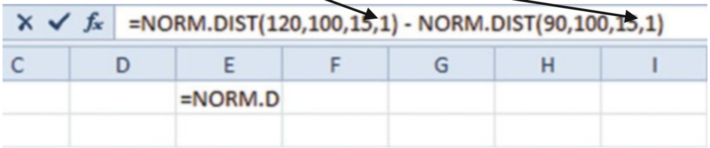
- Input the upper limit value and the lower limit value of the area where you want to calculate the data distribution
- Input the mean and standard deviation from your data set inside both sets of brackets as the next 2 inputs



=NORM.DIST(120,100,15,1) - NORM.DIST(90,100,15,1)

=NORM.D

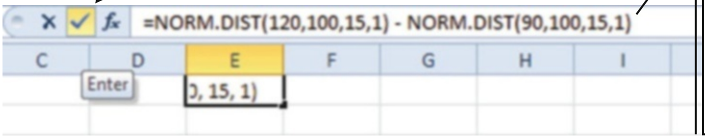
- Input the value of **1** or the word **TRUE** for normal cumulative calculations inside both sets of brackets



=NORM.DIST(120,100,15,1) - NORM.DIST(90,100,15,1)

=NORM.D

- Click the **checkmark** to the left of ***fx***

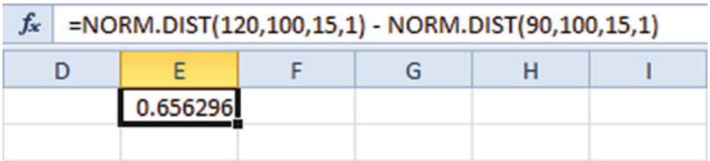


=NORM.DIST(120,100,15,1) - NORM.DIST(90,100,15,1)

), 15, 1)

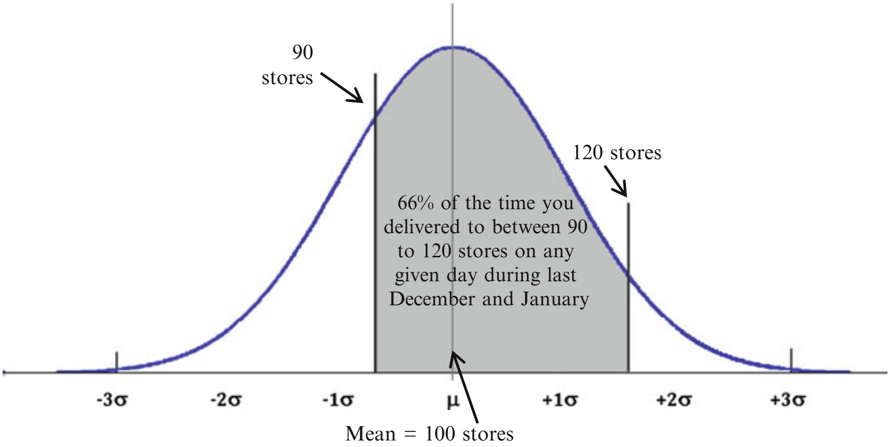
Note: The formula is entered "**NORM.DIST (Limit, Mean, SD, 0/1)**". Enter either **1** for true or **0** for false.

The answer is given as 0.656296 which translates to during last December and January you delivered fire logs to between 90 and 120 stores about 66 % of the time.



=NORM.DIST(120,100,15,1) - NORM.DIST(90,100,15,1)

D	E	F	G	H	I
	0.656296				



90 stores
120 stores
66% of the time you delivered to between 90 to 120 stores on any given day during last December and January
-3σ
-2σ
-1σ
μ
+1σ
+2σ
+3σ
Mean = 100 stores

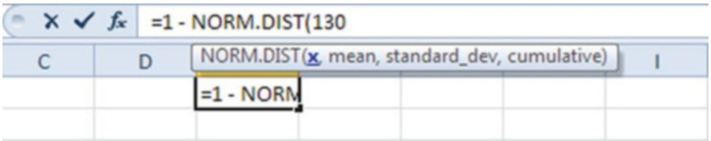
(c) Calculating the area to the right of a value

What % of Time Did You Deliver to 130 or More Stores (≥) During Last December and January? In Other Words How Much Data Is in the Right Hand Tail?

◇ Note in these type of problems the Excel answer will include when you are delivering to exactly 130 stores and when you are delivering to more than 130 stores.

➢ Click on an empty cell where you want the answer
➢ Click on the ***fx*** function

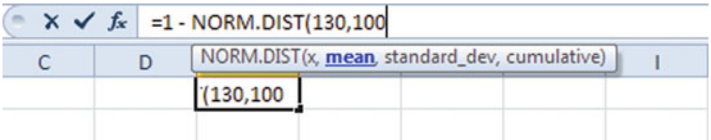
➢ Type the "=" in the white input box
➢ Type the number "**1**" which equate to 100 %
➢ Type the function or highlight the function from the drop down menu and subtract from **1**
➢ Input the upper limit of the right hand tail


=1 - NORM.DIST(130

NORM.DIST(x, mean, standard_dev, cumulative)

=1 - NORM

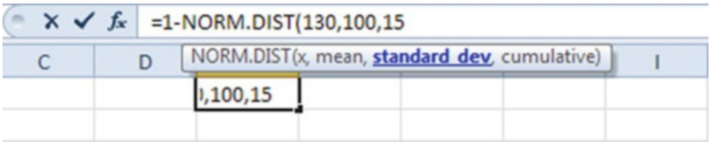
- ➢ Input the mean from your data set


=1 - NORM.DIST(130,100

NORM.DIST(x, mean, standard_dev, cumulative)

'(130,100

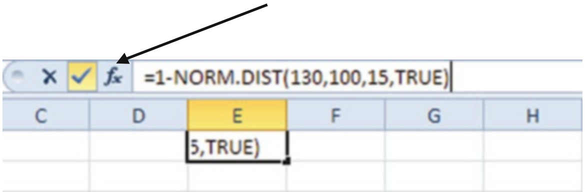
- ➢ Input the standard deviation from your data set


=1-NORM.DIST(130,100,15

NORM.DIST(x, mean, standard_dev, cumulative)

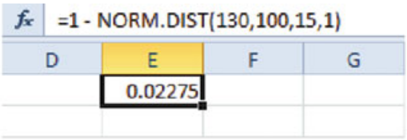
),100,15

- ➢ Input the value of "**1**" or the word "**TRUE**" for the normal cumulative calculation
- ➢ Click the **checkmark** to the left of ***fx***


=1-NORM.DIST(130,100,15,TRUE)

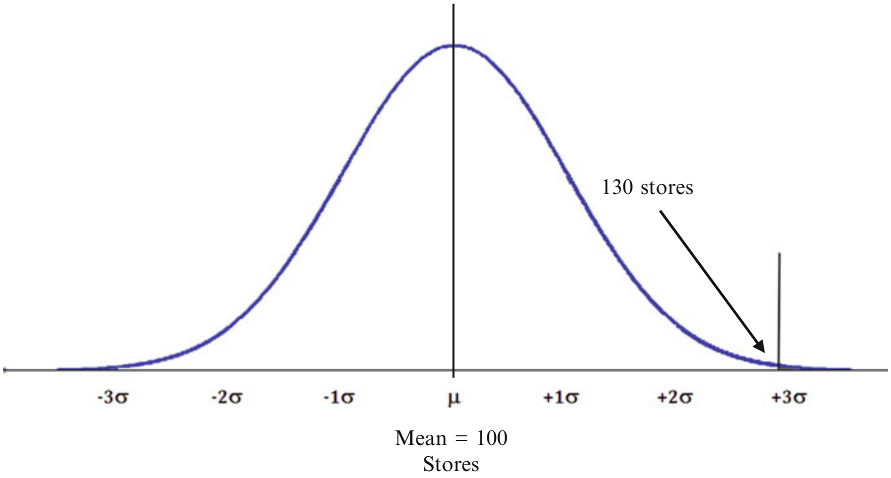
5,TRUE)

The answer is given as 0.02275 which translates to about 2.3 % of the time you delivered fire logs to 130 or more stores on any given day. If we expect the same results this year, we could state that we will be delivering to *at least* 130 stores on any given day about 2.3 % of the time this coming December and January.



=1 - NORM.DIST(130,100,15,1)

0.02275



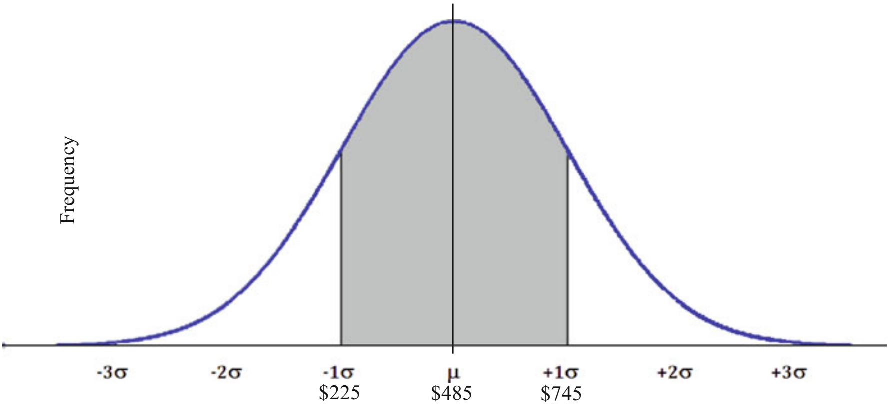
(d) Graphing a normal distribution (Area Graph)

There are three steps in this process:

Step 1. Sort the data into a sequential order (SORT)
Step 2. Calculate the height of the normal distribution for each x value (NORM. DIST)
Step 3. Create the graph

Example Problem

You would like to graph the distribution of the cost of airplane tickets purchased by employees over the past 3 months. The average cost is $485 and the standard deviation is $260. You know that ticket prices are normally distributed based on historical data.



To do this in Excel will require several steps.

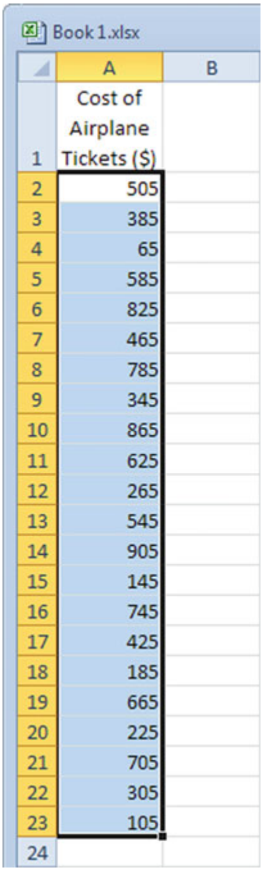
➢ Input the following ticket prices in column A

Cost of airplane tickets ($)
505
385
65
585
825
465
785
345
865
625
265
545
905
145
745
425
185
665
225
705
305
105

Step 1

Sort data into sequential order using the SORT function.

➢ Highlight the column of data to sort

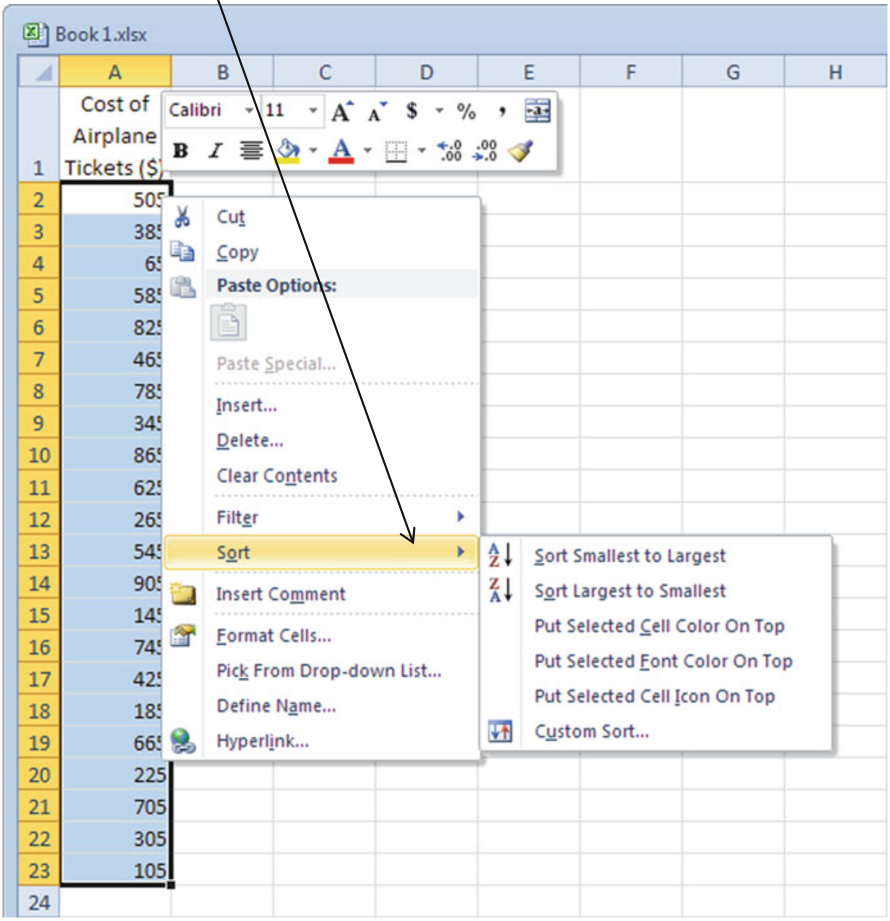


	A	B
1	Cost of Airplane Tickets ($)	
2	505	
3	385	
4	65	
5	585	
6	825	
7	465	
8	785	
9	345	
10	865	
11	625	
12	265	
13	545	
14	905	
15	145	
16	745	
17	425	
18	185	
19	665	
20	225	
21	705	
22	305	
23	105	
24		

- ⟐ If you want to preserve your original data you need to copy and paste it into another column. When Excel sorts the data, it over rides the original order of the data with the sorted order of data.

➢ Right click anywhere in the column

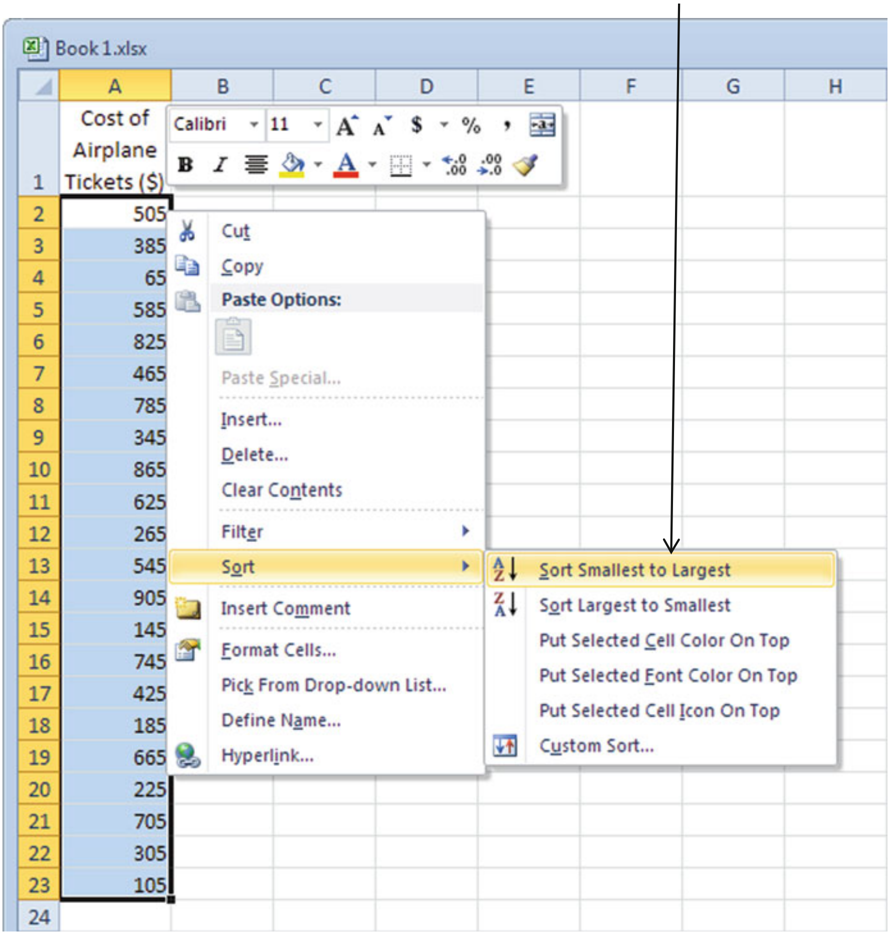
➢ Select the **SORT** function from the menu



Book1.xlsx

	A
1	Cost of Airplane Tickets ($)
2	505
3	385
4	65
5	585
6	825
7	465
8	785
9	345
10	865
11	625
12	265
13	545
14	905
15	145
16	745
17	425
18	185
19	665
20	225
21	705
22	305
23	105
24	

➢ Select which sort option you want. In this case, choose "**Smallest to Largest**"



Book1.xlsx

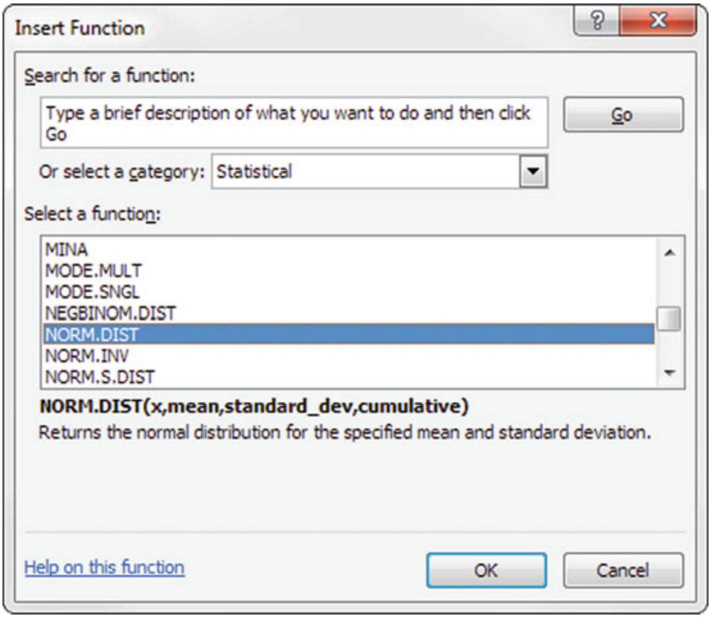
	A
1	Cost of Airplane Tickets ($)
2	505
3	385
4	65
5	585
6	825
7	465
8	785
9	345
10	865
11	625
12	265
13	545
14	905
15	145
16	745
17	425
18	185
19	665
20	225
21	705
22	305
23	105
24	

Step 2

Calculate the height of the normal distribution for each x value using the NORM.DIST function.

➢ Click on an empty cell where you want the output for the first value

➢ Click on the ***fx*** and select **NORM.DIST**


Insert Function

Search for a function:

Type a brief description of what you want to do and then click Go

Or select a category: Statistical

Select a function:

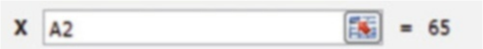
MINA
MODE.MULT
MODE.SNGL
NEGBINOM.DIST
NORM.DIST
NORM.INV
NORM.S.DIST

NORM.DIST(x,mean,standard_dev,cumulative)

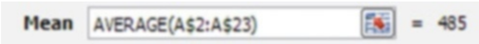
Returns the normal distribution for the specified mean and standard deviation.

➢ Click **OK**

➢ Input the location (cell) of the first x value

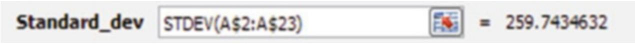

X A2 = 65

➢ Calculate the mean using the **AVERAGE** function


Mean AVERAGE(A$2:A$23) = 485

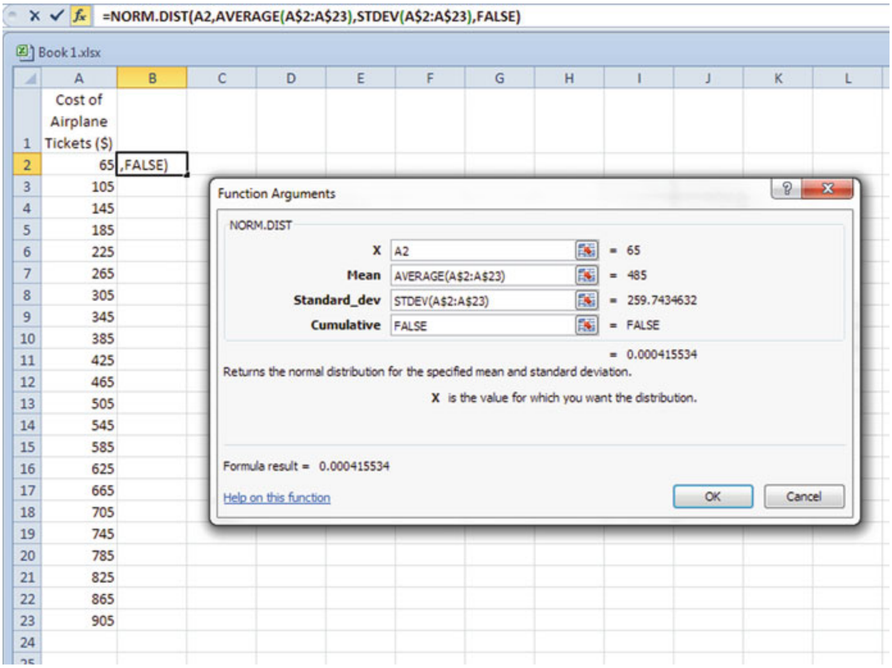
- ◈ Make sure to add “$” with each cell location to treat these cells as constant values when you cut and paste them for use with the other data values.

➢ Calculate the std deviation using the **STDEV** function


Standard_dev STDEV(A$2:A$23) = 259.7434632

Examine the Cell Reference section in Chapter 2 for a refresher on “4” references.

➢ Input **0** (or **FALSE**) to calculate the height of the distribution for each ticket value in this normal distribution



```
=NORM.DIST(A2,AVERAGE(A$2:A$23),STDEV(A$2:A$23),FALSE)
```

Book1.xlsx

	A	B
1	Cost of Airplane Tickets ($)	
2	65	,FALSE)
3	105	
4	145	
5	185	
6	225	
7	265	
8	305	
9	345	
10	385	
11	425	
12	465	
13	505	
14	545	
15	585	
16	625	
17	665	
18	705	
19	745	
20	785	
21	825	
22	865	
23	905	
24		

Function Arguments

NORM.DIST

X	A2	= 65
Mean	AVERAGE(A$2:A$23)	= 485
Standard_dev	STDEV(A$2:A$23)	= 259.7434632
Cumulative	FALSE	= FALSE

= 0.000415534

Returns the normal distribution for the specified mean and standard deviation.

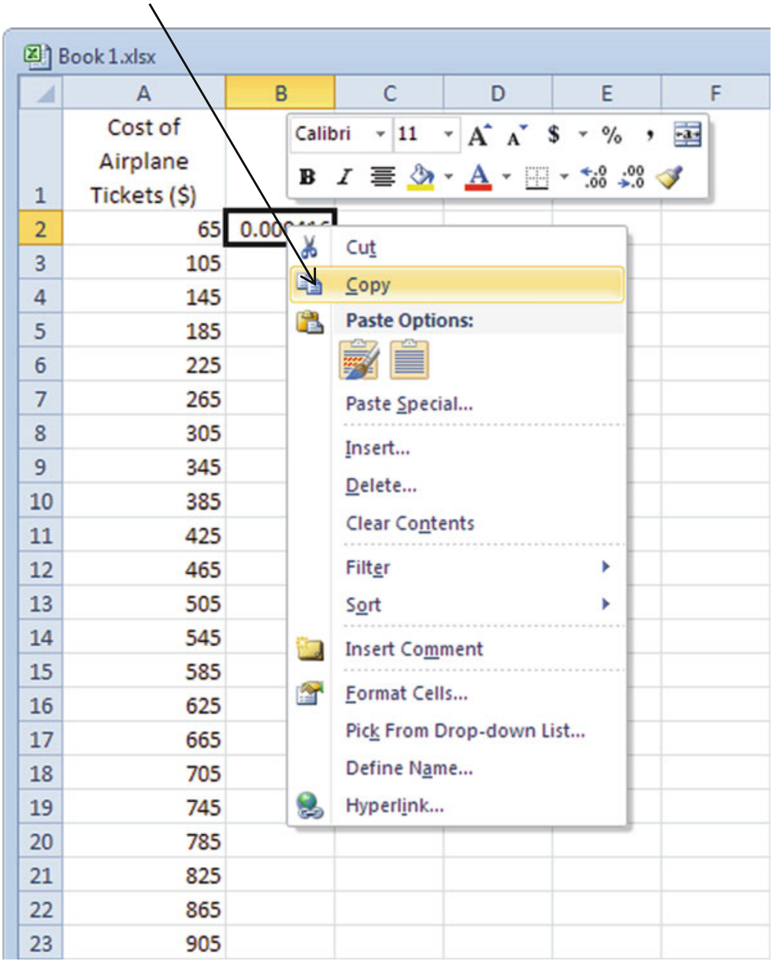
X is the value for which you want the distribution.

Formula result = 0.000415534

➢ Click **OK**

➢ Highlight cell B2 (or wherever your first calculation is located) and right click

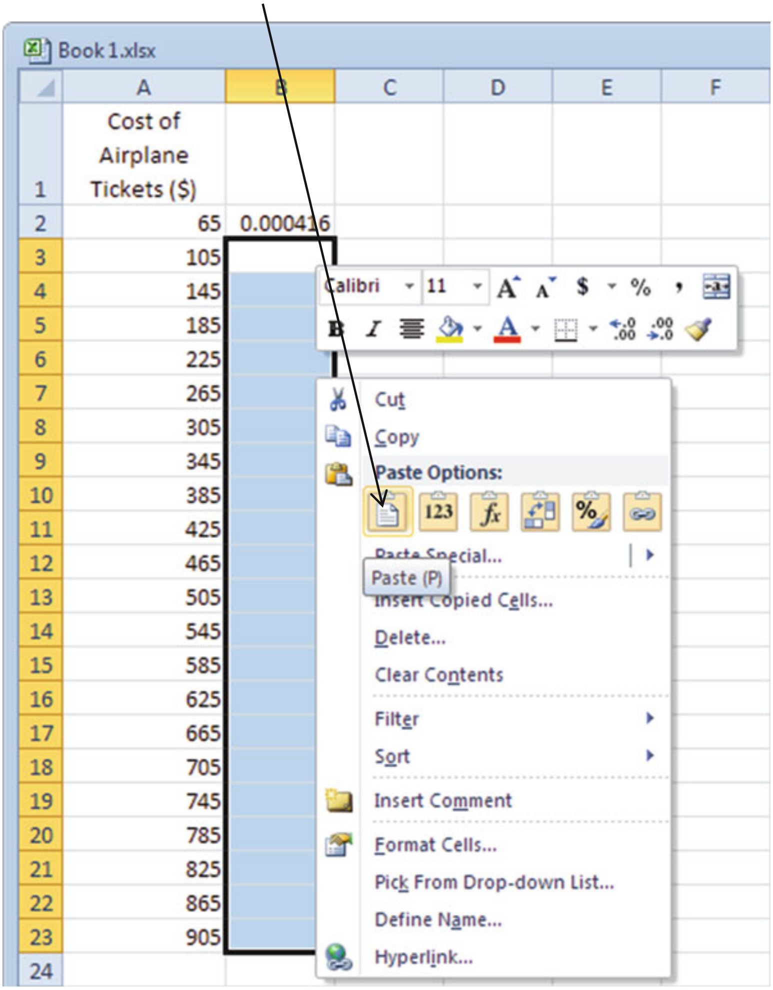
➢ Select the **Copy** function



Book 1.xlsx

	A	B	C	D	E	F
1	Cost of Airplane Tickets ($)					
2	65	0.00				
3	105					
4	145					
5	185					
6	225					
7	265					
8	305					
9	345					
10	385					
11	425					
12	465					
13	505					
14	545					
15	585					
16	625					
17	665					
18	705					
19	745					
20	785					
21	825					
22	865					
23	905					

➢ Highlight the empty cells in this same column across from the other ticket prices, right click, and select the **Paste** function



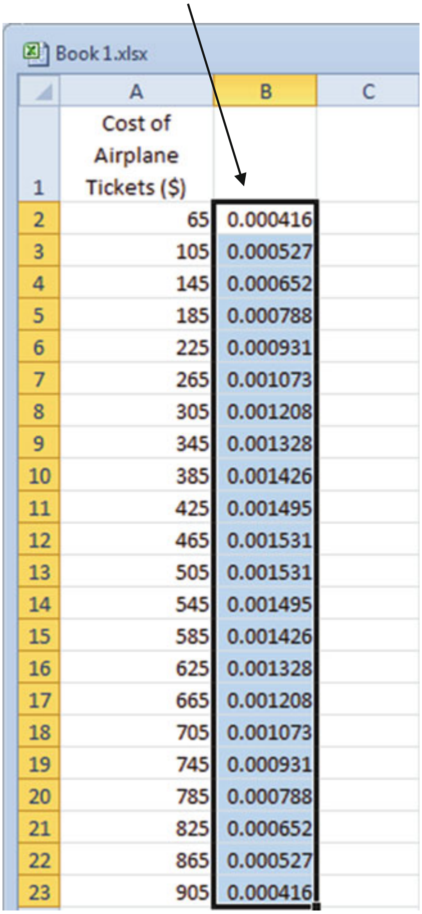
Book 1.xlsx

	A	B	C	D	E	F
1	Cost of Airplane Tickets ($)					
2	65	0.000416				
3	105					
4	145					
5	185					
6	225					
7	265					
8	305					
9	345					
10	385					
11	425					
12	465					
13	505					
14	545					
15	585					
16	625					
17	665					
18	705					
19	745					
20	785					
21	825					
22	865					
23	905					
24						

Step 3

Create the Graph.

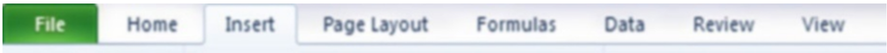
➢ Highlight the second column of values



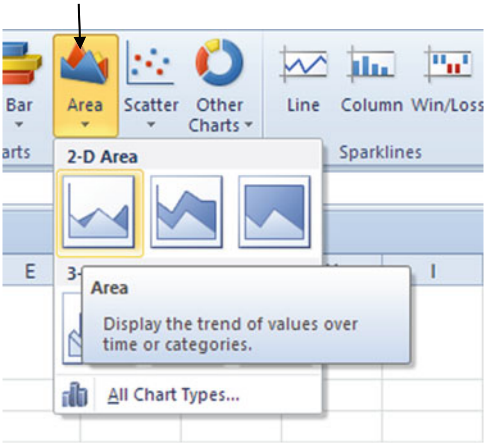
Book 1.xlsx

	A	B	C
1	Cost of Airplane Tickets ($)		
2	65	0.000416	
3	105	0.000527	
4	145	0.000652	
5	185	0.000788	
6	225	0.000931	
7	265	0.001073	
8	305	0.001208	
9	345	0.001328	
10	385	0.001426	
11	425	0.001495	
12	465	0.001531	
13	505	0.001531	
14	545	0.001495	
15	585	0.001426	
16	625	0.001328	
17	665	0.001208	
18	705	0.001073	
19	745	0.000931	
20	785	0.000788	
21	825	0.000652	
22	865	0.000527	
23	905	0.000416	

➢ Click the Insert tab

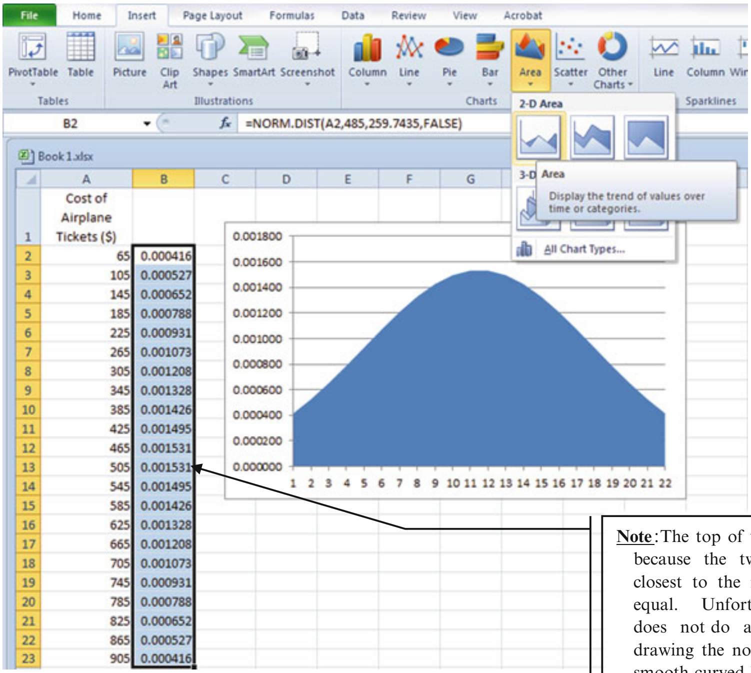


- ➢ Select the **Area** graph



Area

Display the trend of values over time or categories.



B2 f_x =NORM.DIST(A2,485,259.7435,FALSE)

Area

Display the trend of values over time or categories.

	A	B
1	Cost of Airplane Tickets ($)	
2	65	0.000416
3	105	0.000527
4	145	0.000652
5	185	0.000788
6	225	0.000931
7	265	0.001073
8	305	0.001208
9	345	0.001328
10	385	0.001426
11	425	0.001495
12	465	0.001531
13	505	0.001531
14	545	0.001495
15	585	0.001426
16	625	0.001328
17	665	0.001208
18	705	0.001073
19	745	0.000931
20	785	0.000788
21	825	0.000652
22	865	0.000527
23	905	0.000416

Note: The top of the graph is flat because the two data values closest to the mean (485) are equal. Unfortunately Excel does not do a great job of drawing the normal curve with smooth curved lines.

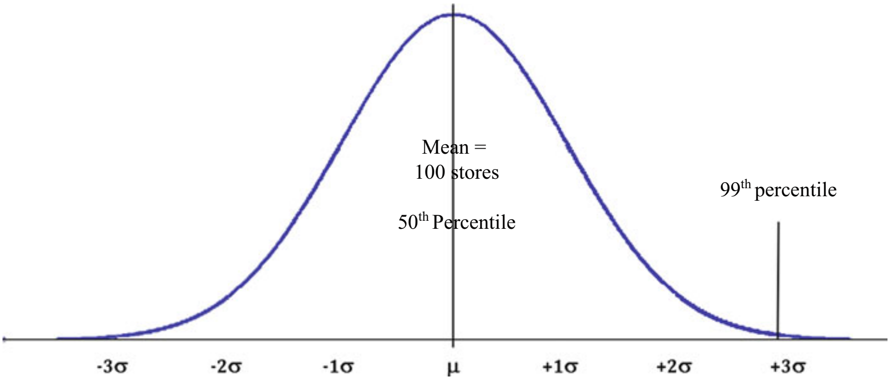
2. Converting Percentiles to Measured Units (NORM.INV)

Excel can covert percentile values in a normally distributed data set into measured units.

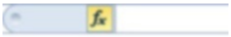
Example Problem

Your distribution company provides delivery service 7 days a week to stores selling synthetic fire logs. You need to determine how many trucks and drivers to have on hand to deliver the product. The data you have collected from last year indicates the number of stores requiring deliveries on a daily basis during the months of December and January. This data seems to follow a normal distribution. The average number of stores requiring deliveries on any given day during this time is 100. The standard deviation is 15 stores. Let's analyze the data.

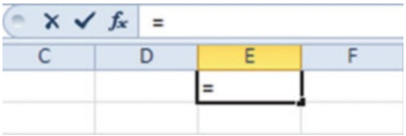
Calculate the Number of Stores Corresponding with the 99th Percentile



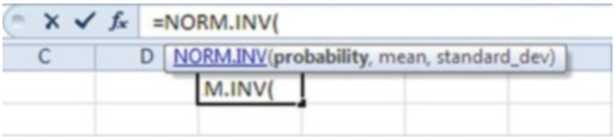
- ➢ Click on an empty cell where you want the answer
- ➢ Click on the ***fx*** function



- ➢ Type "=" in the white input box



- Type the function

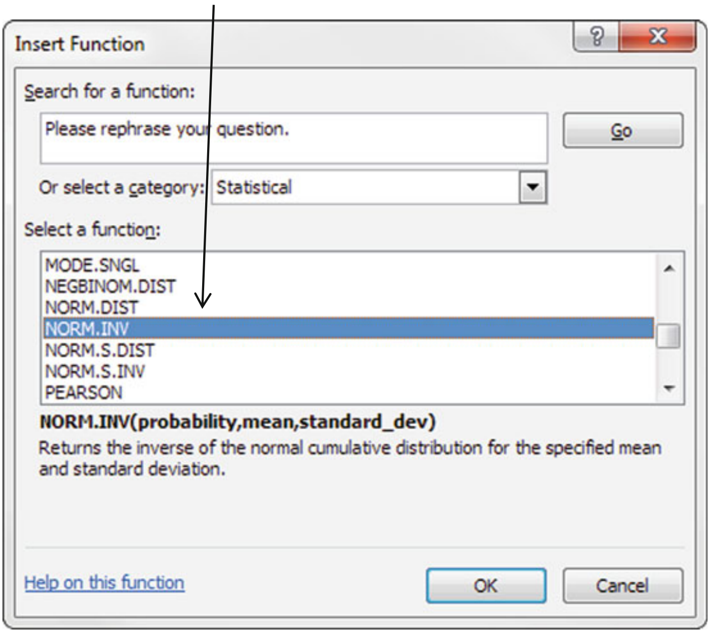


=NORM.INV(

NORM.INV(probability, mean, standard_dev)

M.INV(

OR highlight the function from Insert Function window



Insert Function

Search for a function: Please rephrase your question.

Or select a category: Statistical

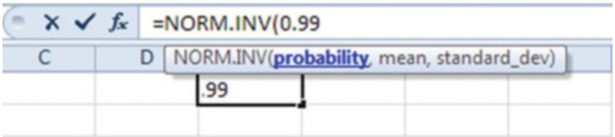
Select a function:

MODE.SNGL
NEGBINOM.DIST
NORM.DIST
NORM.INV
NORM.S.DIST
NORM.S.INV
PEARSON

NORM.INV(probability,mean,standard_dev)

Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation.

- Input the percentile value that you want to convert. **Input as a decimal value.** This is referred to as the probability by Excel.

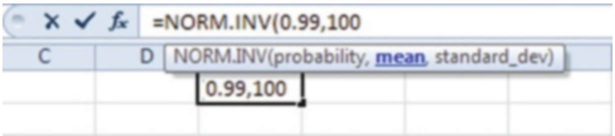


=NORM.INV(0.99

NORM.INV(**probability**, mean, standard_dev)

.99

➢ Input the **mean** from your data set

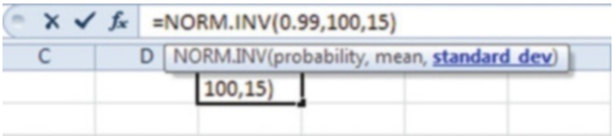


=NORM.INV(0.99,100

NORM.INV(probability, mean, standard_dev)

0.99,100

➢ Input the **standard deviation** from your data set

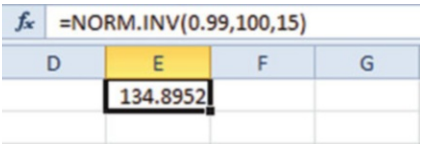


=NORM.INV(0.99,100,15)

NORM.INV(probability, mean, standard_dev)

100,15)

➢ Click the checkmark to the left of ***fx***

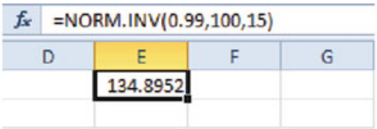


=NORM.INV(0.99,100,15)

134.8952

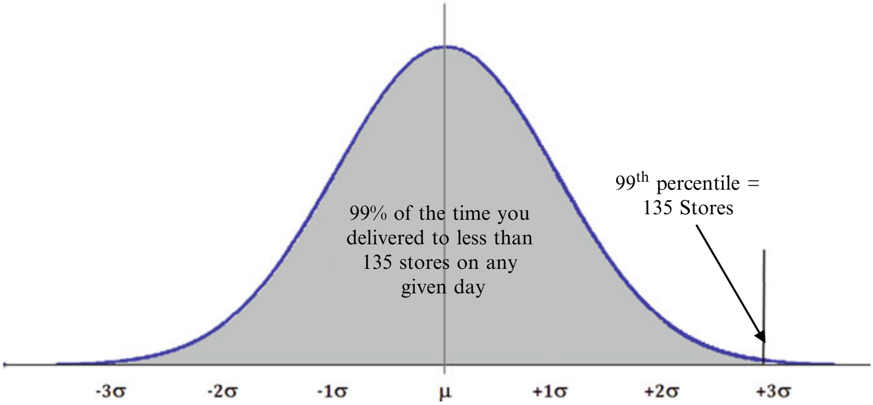
The answer is given as 134.8952 which means that the 99th percentile for this data set is about 135 stores. In other words, about 135 stores or less required deliveries on any given day 99 % of the time period in December and January of last year.

- No matter what the decimal indicates, we need to round up, because we cannot have part of a store.



fx =NORM.INV(0.99,100,15)

D	E	F	G
	134.8952		



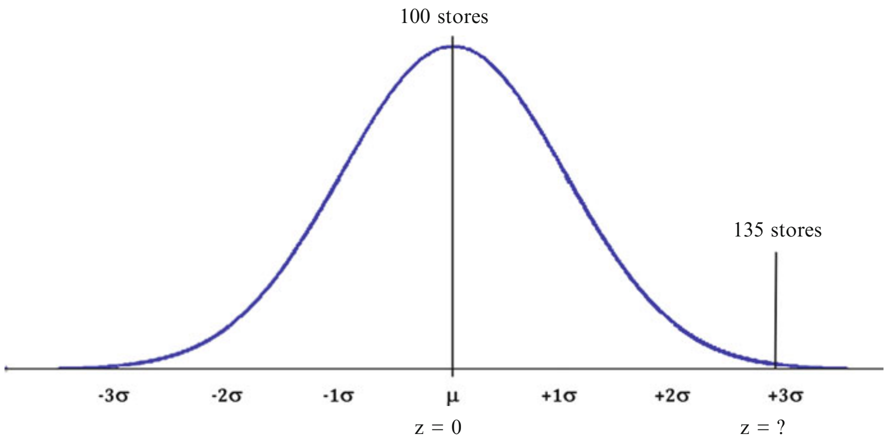
Mean = 100 stores

3. Converting Measured Units to z-Scores (STANDARDIZE)

Example Problem

Your distribution company provides delivery service 7 days a week to stores selling synthetic fire logs. You need to determine how many trucks and drivers to have on hand to deliver the product. The data you have collected from last year indicates the number of stores requiring deliveries on a daily basis during the months of December and January. This data seems to follow a normal distribution. The average number of stores requiring deliveries on any given day during this time is 100. The standard deviation is 15 stores.

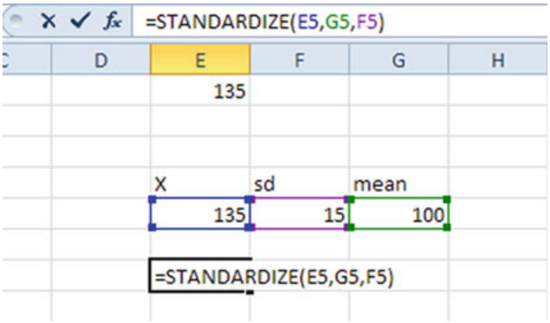
Convert the Measured Value of 135 Stores to a z-Score



- ➢ Input in the **mean** number of stores as **100** into cell G5
- ➢ Input the **standard deviation** as **15** stores into cell F5.
- ➢ Input the value that we want to convert of **135** stores into cell E5.
  - ◇ This problem could also be solved by directly typing in these required input values within the brackets of the function. The **STANDARDIZE** function uses the following format:

*STANDARDIZE(**observed value** in measured units to be converted to a z-score, **mean** of the data set, **standard deviation** of the data set)*

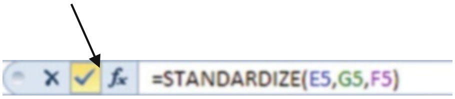
- ➢ Click on an empty cell where you want the answer
- ➢ Input "=" inside the ***fx*** formula box
- ➢ Type in the formula using cell locations for the observed value (x), mean and standard deviation. The formula divides the z-score by the sd and adds it back to the mean.



=STANDARDIZE(E5,G5,F5)

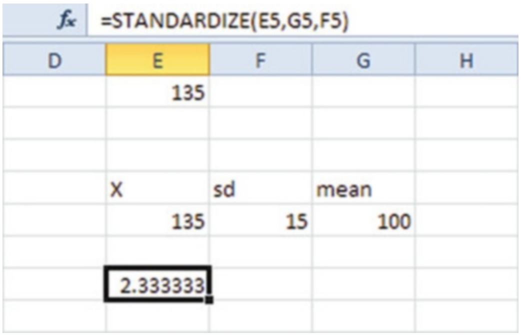
C	D	E	F	G	H
		135			
		X	sd	mean	
		135	15	100	
		=STANDARDIZE(E5,G5,F5)			

➢ Click on the checkmark



=STANDARDIZE(E5,G5,F5)

So in this case, 135 stores equates to a z- score of 2.326347 standard units. We should round the z-score to a value of approximately 2.33.



fx =STANDARDIZE(E5,G5,F5)

D	E	F	G	H
	135			
	X	sd	mean	
	135	15	100	
	2.333333			

4. Calculate Rank and Percentile (Rank and Percentile)

Example Problem

You would like to calculate the rank and percentile costs of airplane tickets purchased by employees over the past 3 months. The average cost is $485 and the standard deviation is $260. You know that the ticket prices are normally distributed based on historical data. Input the following ticket prices that are sorted from smallest to largest

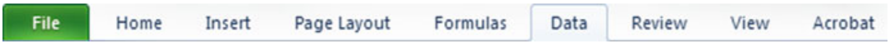
Cost of airplane tickets ($)
65
105
145
185
225
265
305
345
385
425
465
505
545

(continued)

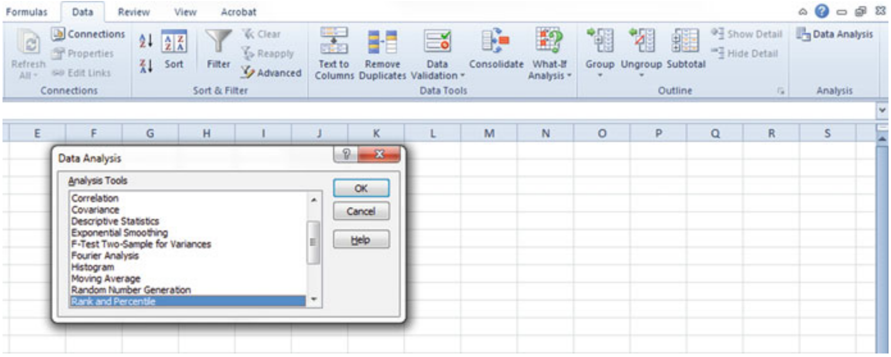
(continued)

Cost of airplane tickets ($)
585
625
665
705
745
785
825
865
905

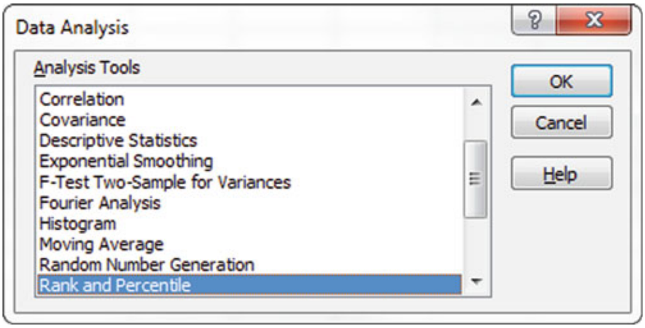
➢ Select the **Data** tab



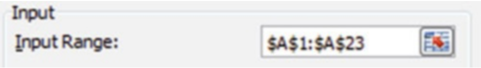
➢ Open the **Data Analysis** function



➢ Select the **Rank and Percentile** function



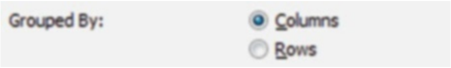
➢ Type in your column of x data



Input

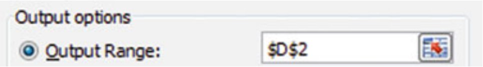
Input Range: $A$1:$A$23

➢ Select the how your data is grouped: **rows** or **columns**



Grouped By: ● Columns ○ Rows

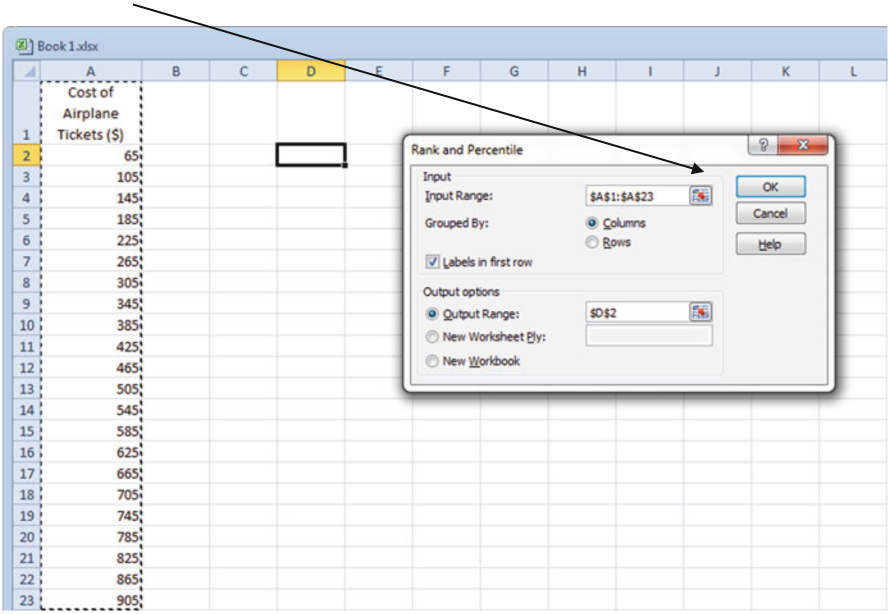
➢ Select an output location on your Excel spreadsheet



Output options

● Output Range: $D$2

➢ Click **OK**



Book1.xlsx

	A
1	Cost of Airplane Tickets ($)
2	65
3	105
4	145
5	185
6	225
7	265
8	305
9	345
10	385
11	425
12	465
13	505
14	545
15	585
16	625
17	665
18	705
19	745
20	785
21	825
22	865
23	905

Rank and Percentile

Input

Input Range: $A$1:$A$23

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: $D$2

○ New Worksheet Ply:

○ New Workbook

Output

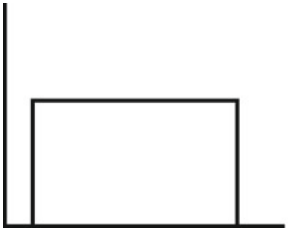
- The first column of data is just the rank order from highest to lowest.
- The second column includes the actual data value.
- The third column has the **Rank** from lowest to highest.

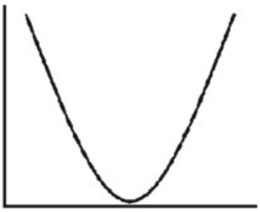
- The final column is the **Percentile** moving from the highest value (right hand side of the distribution) to the lowest x value (left hand side of the distribution) in the distribution.

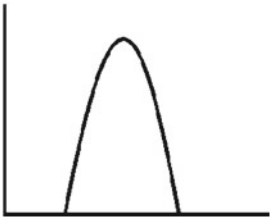
Point	Cost of airplane tickets ($)	Rank	Percent
22	905	1	100.00 %
21	865	2	95.20 %
20	825	3	90.40 %
19	785	4	85.70 %
18	745	5	80.90 %
17	705	6	76.10 %
16	665	7	71.40 %
15	625	8	66.60 %
14	585	9	61.90 %
13	545	10	57.10 %
12	505	11	52.30 %
11	465	12	47.60 %
10	425	13	42.80 %
9	385	14	38.00 %
8	345	15	33.30 %
7	305	16	28.50 %
6	265	17	23.80 %
5	225	18	19.00 %
4	185	19	14.20 %
3	145	20	9.50 %
2	105	21	4.70 %
1	65	22	0.00 %

From the output we can determine that 47.6 % of the prices were at the mean ($485) or less. Similarly we note that about 90 % are $825 or less. The 10th percentile is approximately $145. The ticket price of $705 is the sixth most expensive ticket. There are numerous analytical conclusions that can be drawn from this simple data output.

5. Non-normal Distributions


Uniform


Parabolic


Exponential

Not all distributions are normally distributed, or you simply may not know what the distribution looks like. To err on the side of conservative estimates, it is better to use the Chebyshev approximation in these cases. At least $(1-1/k^2)$ of the items in any data set will be within k standard deviations of the mean, where k is any value greater than 1.

- At least 75 % of the items must be within $k = \pm 2$ standard deviations of the mean.
- At least 89 % of the items must be within $k = \pm 3$ standard deviations of the mean.
- At least 94 % of the items must be within $k = \pm 4$ standard deviations of the mean.

There are two steps in this type of problem:

Step 1. Convert measured units to standard units (STANDARDIZE)
Step 2. Use Chebyshev approximation (*fx*)

Example Problem

The data set of apartment rents has a mean of $490.80 and a standard deviation of $54.75. The data is not normally distributed.

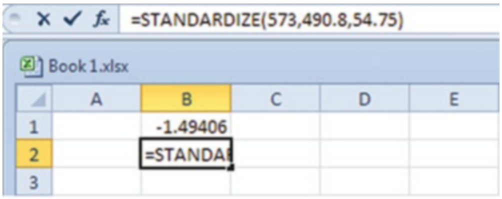
Calculate What Percentage of Rents Fall Between $409 and $573

➢ With this approximation, we can only calculate the minimum amount of data that would fall in the range of interest.

Step 1: Convert the Measured Values to Standard Units

For the first boundary value of 409

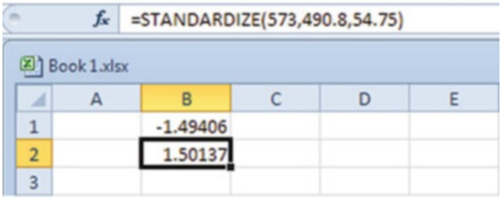
➢ Click on an empty cell where you want the answer
➢ Input "=" inside the ***fx*** formula box
➢ Type in the formula using cell locations for the observed value (x) of 409, mean and standard deviation



=STANDARDIZE(573,490.8,54.75)

Book1.xlsx

	A	B	C	D	E
1		-1.49406			
2		=STANDA			
3					



=STANDARDIZE(573,490.8,54.75)

Book1.xlsx

	A	B	C	D	E
1		-1.49406			
2		1.50137			
3					

The resulting standardized value for 409 is −1.49406 or approximately 1.5

For the second boundary value of 573

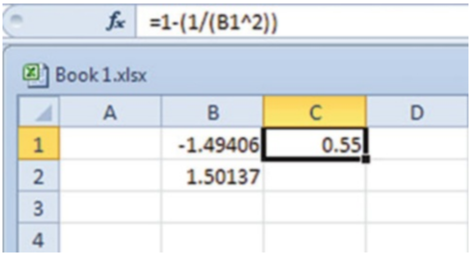
- Click on an empty cell where you want the answer
- Input "=" inside the ***fx*** formula box
- Type in the formula using cell locations for the observed value (x) of 573, mean and standard deviation

The resulting standardized value for 573 is 1.50137 or 1.5

Step 2: Use Chebyshev Approximation $= 1-(1/(k)^2)$ *where k is the boundary value in standard units*

- Click on an empty cell where you want the answer
- Input "=" inside the ***fx*** formula box
- Type in the formula using cell locations for the observed value (x) in standard units

$$\text{In this example the formula} = \mathbf{1} - \left(\mathbf{1}/(\mathbf{k})^{\mathbf{2}}\right) \text{ is input to Excel as}$$
$$= 1 - (1/((-1.5)^{\wedge 2}))$$



fx =1-(1/(B1^2))

Book1.xlsx

	A	B	C	D
1		-1.49406	0.55	
2		1.50137		
3				
4				

Answer: At least 55 % of the rents fall between $409 and $573.

Note in this case the standard units were equal on either side of zero. For non-symmetric problems where the two z scores are different distances away from the mean you can perform a series of subtractions just as you would if the distribution was normal. For example if the lower z score was +1 and the upper z score was +2, calculate how much data falls between ± 1 using the Chebyshev formula and subtract that amount from what falls between ± 2. Then divide by 2 so you have just the part greater than the mean. But this approximation only indicates the minimum amount of data that exists in this region.

Common Pitfalls

- Many of the Excel functions only differ by one letter, so make sure you have the right normal distribution function.
- Remember, Excel gives percentages as decimals. Multiply by 100 to get a whole number percentage.
- Some versions of Excel may have a period in the function name so make sure you include this to get the right function.
- Chebyshev approximation does not work with small standard deviations (e.g., the sd must be greater than 1).
- Remember the Chebyshev approximation only calculates the minimum amount of data (at least. . .).
- NORM.INV does not provide the percent of distributions nor frequency on the y-axis. It is used for graphing purposes only.

Final Thoughts and Activities

Practice Problems

1. The data in the file **Amusementparks** contain the starting admission price (in $) for one-day passes to 15 amusement parks in the United States.

   (a) Compute the mean, range, variance, standard deviation.
   (b) Compute the z-scores.
   (c) Based on the results from (a) and (b), what conclusions can you reach concerning the starting admission price for one-day passes.

2. The data in the file **Candy** represent the cost per pound ($) for a sample of candy.

   (a) Compute the mean, range, standard deviation.
   (b) Compute the z-scores.
   (c) Is the data skewed? If so, how?

3. You are an operations analyst for AT&T. The length of long distance telephone calls may be normally distributed but you are not certain. You select a random sample of 25 calls. The average length is 8 min and the SD is 2 min. What % of these calls should be between 4.8 and 11.2 min?

Discussion Boards

1. IQ scores are normally distributed with a mean of 100 and a standard deviation of 15. MENSA is an organization for people with high IQ's; eligibility requires

an IQ about 131.5. New York Times reporter Trish Hall notes that people are scoring substantially higher on IQ tests than in the past. Discuss.

2. Discuss some of the cultural and political factors that could affect a normal distribution of voter ages in government election results in different countries around the world.

Group Activity

1. Queuing Theory is a branch of statistics that is very important to businesses such as McDonald's, Safeway and Disney. Disney conducts extensive studies of lines at its amusement parks to keep patrons happy and to assist with expansion plans. Discuss how all of this relates to the normal distribution.
2. Ergonomics is the study of problems associated with people adjusting to their environments. A case in point was faced by the US Air Force in studying the cockpits of fighter jets. These were originally designed for men but of course now we also have women pilots. Research information on the web in conjunction with a normal distribution approach to address this issue.

Parting Thought

Smoking is one of the leading causes of statistics.

Problem Solutions

1. The data in the file **Amusementparks** contain the starting admission price (in $) for one-day passes to 15 amusement parks in the United States.

   (a) Compute the mean, median, range, variance, standard deviation.

Admission	
Mean	49.80
Median	45.50
Standard deviation	11.10
Sample variance	123.29
Range	34.00
Count	10.00

(b) Compute the z-scores.

Location	Admission ($)	z-score
Boo's Gardens	61	1.01
Animal Adventures	66	1.46
Dolly's Wonderland	44	−0.52
Chocolate Heaven Park	45	−0.43
Bob's Children's Park	32	−1.60
Mouse's Great Adventure	53	0.29
Fairy Lake Park	65	1.37
Silver Lining Park	46	−0.34
Adventure City Rides	43	−0.61
Mountain Explorer's Park	43	−0.61

(c) Based on the results from (a) and (b), what conclusions can you reach concerning the starting admission price for one-day passes?

The admission price for a one-day pass is slightly skewed to the right because the mean is slightly greater than the median.

2. The data in the file **Candy** represent the cost per pound of ($) for a sample of candy.

(a) Compute the mean, median, range, standard deviation.

Cost($)	
Mean	1.68
Median	1.63
Standard Deviation	0.33
Sample Variance	0.11
Range	0.96
Count	14.00

(b) Compute the z-scores. Are there any outliers? Explain

Cost($)	z-score
1.43	−0.75
1.47	−0.63
1.67	−0.02
1.89	0.65
2.17	1.51
1.69	0.04
1.52	−0.48
1.32	−1.09
2.26	1.79
1.32	−1.09
1.30	−1.15
1.61	−0.20
2.16	1.48
1.65	−0.08

There are no outliers because none of the Z-Scores has an absolute value that is greater than 3.0.

(c) Are the data skewed? If so, how?

The data appears to be skewed to the right because the mean is greater than the median.

3. Given: avg = 8 min sd = 2 min

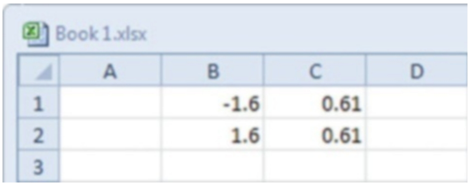
- Convert 11.2 and 4.8 min to std units

  Enter "=STANDARDIZE(11.2,8,2)" in a cell
  Answer: 1.6

  And enter "=STANDARDIZE(4.8,8,2)" into another cell
  Answer: −1.6

- If k = 1.6 AT LEAST $\{1-(1/k^2)\}$ will be within ± 1.6 standard deviations of the mean

  Enter "=1−(1/(1.6^2))" in one cell
  =1−(1/2.56) = .61 or AT LEAST 61 % of the calls will be between 4.8 and 11.2 min



	A	B	C	D
1		-1.6	0.61	
2		1.6	0.61	
3				

Chapter 5
Survey Design

Case Study: Sampling for Potential Growth Areas

Infinity Auto Insurance is looking to expand their customer base starting with large metropolitan areas across the United States. Because the Latino/Hispanic population is the largest growing population in the United States, Infinity Auto Insurance wishes to better understand this market segment and expand its advertising campaigns to include this population. They have decided to concentrate their strategic efforts to large metropolitan areas in four different states: California, Arizona, Texas, and Florida. For the initial study, they sampled 200 people in Los Angeles, California.

Food for Thought

What other states/cities should they target?
What other demographic segments should they consider?
From what you understand of sampling, what sample size should they use to gather a statistically valid sample?

Possible Answers

What other states/cities should they target?
Houston, Texas; Phoenix Arizona; Miami, Florida; Austin, Texas
What other demographic segments should they consider?
Lower income or less educated individuals

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_5,

Over the years, Infinity Auto Insurance has increased their sample size from 200 to 400, or 600 depending on the depth and detail they wish to obtain from the sample. With a sample size of 400, Infinity Auto Insurance is able to statistically calculate valid information at a 95 % level of confidence (±5 points); with 600, 95 % level of confidence (±4 points).

Food for Thought

Would you suggest they gather a larger sample or is their current sample sufficient?
If you were working for Infinity Auto Insurance, how would you go about gathering a valid sample?
What issues would help or hinder you from creating a statistically valid sample?

Possible Answers

Would you suggest they gather a larger sample or is their current sample sufficient?
A sample size of 400 or 600 should be sufficient to capture the necessary data.

If you were working for Infinity Auto Insurance, what types how would you go about gathering a valid sample?
Use a random sample by gathering a database of randomly selected individuals. Contact the individuals in the database randomly to take the survey.

What issues would help or hinder you from creating a statistically valid sample?
Non response, non-random sampling, etc.

INFINITY DriverClub®

It isn't often that a company puts your needs above theirs. But at Infinity, we believe that no one should be stranded on the side of the road. Our core values focus on providing protection for you and your family from everyday road hazards, even if you are not an Infinity insured. Join Infinity DriverClub® and receive access to 24/7 Nationwide Roadside Assistance. You are covered:

- Any time, in any car
- No membership fees
- Pay only for the services you use
- Coverage available everywhere in the USA

Signing up is easy! Call 800-516-6894 or sign up at InfinityDriverClub.com

Key Concepts

Bias, Bipolar scale, Extreme checking, Likert scale, Nonprobability sampling, Primary data, Probability sampling, Random sampling, Random sampling error, Response rates, Reversed items, Secondary data and Systematic error.

Discussion

This chapter provides a framework for effective survey design. All statistical analyses require high quality input data. The first step in collecting unbiased, reliable and valid survey data is having an unbiased, reliable and valid survey instrument. Several effective examples are provided as well as some of the more common pitfalls to avoid.

Basic Concepts

Primary data is collected by the investigator conducting the research. **Secondary data** is collected by someone other than the investigator. The most common way to generate primary data is through surveys. Surveys provide quick, relatively inexpensive, and efficient means of assessing information about a population. The tasks of determining the survey distribution method, determining the list of questions, and designing the correct format of the survey are essential aspects of the development of a reliable and robust survey.

Most surveys are designed to collect information from a sample of the population. Sampling is the use of a subset of the population to represent the whole population. **Probability sampling**, or **random sampling**, is a sampling technique in which the probability of getting any particular respondent is equally as likely to occur. **Nonprobability sampling** does not meet this criterion and should be used with caution. Regardless of how the data is collected, it is important to minimize errors in the data collection process. This topic is covered in more detail in Chapter 6.

- When a selection procedure is biased, taking a larger sample does not help. This just creates the bias on a larger scale.
- To minimize bias, an impartial and objective probability method should be used to select the sample.

Although most surveys are conducted to quantify factual information, certain aspects of surveys may be qualitative. Although most surveys are descriptive, they can also be designed to provide causal explanations.

Survey Design

The survey designing process is a very important step to get your future survey produced. The first step is to start planning the purpose of the survey. Start by listing the survey objectives. As you start drafting questions, make sure that your survey items correspond to an initial objective listed. A common mistake is that researchers do not get the information needed because they do not ask the right questions in the survey. Finally, keep in mind that demographic questions can be critical to analyze results.

Scale

The **Likert scale** is one of the most common scales you will come across in surveys. This scale was introduced by Likert (England; 1932). Likert scaling is a uni-dimensional scaling method or "summative" scale.

One of the first questions in designing a survey is how many choices should be offered on the Likert scale. The most common are 5 or 7-point Disagree-Agree response scales. Examples are given in Fig. 5.1.

There are tradeoffs between choosing a 5-point scale and a 7-point scale. The 7-point scale provides more choice for the respondent and a more detailed response set for the analyst. However, more choices may lengthen the time to complete the survey, as respondents try to decide on their answer. Too detailed responses may require more time to complete: "Was I really somewhat satisfied or just satisfied the last time I responded to this survey?"

1. = Strongly unfavorable to the concept
2. = Somewhat unfavorable to the concept
3. = Neutral
4. = Somewhat favorable to the concept
5. = Strongly favorable to the concept

or

1. = Extremely unfavorable to the concept
2. = Strongly unfavorable to the concept
3. = Somewhat unfavorable to the concept
4. = Neutral
5. = Somewhat favorable to the concept
6. = Strongly favorable to the concept
7. = Extremely favorable to the concept

or

1. Agree
2. Neutral
3. Disagree

Fig. 5.1 Likert scales

However, having too few choices, less than 5, can be frustrating to the respondents; none of the answers really fit. This smaller scale also limits a more detailed understanding of the responses. A 3-point scale is also given in Fig. 5.1.

The final score for the respondent on the scale is the sum of their ratings for all of the items (this is why it is sometimes called a "summated" scale). On some scales, you will have items that are reversed in meaning from the overall direction of the scale. These are called **reversed items**. You will need to reverse the response value for each of these items before summing for the total.

Bipolar questions offer positive, neutral and negative choices. Whenever you are dealing with bipolar answers, it is important to have an odd number of choices no matter how large the scale. This becomes even more important when using a Likert scale with qualitative data. In this situation, the respondent is forced to decide whether they lean more towards the agree or disagree end of the scale for each item. Figure 5.2 illustrates bipolar and unidirectional Likert scales.

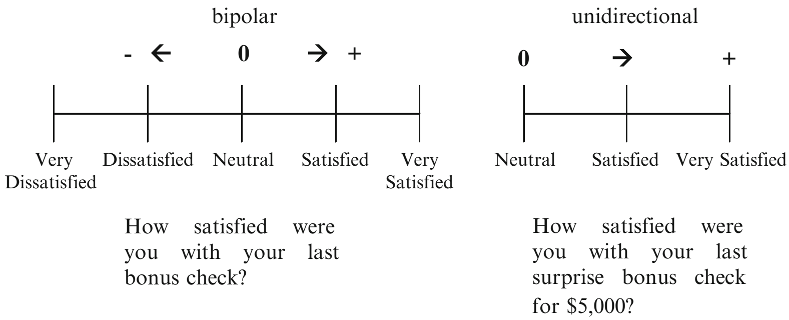


Fig. 5.2 Polarity with Likert scales

In Fig. 5.2 the first bipolar scale is designed to allow respondents to agree, disagree or remain neutral about their level of satisfaction with the recent bonus check. The second unidirectional scale is designed to only capture positive responses. In other words, it is safe to say everyone would be positive to some degree to get a surprise bonus check for $5,000. These scales are easily converted to quantitative scales if they are designed correctly, which allows for more rigorous analysis.

If the bipolar scale had only been a 4-point scale we would not have been able to collect neutral responses. However, often less rigorous surveyors will simply average the responses to synthetically create a neutral response, as shown in Fig. 5.3. This is inaccurate data and should be avoided.

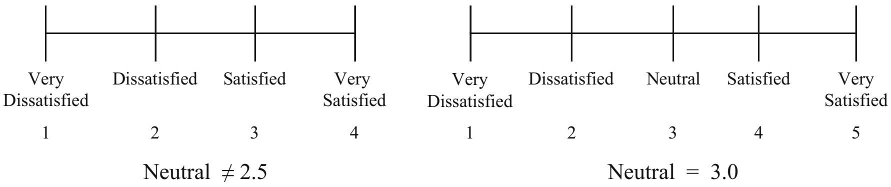


Fig. 5.3 Offering a neutral choice

Types of Questions

Single Response/Select

Single select items are questions that ask the respondent to select one specific response to fulfill the question requirement. For example most demographic questions are single response questions. The question asks for a specific characteristic of the person's demographic status (i.e. age, gender, household income, etc.), and the respondent is allowed to respond with one specific selection.

Example

1. What is your age?

○ 18–24
○ 25–34
○ 35–44
○ 45–54
○ 55–64
○ 65 or over

Because the respondent is a specific age, he/she will respond with only one item.

◈ For range items, such as age, make sure the numbers do not overlap and that the entire range is covered. Otherwise, this might cause confusion and inaccurate responses (i.e. age 18–25, age 25–30, age 30–40, etc.)

Multiple Response/Select

Multiple select questions are ones that ask the respondent to "select all that apply." A respondent can choose one or more responses. Sometimes the option of "*None of the above*" is available when none of the other responses seem to fit.

Example

1. What are your favorite colors?

□ Red
□ Blue
□ Green
□ Yellow
□ Orange
□ None of the above

When a respondent chooses the different responses, he/she is allowed to choose as many responses as they deem fit. However, if the respondent chooses the "*None of the above*" option, it should deselect any previous answer selections for this question.

Structured Questions

Structured questions are questions that have a closed set of responses from which to choose. Structured questions are usually preferred because they make data collection and analysis much easier and they are quicker for the respondent to answer. Structured questions are recommended when you are not trying to capture new ideas or thoughts from the respondent. You need a good understanding of the responses so that you can develop the appropriate answer choices.

Example

Do you have an MBA?
○ Yes
○ No
○ In Progress

Which international city where we have offices would you like to accept an assignment?
○ New York
○ Singapore
○ Dubai
○ Hong Kong
○ Frankfurt
○ Stonetown

For structured questions make sure that the list of answers includes every possible answer and that each of the answers is unique.

To ensure that the data you are collecting is indeed complete, you might want to include *"Other"* as the last answer choice. Including the answer choice "*Don't know*" to a response list provides a valid choice for respondents who are not capable of answering the question.

Do not include answer choices that are irrelevant to the question. You want to keep the survey as short and as easy to complete as possible.

Example of an Irrelevant Answer Choice

To which city should we relocate the head office?

○ San Francisco
○ Atlanta
○ Jacksonville
○ Detroit
○ Disneyland

The last answer choice is not a real city and may be a very popular answer but is certainly not going to add much to your analysis. Although we might think it good to include humorous answers every once and awhile, keep in mind that when someone actually chooses these, you lose some real data.

All of the responses should be consistent so that no single response stands out from the other responses. Consistency ensures that you are not leading respondents to a particular answer by making that answer different from the others.

Example of an Inconsistent Answer Choice

To which city should we relocate the head office?

○ San Francisco
○ Atlanta
○ Jacksonville
○ Detroit
○ Alaska

The last answer choice is not a city but a state. If we really did not want the respondent to choose a city in Alaska this would be the way to ensure that. This confuses the respondent and distracts them from completing the survey in a valid manner.

Ranking and Rating

Surveys often gather opinion information, by way of asking a person's opinion on a topic. Rating or ranking questions capture varying degrees of emotion about a subject. A rating question asks respondents to indicate the degree to which they feel about a certain topic. These are often referred to as Likert scales.

For Example

How do you feel about the new flexible spending account?

Unsatisfied	Somewhat Satisfied	Satisfied	Very Satisfied	Extremely Satisfied
(1)	(2)	(3)	(4)	(5)

A **ranking** question asks respondents to explain how they feel about something by comparing it to other items in a list.

Example

Please rank the following healthcare benefits (1 being highest rank and 4 being the lowest)
___Flexible healthcare savings account
___Employee assistance program
___Long term insurance for retiree healthcare
___Exercise program subsidy

Non-structured (Open-Ended) Questions

Non-structured questions, or open-ended questions, are questions that have no list of answer choices from which to choose. Respondents write in their response to a question.

Example

What do you like best about our new CEO?

____

____

Use non-structured questions when you are exploring new ideas and you don't really know what to expect from the respondents. This format gives them an opportunity to provide you with information that you may not have considered before. However, keep in mind that these types of questions take much longer to analyze.

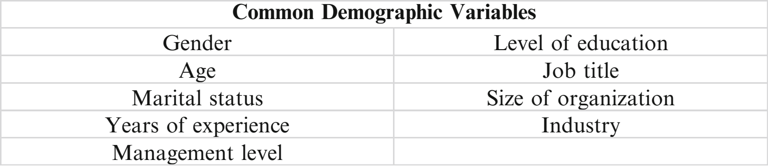
Data

Labels

Another issue involves thinking you have transformed qualitative data into quantitative data because the labels have become numeric. By translating the label "satisfied" to a "3" does not make that data quantitative. Recall from our earlier chapters on descriptive statistics, we are much more limited in what type of analysis we can complete with qualitative data, regardless of whether the labels are numeric or not.

Demographic Data

To make sure the survey sample is representative of a wider population we need to collect demographic information. Demographic items are typically categorical data. Some common demographic variables are presented in Fig. 5.4.



Common Demographic Variables	
Gender	Level of education
Age	Job title
Marital status	Size of organization
Years of experience	Industry
Management level	

Fig. 5.4 Common demographic variables

In addition to those variables listed in Fig. 5.4 and depending on the purpose of the survey, you may also want to collect demographic data on national culture and religion; therefore, make sure you have included all the necessary demographic variables.

Response Rates

It is important to achieve a high response rate; otherwise, we may have non-response bias. In other words, there might be some reason we would suspect that so many people did not respond. For example, if the survey was 15 pages long, people might not complete the survey because of how much time would be required. Those who do reply may also represent a bias. Perhaps, this group has more time on their hands, they are personal friends with the researcher, or they worked for the manager who initiated the survey. This type of bias weakens the usefulness of the data. Figure 5.5 summarizes some key design elements that may improve response rates.

1. Keep the survey as brief as possible.
2. Ensure all questions are clear and easy to understand. Avoid ambiguity ;it creates respondent frustration.
3. Maintain a consistent scale on short surveys; mix it up on longer surveys to keep it interesting for respondents.
4. Provide confidentiality and anonymity.
5. Don’t ask the same question in different ways more than three times.
6. Explain what the survey results will help provide.

Fig. 5.5 Response rate factors

Of course, the most influential factor is the subject matter of the survey. If you ask employees to complete a survey on their pay and benefits, there is usually a high response rate, because it is of great interest and valence to employees. However, if you ask employees to complete a survey about a new ad campaign, there may be a very low response rate, because they have been asked to do this every month for the past 6 months.

Editing: Data Quality

Before you start generating statistics from your data, you need to have a quick look at the raw data. Generate an Excel output of the raw data to check for unusual patterns in the data. Figure 5.6 gives you some examples.

		Questions						
Respondent		**1**	**2**	**3**	**4**	**5**	**6**	**7**
1		5	1	5		1	1	5
2		3	2	4		3	4	6
3		3	3	3		5	2	3
4		1	1	1		1	1	1
5		3	1	2		5	5	5
6		3	2	4		3	1	5
7		3	1	5		2	5	2
8		3	3	1		5	2	3

Fig. 5.6 Patterns in response

In Fig. 5.6, there are 7 questions and 8 respondents. There are several patterns in these responses that may concern us.

- Respondent 1 has only checked 1 and 5 on a 5-point Likert scale. This type of response pattern is referred to as **extreme checking** and creates a bias in the data. Some respondents only really agree with something or really disagree; they just cannot take a neutral position.
- Question 1 seems to have an overwhelming "3" or neutral response on the 5-point Likert scale. This may mean the question was worded in a confusing way or there really was no way to disagree or agree with the question. Consider the question "*Over half of our employees were "satisfied" in an earlier version of the survey and this is important. Do you agree or disagree?"* You may not have heard about an earlier version of the survey, let alone the survey results, so how can you agree or disagree. You don't know if the first part of the question is true or not. A remedy in this case would be to insert the word "*If*" at the start of the question.
- Question 4 has a large number of non-responses. Again a source of potential bias.
- Respondent 2 shows a "6" as the response to question 7. This needs to be checked, since the possible answers change only from 1 to 5.
- Respondent 4 seems to have only selected "1" as a response to all questions. This respondent may not have bothered to read any of the questions. If you asked some of the questions with **reversed scales** you can check if the answers are consistent. Consider the example.

How do you feel about adding a new cafeteria?

[✓ 1]	2	3	4	5
Strongly Disagree		Neutral		Strongly Agree

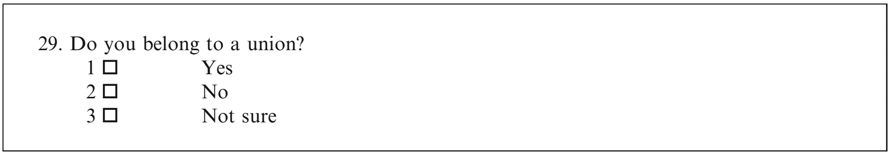
How do you feel about not adding a new cafeteria?

1	2	3	4	[✓ 5]
Strongly Disagree		Neutral		Strongly Agree

In this example, someone who strongly disagrees in the first question should strongly agree with not adding the cafeteria in the reversed question.

Coding

Coding is the process of identifying and classifying each answer with a numerical score or other character symbol. Before data can be tabulated, meaningful categories and character symbols must be established for groups of responses. In Fig. 5.7 the answers are coded as **1 for Yes**, **2 for No** and **3 for Not sure**.

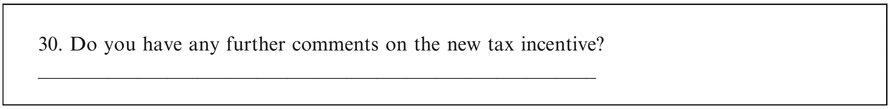


29. Do you belong to a union?

1 □ Yes
2 □ No
3 □ Not sure

Fig. 5.7 Basic coding of answers

Open ended questions pose more work in terms of coding as shown in Fig. 5.8. You need to complete a qualitative word categorization. There may be many comments concerning "*inequality*" or "*difficult to understand*." These categories can then be coded accordingly.

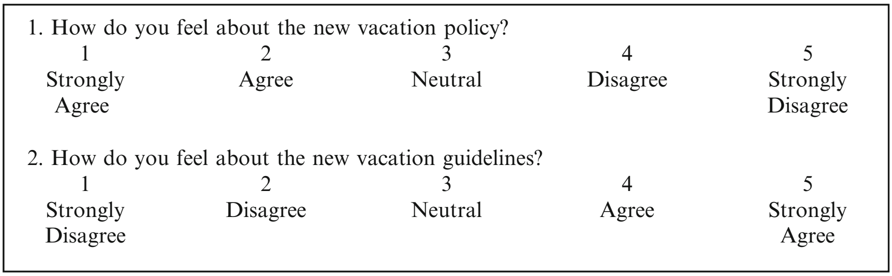


30. Do you have any further comments on the new tax incentive?

____

Fig. 5.8 Coding open-ended questions

Be careful with **reversed scale** questions. As shown in Fig. 5.9, the answers for question 2 need to be reversed so the high agreement score still gets coded as a 1, rather than the 5.



1. How do you feel about the new vacation policy?

1	2	3	4	5
Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree

2. How do you feel about the new vacation guidelines?

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Fig. 5.9 Reversed scale questions without transformation

Reversing the order of the codes for negative statements requires a simple data transformation, so the codes reflect the same direction and order of magnitude as the positive statements. Figure 5.10 shows the same questions with question 2 recoded appropriately

1. How do you feel about the new vacation policy?

1	2	3	4	5
Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree

2. How do you feel about the new vacation guidelines?

5	4	3	2	1
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Fig. 5.10 Reversed scale questions with transformation

The 0/1 coding is sometimes referred to as dummy variables or indicator variables. One of the most common dummy variables is gender, where we assign a 1 for female and a 0 for male. In some situations, we may have more than 2 categories such as the quarters in the fiscal year or the seasons. In these cases, we use a number of indicator variables always starting with the number "0" as the first category.

Errors in Survey Question Creation

Loaded Questions

This type of question usually produces a socially desirable response or contains emotionally charged words. For example, "*Are you a cheater*?" These types of questions make assumptions about the respondent and tend to evoke certain responses from them. To avoid asking leading questions, use agree/disagree questions that are phrased with words that have neutral connotations and do not assume specific details about the respondent.

Leading Questions

This type of item phrases the question in a way that suggests to the respondent that he/she is supposed to answer in a specific manner (i.e., it leads the respondent to a particular answer). For example, "*Do you agree with Doctors world-wide that say* *obesity has become a chronic pandemic?*" To avoid leading questions, remove the leading words or phrases (e.g., the words that make a judgment or imply stupidity if the respondent disagrees). In this example, "*Doctors world-wide*" is leading the respondent to agree with "experts."

To avoid leading questions, phrase all questions in a neutral way. It is a common error for surveyors to word the questions in a manner that reflects their underlying opinion.

Example

Bad question: leading	Good question: neutral
Do you think that the new CEO who was hired from a completely different industry can effectively lead our company?	Do you think our new CEO can effectively lead our company?
○ Yes	○ Yes
○ No	○ No
○ No opinion	○ No opinion

The leading question hints that since the new CEO is from another industry this may influence his/her ability to lead. The neutral question presents a better way to phrase this question, and removes the bias.

Double-Barreled Questions

This type of question asks about more than one idea or issue in one question, which may lead to inaccuracies in the attitudes and opinions being measured. For example, "do you agree or disagree that ice cream and cake are great?" The respondent may like ice cream but not cake, vice versa, both, or none. To avoid double-barreled questions ask each idea separately.

⟐ Make sure you are only asking one question at a time, this is different from multiple select type questions. These types of questions do not provide you with valid and reliable data.

Example

Bad question: double-barreled question	Good question
How have scientists and administrative staff at your research institute responded to the new parking policy? ○ Satisfied ○ Unsatisfied	How have scientists at your research institute responded to the new parking policy? ○ Satisfied ○ Unsatisfied How have administrative staff at your research institute responded to the new parking policy? ○ Satisfied ○ Unsatisfied

The double barreled question is phrased in such an ambiguous way, that you will not know what the respondent intended with their response (e.g. are the scientists, the administrative staff, or both satisfied/dissatisfied?). The key word (which differentiates the two questions) should be underlined to help ensure that the respondents are reading the questions correctly when the structures are so similar.

Errors in Survey Data Collection

Random Sampling Error

Even with probability sampling methods, statistical errors will occur because of chance error. **Random sampling errors** can be estimated, and sample size can be increased to decrease the effect of this type of error in the data.

Systematic Error

This type of error results from some imperfect aspect of the survey or from a mistake in executing the survey. The two most common forms of **systematic error** are referred to as respondent error and nonresponse error. Nonresponse error occurs when certain groups are no longer represented in your survey because they did not respond. For example, if you were conducting a door-to-door poll in a neighborhood where typically both people work during the day, your sample will only include respondents such as retirees or people on sick leave or vacation.

If your nonresponse is high, you need to rethink your survey questions, format, and/or delivery method. Self-selection bias is another form of error in which allows extreme position to be overrepresented while other differing positions are underrepresented.

Response Bias

There are six specific categories of response bias: acquiescence, auspices, central tendency, extremity, interviewer, and social desirability. Take note that these biases are not mutually exclusive, so you can have several of these biases operating within the same survey. Surveyors should also be alert for culturally based response biases in international business research. One might expect more acquiescence bias within Japanese responses, as they have a more culturally homogeneous society and tend not to contradict their superiors.

- **Acquiescence Bias.** This is a response bias due to the respondents' tendency to agree with questions when having doubt. Some respondents are very agreeable whereas others may be very disagreeable throughout the survey. We typically find positive acquiescence when asking about a new product.
- **Auspices Bias.** Answers may be influenced by the actual group doing the survey. If EarthWatch was conducting a survey about open pit mining the respondent may be inclined to provide answers that align with the mission of the organization.
- **Central Tendency Bias.** This response bias usually occurs with respondents who wish to remain neutral or have no specific opinion. Often, participants who respond this way want to get through the survey quickly and don't care about the subject matter of the survey.
- **Extremity Bias.** Some respondents will indicate extreme answers either in a positive and/or negative manner. These are the folks who only check the extreme ends of the scale, which may cause a bias in the data.
- **Interviewer Bias.** There may be a bias introduced because of the interactions between the interviewer and the interviewee. Often socially acceptable answers are given to impress the interviewer. The physical appearance of the interviewer may also affect the respondents. If the interviewer smiles after a particular answer is given, the respondent may tend to provide similar answers for further positive feedback. The gender of the interviewer may hamper candid responses on gender specific issues. If the interviewer takes too long to ask the question the respondent may feel like their time is being wasted and will start answering abruptly.
- **Social Desirability Bias.** The respondent may either consciously or unconsciously provide an answer to create a favorable impression. For example in asking if someone voted in the last election, you might create a social desirability bias, as most respondents want to be considered as responsible citizens and will report they voted. A more accurate question might be to ask if they know someone that did vote in the last election.

Checklist

✓ **Clearly state your intentions with the research.**

- At the top of your survey, write a brief statement explaining why you are collecting the information.

✓ **Reassure privacy of the data.**

- Reassure each respondent that the information is entirely anonymous and confidential.

✓ **Include instructions with your survey**

- What may seem obvious may not be obvious to respondents; especially, if the survey is being completed outside of your company. Include a short introductory set of instructions at the top of the survey and additional instructions for specific questions, as needed. Don't forget to include how the completed surveys should be submitted.

✓ **Don't ask for personal information unless you need it.**

- Asking individuals to provide you with personal or demographic information may irritate some respondents and prevent them from completing your survey. (i.e., sexual orientation)

✓ **Keep the questions short and concise**

- The wording for survey questions should be short, concise, and easy to understand and complete the question.

✓ **Order/group questions according to subject**

- A good way to minimize respondent distraction is to group questions together by subject. This way your respondents can focus their thoughts and answer a series of questions around these thoughts. However, be careful how you word and order your questions because order effects may also cause response bias.

✓ **Test the survey questionnaire**

- Once you have developed your survey, have a few colleagues or subject matter experts complete the survey to make sure there are no problems.

Excel

Below is an example of a survey written in Microsoft Word. With the example below, you will need to create a data dictionary which will enable you to properly code the data in Microsoft Excel. The actual data dictionary follows the survey.

- **Survey : Trends in Associating**

Q1. How do you stay connected to others in your industry? (Select all that apply.)
□ Association/society conferences
□ Alumni networks
□ Social networking websites
□ Phone
□ Email
□ Blogs
□ Twitter
□ Other: ___
□ None of the Above

Q2. How do you use social networking sites? (Select all that apply.)
□ Staying connected with friends and family
□ Communicating with others in my industry
□ Communicating with clients
□ Marketing my company
□ Networking/finding professional opportunities
□ Staying connected to a professional community (i.e. Facebook group)

Q3. Which of the following social networking sites do you use? (Select all that apply.)
□ Facebook
□ Twitter
□ LinkedIn
□ FourSquare
□ Other______

Demographic Information

D1. What is your gender?
○ Male
○ Female

D2. What is your age?

D3. In which industry do you work?

○ Arts/Entertainment/Recreation	○ Automotive
○ Construction	○ Energy
○ Health Care	○ Hospitality

(continued)

(continued)

○ Information Technology
○ Manufacturing and Distribution
○ Marketing
○ Miscellaneous
○ Professional Services
○ Real Estate
○ Retail
○ Telecommunications
○ Transportation
○ Other______

D4. How many years have you been in this industry?

A data dictionary assigns a value for each response in a question. For example, Question 1 (Q1) is a multiple select question and may cause the respondent to select multiple items. For this type of question, each response will be categorized as an individual question with the coding being 1, if the item is selected, and 0, if the item is not selected.

- **Data Dictionary**

Below is an example of how a data dictionary would appear in an Excel for the survey example.

ID. Respondent Number		
[this is the number assigned to the respondent for record keeping]		
Q1. How do you stay connected to others in your industry?		
(a) Association/society conferences	1 = Selected	0 = Not Selected
(b) Alumni networks	1 = Selected	0 = Not Selected
(c) Social networking websites	1 = Selected	0 = Not Selected
(d) Phone	1 = Selected	0 = Not Selected
(e) Email	1 = Selected	0 = Not Selected
(f) Blogs	1 = Selected	0 = Not Selected
(g) Twitter	1 = Selected	0 = Not Selected
(h) Other	1 = Selected	0 = Not Selected
h_o.	[text response]	
(i) None of the Above	1 = Selected	0 = Not Selected
Q2. How do you use social networking sites? (Select all that apply.)		
(a) Non-professionally (to stay connected with friends and family)	1 = Selected	0 = Not Selected
(b) Communicating with others in my industry	1 = Selected	0 = Not Selected
(c) Communicating with clients	1 = Selected	0 = Not Selected
(d) Marketing my company	1 = Selected	0 = Not Selected
(e) Networking/finding professional opportunities	1 = Selected	0 = Not Selected
(f) Staying connected to a professional community (i.e. Facebook group)	1 = Selected	0 = Not Selected

(continued)

(continued)

Q3. Which of the following social networking sites do you use? (Select all that apply.)

(a) Facebook	1 = Selected	0 = Not Selected
(b) Twitter	1 = Selected	0 = Not Selected
(c) LinkedIn	1 = Selected	0 = Not Selected
(d) FourSquare	1 = Selected	0 = Not Selected
(e) Other	1 = Selected	0 = Not Selected
e_o.	[text response]	

D1. What is your gender?
1 = Male
2 = Female

D2. What is your age?
[text response]

D3. In which industry do you work?
1 = Arts/Entertainment/Recreation
2 = Automotive
3 = Construction
4 = Energy
5 = Health Care
6 = Hospitality
7 = Information Technology
8 = Manufacturing & Distribution
9 = Marketing
10 = Miscellaneous
11 = Professional Services
12 = Real Estate
13 = Retail
14 = Telecommunications
15 = Transportation
16 = Other

D3o. Other:
[text response]

D4. How many years have you been in this industry?
[text response]

◇ **Note**: It is a standard practice to code your questions with a **Q** for the general survey questions, **o** for other, and a **D** for demographic questions.

After creating the data dictionary you can begin entering and coding your response data according to the variables you have established.

The following example includes an example of 10 responses to the previous survey. The question labels are in the first column and are read across each row. Each respondent is a separate column.

ID	1	2	3	4	5	6	7	8	9	10
Q1a	1	0	1	0	1	1	1	1	0	1
Q1b	0	0	0	0	0	0	0	0	1	0

(continued)

(continued)

ID	1	2	3	4	5	6	7	8	9	10
Q1c	1	1	1	1	1	1	1	1	1	1
Q1d	0	1	0	0	0	1	0	0	0	1
Q1e	1	1	0	1	0	1	1	0	1	1
Q1f	0	0	0	0	1	0	0	0	0	1
Q1g	1	1	1	1	0	1	0	1	1	1
Q1h	0	0	0	0	0	0	0	0	0	0
Q1h_o										
Q1i	0	0	0	0	0	0	0	0	0	0
Q2a	1	1	0	1	1	1	1	1	1	1
Q2b	1	1	1	1	1	1	1	1	1	1
Q2c	1	0	0	0	1	0	0	1	0	0
Q2d	1	0	1	0	0	0	0	0	1	0
Q2e	0	1	0	1	0	0	0	0	0	1
Q2f	0	1	0	1	1	0	1	1	1	0
Q3a	1	1	1	1	1	1	1	1	1	1
Q3b	0	1	0	1	1	1	1	1	1	0
Q3c	1	1	1	1	1	1	1	1	1	1
Q3d	0	0	0	0	1	0	0	1	0	0
Q3e	0	0	0	0	0	0	0	0	0	0
Q3e_o										
D1	1	2	2	1	2	1	2	1	1	2
D2	25	32	56	45	22	60	38	42	54	35
D3	4	5	6	7	9	11	1	3	13	16
D3o										Tech
D4	2	5	7	15	2	40	10	24	30	6

After you have entered the data, you can analyze the numbers using descriptive statistics, frequency counts, and various other methods.

Final Thoughts and Activities

Practice Problems and Case Studies

1. Below is an example of a survey. Take a look at the questions below and describe how you can improve them.

> 1. Our in store greeting
>   Unacceptable—Bad—Good—Great—Excellent
> 2. Store cleanliness
>   Excellent—Great—Good—Bad—Unacceptable

(continued)

(continued)

3. Our awesome in store merchandise
   Not appropriate—Appropriate—Awesome
4. Our overall store selection and bathrooms
   Unacceptable—Bad—Good—Great—Excellent
5. Wait time at the checkout counter
   Too long—Just Right—Perfect
6. Our sincere "Thank you" upon leaving our store
   Yes/No
7. Your overall store experience
   Awesome—Good—Not Good——Poor

2. We ask the survey question "What do you remember about advertising for the Ford pickup trucks during the Super Bowl?" How should the codes be structured for this question?

Discussion Boards

1. How difficult is it to recruit households for the Nielsen panel? Why?
2. The use of Survey Monkey as an online tool has created some interesting ethical issues. Discuss.

Group Activity

1. Compile a list of at least five Census Bureau activities, and explain what type of information is derived from these surveys.
2. Build your own survey around employee satisfaction and test it out on a small sample to get feedback.

Parting Thought

A statistician is a professional who diligently collects facts and data, then carefully draws confusions about them.

Problem Solutions

Answers may vary but here is an example of how the survey items can be improved
How good would you rate the following store experiences:

1. Our in-store greeting
Very Unacceptable —Unacceptable—Neutral—Acceptable—Very Acceptable

2. Store cleanliness
Very Unacceptable —Unacceptable—Neutral—Acceptable—Very Acceptable

3. Our in store merchandise (loaded question)
Very Unacceptable —Unacceptable—Neutral—Acceptable—Very Acceptable

4A. Our overall store selection (double-barreled question)
Very Unacceptable —Unacceptable—Neutral—Acceptable—Very Acceptable

4B. Our bathrooms
Very Unacceptable —Unacceptable—Neutral—Acceptable—Very Acceptable

5. Wait time at the checkout counter
Very Unacceptable —Unacceptable—Neutral—Acceptable—Very Acceptable

6. Our "Thank you" upon leaving our store (Leading loaded question)
Very Unacceptable —Unacceptable—Neutral—Acceptable—Very Acceptable

7. Your overall store experience
Very Unacceptable —Unacceptable—Neutral—Acceptable—Very Acceptable

Note: (1) The items all should have the same rating scale; should not switch between rating scale and yes/no question types. (2) All questions should have the same directional scales; should not switch between starting with very acceptable to staring with very unacceptable. Pick a direction and stick with it.

2. This question can be set as:

- an open ended question, thus the responses should be coded based on recurring themes and ideas. OR
- a nominal multiple response select question, thus the responses would be coded as 1 for remember an aspect of the commercial or 0 for no response to the item. OR
- a nominal single select question: the responses can be coded using numbers, but the numbers are not scale able and are only categorical labels.

Chapter 6
Sampling

Case Study: Customizing a Meaningful Report for Infinity Auto Insurance

Roughly 2.1 Million Hispanics live in Houston, Texas. Infinity Auto Insurance has asked you to help them create a market research survey for the Houston, Texas area. The Infinity marketing group wants to better understand the Hispanic automobile insurance market in this large metropolitan area. Infinity Auto Insurance has decided that a telephone survey would reach their target population (less acculturated Hispanics) better than any other data collection method. To get a representative sample, Infinity Auto Insurance has decided that 500 respondents would be sufficient to draw some conclusions.

Infinity Auto Insurance wants to better understand some of the following information; what percentage of Hispanics are insured, what companies are Hispanics insured with, how willing are Hispanics to buy or switch insurance companies, etc. Finally, Infinity Auto Insurance would like to know what attributes keep Hispanics buying auto insurance in the Houston area.

Food for Thought

What types of questions would you ask?
What are some questions you would develop?
How would you ensure the questions were valid?
How long would your survey be?
What demographic attributes would you try to capture?

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_6,

Possible Answers

What types of questions would you ask?
Likert questions
Single response questions
General demographic questions

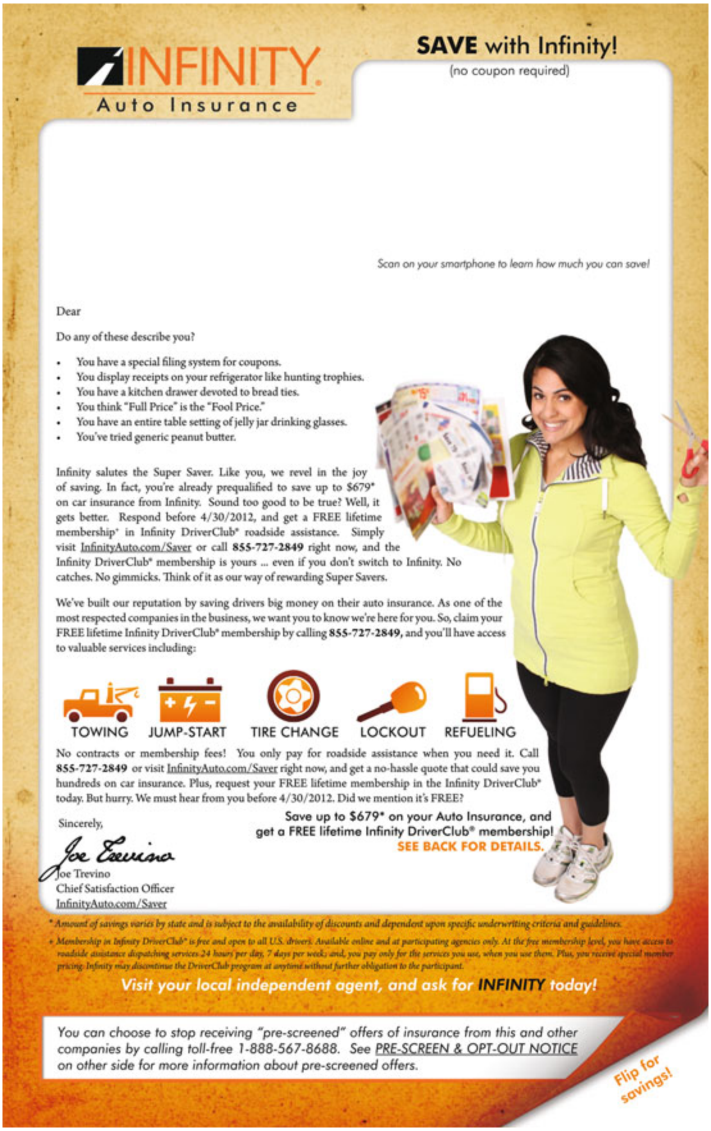
What are some questions you would develop?
Do you own a vehicle?

Do you have auto insurance?
How likely or unlikely would you be to switch auto insurance companies?
Have you heard of [___, ___, Infinity Auto] Insurance?
Have you seen ads for [___, ___, Infinity Auto] Insurance?
How favorable or unfavorable would you rate [___, ___, Infinity Auto Insurance] based on their reputation?

How would you ensure the questions were valid?
Talk to Subject Matter Experts about questions
Check to make sure the questions are not double barreled
Check to make sure there are no leading questions
Check to make sure the questions do not have any biases

How long would your survey be?
Roughly 5–10 min long

What demographic attributes would you try to capture?
Age
Language spoken at home
Hispanic/Non-Hispanic
Preferred Advertising Language
Highest Level of Education



INFINITY Auto Insurance

SAVE with Infinity!

(no coupon required)

Scan on your smartphone to learn how much you can save!

Dear

Do any of these describe you?

- You have a special filing system for coupons.
- You display receipts on your refrigerator like hunting trophies.
- You have a kitchen drawer devoted to bread ties.
- You think "Full Price" is the "Fool Price."
- You have an entire table setting of jelly jar drinking glasses.
- You've tried generic peanut butter.

Infinity salutes the Super Saver. Like you, we revel in the joy of saving. In fact, you're already prequalified to save up to \$679* on car insurance from Infinity. Sound too good to be true? Well, it gets better. Respond before 4/30/2012, and get a FREE lifetime membership+ in Infinity DriverClub® roadside assistance. Simply visit InfinityAuto.com/Saver or call **855-727-2849** right now, and the Infinity DriverClub® membership is yours ... even if you don't switch to Infinity. No catches. No gimmicks. Think of it as our way of rewarding Super Savers.

We've built our reputation by saving drivers big money on their auto insurance. As one of the most respected companies in the business, we want you to know we're here for you. So, claim your FREE lifetime Infinity DriverClub® membership by calling **855-727-2849,** and you'll have access to valuable services including:

TOWING JUMP-START TIRE CHANGE LOCKOUT REFUELING

No contracts or membership fees! You only pay for roadside assistance when you need it. Call **855-727-2849** or visit InfinityAuto.com/Saver right now, and get a no-hassle quote that could save you hundreds on car insurance. Plus, request your FREE lifetime membership in the Infinity DriverClub® today. But hurry. We must hear from you before 4/30/2012. Did we mention it's FREE?

Sincerely,

Joe Trevino

Joe Trevino
Chief Satisfaction Officer
InfinityAuto.com/Saver

Save up to \$679* on your Auto Insurance, and get a FREE lifetime Infinity DriverClub® membership!
SEE BACK FOR DETAILS.

** Amount of savings varies by state and is subject to the availability of discounts and dependent upon specific underwriting criteria and guidelines.*

+ Membership in Infinity DriverClub® is free and open to all U.S. drivers. Available online and at participating agencies only. At the free membership level, you have access to roadside assistance dispatching services 24 hours per day, 7 days per week; and, you pay only for the services you use, when you use them. Plus, you receive special member pricing. Infinity may discontinue the DriverClub program at anytime without further obligation to the participant.

Visit your local independent agent, and ask for INFINITY today!

You can choose to stop receiving "pre-screened" offers of insurance from this and other companies by calling toll-free 1-888-567-8688. See PRE-SCREEN & OPT-OUT NOTICE on other side for more information about pre-screened offers.

Flip for savings!

Key Concepts

Accuracy, Confidence level, Finite population, Infinite population, Proportion, Homogeneity, Population, Sample size, and Standard error.

Discussion

Sampling is not a desired activity. We only sample when we have certain limitations such as limited access to the entire population of data and/or limited resources in terms of time and money.

We achieve more accurate results when we can avoid sampling and analyze the entire population of data. Sampling is a last resort as it introduces complexities in analyzing the data and introduces approximations in the final results.

This chapter will assist you in determining how large your sample should be to achieve the desired level of statistical rigor. But remember, the sample data is only ever as good as the population data it was derived from. The goal is to have a large enough sample to be representative of the entire population. The statistical values you calculate using just the sample data should also describe the population data if the correct sample size has been used.

If we under sampled the picture in Fig. 6.1 we might end up with an inaccurate understanding of what it actually depicts. Can you see both a face and an Eskimo? If samples are too small they may not provide accurate generalizations of the entire population. If we had selected too small a sample of pixels from the image, we might see just the face or just the Eskimo.

The information to calculate the required sample size includes:

1. **Homogeneity** of the population – proportion of the variable of interest (p)
2. **Accuracy** or the range of error you can live within your sample statistics (±e)
3. **Confidence Level** or chance that you have selected a representative sample; most common is 90.0 %, 95.0 %, 99.7 % confidence level.

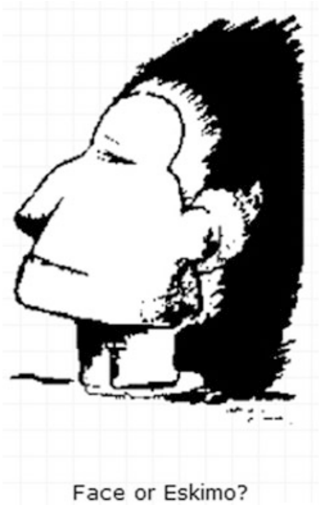


Fig. 6.1 Seeing both a face and an Eskimo in the same image

To minimize bias, an impartial and objective probability method should be used to select the sample. The most common example of this is a simple random sample. Each data point in the sample has an equal chance of being selected.

Homogeneity refers to the diversity of the population. Consider a company of employees from 10 different national cultures. The employees are surveyed about the acceptability of nepotism in the workplace. If the sample is too small you may have certain cultures underrepresented or not represented at all. Since national culture may influence the answers, it is important to have a large enough sample to be representative across all 10 cultures. If on the other hand our population data sample included only Americans, our sample would not have to be as large to remain representative of an American population. The higher the degree of homogeneity in the population, the smaller the sample required.

Accuracy of the data refers to the size of the standard error. The **standard error (SE)** is the likely size of a chance error and is often referred to as the "give or take" number. The idea is that the sample results are also the expected results for the population but will be off from this number by a certain amount. The SE indicates the likely size of that amount the number is off. This is described in more detail in Chap. 7.

In Chap. 3, we also defined **accuracy** as a measure of uncertainty around the mean. Accuracy answers the question: How close do you want your results to be to the true population? If we were willing to accept data with an associated standard error of 2, we could express 2 standard errors (SE) as a percent (%). To do this calculation we convert 2 SE to measured units and express that amount as a percent of the population. It is generally stated as "plus or minus 'X' percent".

The **confidence level or confidence interval** can also be expressed as a percentage. A 95.0 % confidence level means that if 100 samples are taken, 95 will contain the actual statistics of the population, in other words, "*would be representative of the population in question*." We can also express this by stating there is a 5 % probability that the sample is not representative of the population. Most workplace data is analyzed with a 95.0 % (within 2 SEs) or 99.7 % (within 3 SEs) confidence level. Remember the higher the percentage value of the confidence, the wider the range of acceptable data. This may seem counterintuitive but to get more data (99.7 % vs. 95.0 %) you need to widen the range of data values in your accepted region.

Here are some examples in which these three criteria are combined.

- **Example 1**
  If a survey uses a plus or minus 3 % accuracy and a 95.0 % confidence level, that means there is a 5 in 100 chance that the survey results are MORE than 3 % away from the results you would get if you had surveyed the entire population.
- **Example 2**
  If the average GPA of the sample is 3.2 ±0.3, the actual population could have a GPA anywhere from 2.9 to 3.5. If we wanted a 95.0 % confidence level, we would expect these results in 95 out of 100 samples.

- **Example 3**
  We can increase the confidence by adding observations to our sample (increase the sample size), or by widening the range (instead of 3.2 ±0.3, we have 3.2 ±0.5). To go from a 95.0 % to a 99.7 % confidence level, our range of data would increase from ±2 SEs to ±3 SEs.

Types of Problems

Mean versus proportion problems require slightly different treatment

This chapter will address calculating the appropriate sample size for **mean** and **proportion** problems. **Mean** problems involve collecting enough sample data to accurately solve for an average that represents the population. For example, you might want to collect enough data to solve for the average salary of a population. **Proportion** problems involve collecting enough sample data to solve for the proportion of a variable. For example, you might want to collect enough data to solve for the percentage of women over 21 years old in a given population to explore female alcohol consumption.

Finite versus infinite population size is another important factor in determining the appropriate sample size

Finite population means you know the size of the population. However if the sample size ends up being less than 5 % of the known population size, then the population size would be considered infinite. But more generally, an **infinite** population means you do not know the size of the population. In this chapter we will assume we have an infinite population. However a finite population correction factor will also be provided for those who may want to fine tune their estimates with known population size.

Rules of thumb

The **standard deviation estimate** may be useful in **mean** problems where the standard deviation is required. If the standard deviation of the population is unknown, you can estimate it by determining the largest expected observed value minus the smallest expected observed value. Some processes include a value called the **upper reasonable limit** which could also be used as the maximum value.

Although there is a not an explicit relationship between the range and standard deviation, there is a rule of thumb that can be useful is making a rough estimate of the standard deviation. This is referred to as the Range Rule:

$$\text{standard deviation estimate} = \frac{\text{largest possible value} - \text{smalest possible value}}{4}$$

While this estimate is not as reliable as an estimate based on calculations over a large number of data points, it is often useful as a preliminary estimate. The denominator is based on the Gaussian Rule that says 95 % of the data falls from two standard deviations below the mean to two standard deviations above the mean. Thus nearly all of our normal distribution would stretch out over a line segment that is a total of four standard deviations long. Not all data is normally distributed and bell curve shaped. But most data is well behaved enough that going two standard deviations away from the mean on either side should capture nearly all of the data. We estimate and say that four standard deviations is approximately the size of the range, and so the range divided by four is a rough approximation of the standard deviation.

Statistic	SD known	Formula
Mean	yes	$n = \frac{z^2\sigma^2}{e^2}$
	no	Use Range Rule to estimate σ
Proportion	NA	$n = \frac{z^2 p(1-p)}{e^2}$

Fig. 6.2 Sample size formula for infinite population problems

Proportion problems require knowing the statistic proportion of the population. If this is unknown we assume it to be .5 (50 %) to ensure we don't err on the side of underestimating the sample size. By using .5 for p in the formula will produce the largest sample size and we err on the side of having too large of a sample, rather than too small of a sample.

We provide a summary of the formulae based on mean and proportion in Fig. 6.2.

Where:

n = required sample size
z = z value for desired level of confidence
p = estimated value of the population proportion
e = accuracy, the +/− error you can live with (chance error)
σ = standard deviation for the population (known or estimated)

Excel

Problem Type: Infinite Mean

This is the formula for calculating a sample size for a problem in which you are calculating a mean for an infinite population.

Use this equation: $n = \frac{z^2\sigma^2}{e^2}$

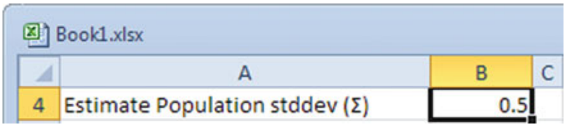
For example, let's suppose you what to estimate the mean of the population and the size of the population is infinite. We want to determine how many samples to take to ensure the mean of the sample matches the mean of the (infinite) population.

Example
Our highest cost for a bottle of water at a sports arena in the United States is $5.00 and our lowest cost is $3.00. We are willing to live with a $0.03 error value and require a confidence level of 95 %. Calculate the required sample size to estimate the average price of a bottle of water sold at this arena.

➢ Input estimated population standard deviation in cell B4. Remember if you don't have a clue about its size use the Range Rule from page 5

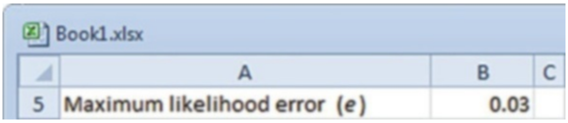
$$\text{standard deviation estimate} = \frac{\text{largest possible value} - \text{smalest possible value}}{4}$$

Based on the data, the largest value is $5.00 and the smallest possible value is $3.00. This results in a standard deviation estimate of .5.



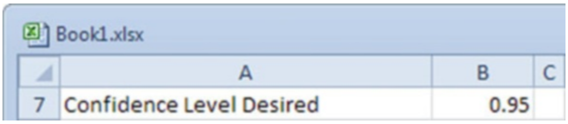
	A	B	C
4	Estimate Population stddev (Σ)	0.5	

➢ Input maximum likelihood error (the ± error you can live with) in the given units of that error in cell B5



	A	B	C
5	Maximum likelihood error (*e*)	0.03	

➢ Input the desired confidence level. This needs to be input as a decimal. Enter a **95 %** confidence level as **0.95** in cell B7

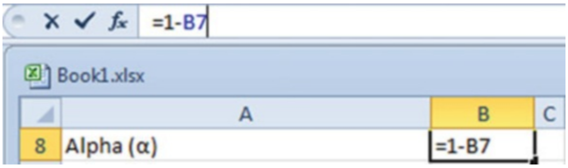


	A	B	C
7	Confidence Level Desired	0.95	

- ➢ Click on cell B8
- ➢ Click on the formula bar at the top of the page and type in

=1−B7

- ◈ Alpha is the failure rate rather than the success rate (confidence level). It is the same information but the convention is to use alpha in the formula.

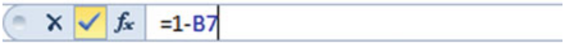


=1-B7

Book1.xlsx

	A	B	C
8	Alpha (α)	=1-B7	

- ➢ Click the green checkmark



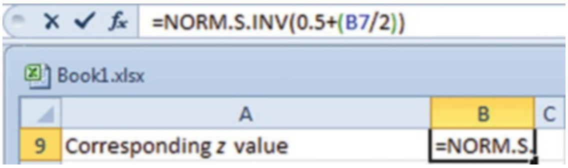
=1-B7

- ➢ Click on cell B9
- ➢ Click on the formula bar at the top of the page and type in

=NORM.S.INV(0.5+ (B7/2))

↑

- ◈ This is the z score. Watch that “**S**” in the formula.

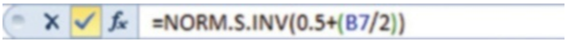


=NORM.S.INV(0.5+(B7/2))

Book1.xlsx

	A	B	C
9	Corresponding *z* value	=NORM.S	

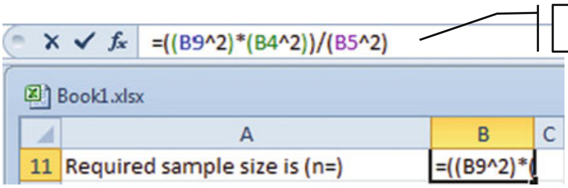
- ➢ Click the green checkmark



=NORM.S.INV(0.5+(B7/2))

- ➢ Click on cell B11
- ➢ Click on the formula bar at the top of the page and type in

= **((B9^2)*(B4^2))/(B5^2)**

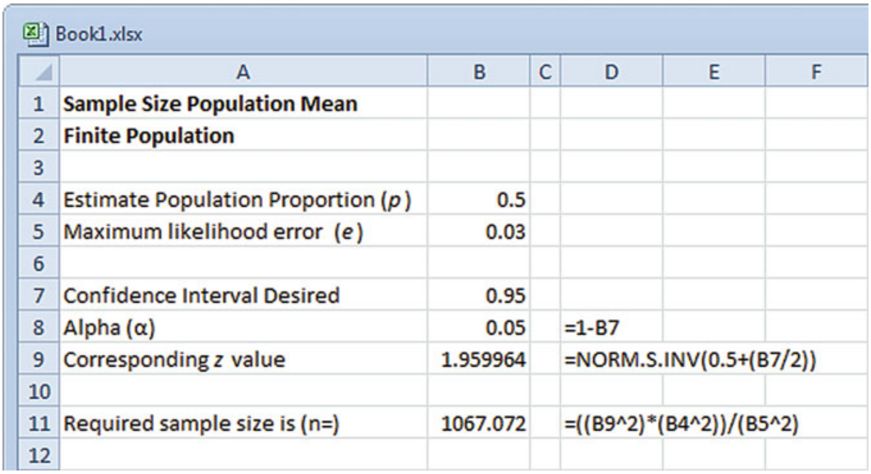


=((B9^2)*(B4^2))/(B5^2)

	A	B	C
11	Required sample size is (n=)	=((B9^2)*(	

Note: This is the sample size

➢ Click the green checkmark



	A	B	C	D	E	F
1	**Sample Size Population Mean**					
2	**Finite Population**					
3						
4	Estimate Population Proportion (p)	0.5				
5	Maximum likelihood error (e)	0.03				
6						
7	Confidence Interval Desired	0.95				
8	Alpha (α)	0.05		=1-B7		
9	Corresponding z value	1.959964		=NORM.S.INV(0.5+(B7/2))		
10						
11	Required sample size is (n=)	1067.072		=((B9^2)*(B4^2))/(B5^2)		
12						

Answer: We would round this answer up to 1,068 bottles, as we cannot have part of a bottle.

Practice Problem for Infinite Mean

A company is concerned about employee overtime and has hired you to survey the employee records. You are an outside consultant and do not have access to the company population, as this is a large multi-national corporation. The estimated population standard deviation has been reported as 100 h of annual overtime per non-exempt employee (σ)(B4). How many employee records should you include in your sample to calculate the average annual overtime per employee with a desired confidence level of 95 % (B7), within an accuracy of ±50 (B5) hours of overtime per employee (e)?

Excel		
Estimate for std dev (σ)	100	Type this in from the problem
Accuracy or maximum likelihood error (e)	50	Type this in from the problem

(continued)

(continued)

Excel		
Confidence level desired	0.95	The problem may state this as a % so convert to a decimal
The corresponding z-value is	1.96	Excel will calculate this from the Confidence level
The required sample size is n=	15.4	Excel will calculate

Answer: Always round up to be conservative on your sample size, so you need to survey 16 records.

Problem Type: Infinite Proportion

This is the formula for calculating a sample size for a problem in which you are calculating a proportion for an infinite population.

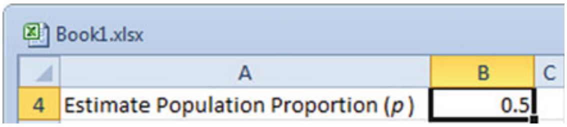
Use this equation: $\mathrm{n} = \dfrac{z^2 p(1-p)}{e^2}$

For example, let’s suppose you what to estimate the proportion of water bottles in the population that sold for more than $4.00; the size of the population is infinite. We want to determine how many samples to take to ensure the proportion of the sample matches the proportion of the (infinite) population.

Example

Our highest cost for a bottle of water at a sports arena in the United States is $5.00 and our lowest cost is $3.00. We are willing to live with a $0.03 error value and require a confidence level of 95 %. Our standard deviation estimate (p) is .5 based upon our previous example problem for calculating an estimate for an unknown standard deviation of the population. What is the required sample size to determine the proportion of bottles of water sold for more than $4.00 per bottle?

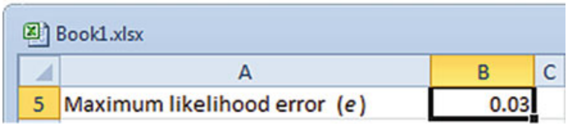
- Type in the population proportion in cell B4
  - If you do not know this use the most conservative estimate of 0.5.



Book1.xlsx

	A	B	C
4	Estimate Population Proportion (*p*)	0.5	

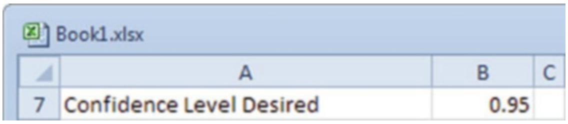
- Type in the maximum likelihood error (the ± error you can live with) in the given units of that error in cell B5



Book1.xlsx

	A	B	C
5	Maximum likelihood error (*e*)	0.03	

- ➢ Type in the desired confidence level desire. This needs to be input as a decimal. A **95.0 %** confidence level would be entered as **0.95** in cell B7

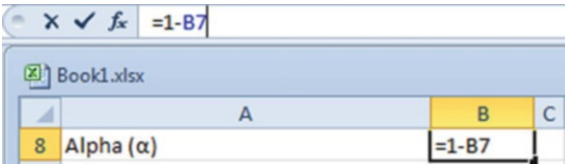


	A	B	C
7	Confidence Level Desired	0.95	

- ➢ Click on cell B8
- ➢ Click on the formula bar at the top of the page and type in

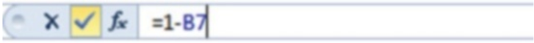
$$=\mathbf{1-B7}$$

(This is the failure rate rather than the success rate)



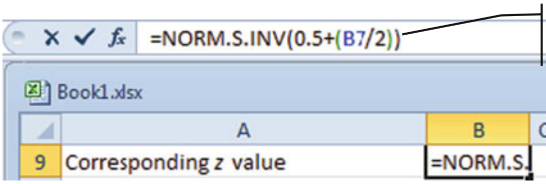
	A	B	C
8	Alpha (α)	=1-B7	

- ➢ Click the green checkmark



- ➢ Click on cell B9
- ➢ Click on the formula bar at the top of the page and type in

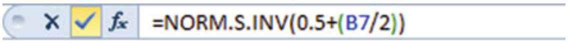
$$=\mathbf{NORM.S.INV(0.5+\ (B7/2))}$$



	A	B	C
9	Corresponding *z* value	=NORM.S.	

Note: This is the z-score.

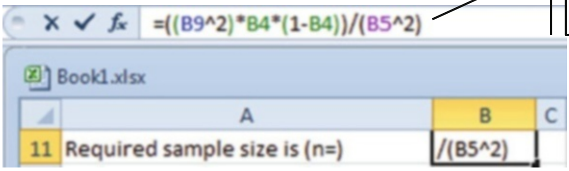
- ➢ Click the green checkmark



- ➢ Click on cell B11

➢ Click on the formula bar at the top of the page and type in

$$= ((B9\wedge2)*(B4)*(1-B4))/(B5\wedge2)$$



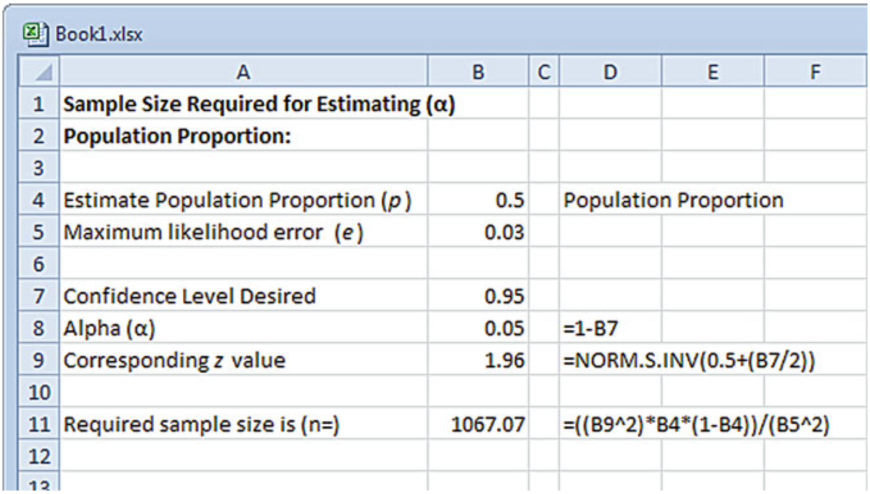
=((B9^2)*B4*(1-B4))/(B5^2)

Book1.xlsx

	A	B	C
11	Required sample size is (n=)	/(B5^2)	

Note: This is the sample size

➢ Click the green checkmark



Book1.xlsx

	A	B	C	D	E	F
1	**Sample Size Required for Estimating (α)**					
2	**Population Proportion:**					
3						
4	Estimate Population Proportion (*p*)	0.5		Population Proportion		
5	Maximum likelihood error (*e*)	0.03				
6						
7	Confidence Level Desired	0.95				
8	Alpha (α)	0.05		=1-B7		
9	Corresponding *z* value	1.96		=NORM.S.INV(0.5+(B7/2))		
10						
11	Required sample size is (n=)	1067.07		=((B9^2)*B4*(1-B4))/(B5^2)		
12						
13						

Answer: We would round this answer up to 1,068.

Practice Problem for Infinite Proportion

You are an outside consultant. The Chinese company who hired you wants to know the proportion of full time employees who might be willing to work overseas. The company population is unknown as records are not accurate about full time versus part time help. We have no idea how many employees may currently want to work overseas, so we assume 50 % (B4) of the entire population would accept an overseas assignment. How large a sample do I need with a confidence level of 99.7 % (B7) and accuracy of ± 4 % (B5)? My new survey needs to be a representative reflection of the entire population.

Excel		
Pop proportion p	0.5	Type this in from the problem; if expressed as a % change to a decimal. If unknown use .5
Maximum likely error, e:	0.04	Type this in from the problem (accuracy); this will be expressed as a % so convert to a decimal
Confidence level desired:	0.997	Input from problem
Corresponding z value:	2.97	Excel calculates this from the CI
Required sample size:	1378.3	

Answer: Always round up to be conservative on your sample size, so we need to survey 1,379 employees.

Finite Population Correction Factor (fpc)

This factor can be used to reduce the **sampling error** when you know the population size. The formula for the fpc follows:

$$\text{fpc} = \sqrt{\frac{(N-n)}{(N-1)}}$$

where **N** is the population size and **n** is the sample size.

Let's assume we have a population of 1,000 and a sample size calculated from our infinite sample size formula of 77. The fpc would equal .00961 or as a percentage .096 %. That means by using the fpc our sampling error will be reduced by almost 1 %.

The fpc can also be directly applied to the sample size by using it in the following format:

$$\text{n} = \frac{n_0 N}{(n_0 + (N-1))}$$

where **N** is the population size and **n** is the sample size and $\mathbf{n_o}$ is the sample size calculated using the infinite sample size formula.

Let's assume we have a population of 1,000 (N) and a sample size calculated from our infinite sample size formula of 77 (n_0). The new sample size n from the formula would equal 71.56 rounded up to 72. By applying the fpc we have reduced the required sample size from 77 to 72.

Final Thoughts and Activities

Practice Problems

1. There are several thousand machinery rebuilding and repairing companies in the USA. A tool manufacturer wishes to survey a simple random sample of these firms to find out what proportion of them is interested in a new tool design. If the tool manufacturer would like to be 95.0 % confident that the sample proportion is within .04 of the actual population proportion, how many machinery repairing and rebuilding companies would he include in the sample?
2. A state politician would like to determine the average amount earned during summer employment by state teenagers during the past summer's vacation period. She wants a 95.0 % confidence that the sample mean is within $50 of the actual population mean. Based on past studies she has estimated the population std. deviation to be about $400. How large should the sample be?
3. There are numerous CEOs in the United States. A political scientist wants to estimate, with 99.7 % confidence and within $3 what the average cost for their support staff is each year. Based on historical records the population standard deviation seems to be only $10. How many CEOs should be interviewed about their support staff?

Discussion Boards

1. What is RDD and what are the issues with using it?
2. Marine biologists often use the capture-recapture method for sampling. How does this work and how effective a sampling method is it?
3. What changes have occurred between the 1984 Gallup Survey and the most recent version?
4. Statistics as a discipline remains sharply divided on the fundamental definition of probability.

Group Activity

In 1936, the Literary Digest magazine predicted an overwhelming presidential victory for Alfred Landon who was running against Franklin Roosevelt. They predicted that Roosevelt would only get 43 % of the vote based on a sample of 2.4 million people. However, Roosevelt won the election with a landslide vote of 62–38 %. This is the largest error ever made by a major poll. George Gallup was just beginning his company. Using his own statistical methods, he predicted what the Digest's predictions were going to be in advance of their publication with an error of only 1 %. Find out where the Digest went wrong. Discuss how Gallup predicted the Digest results so well. After all he only had 36,000 people in his random sample.

Parting Thought

A frequentist is a person whose lifetime ambition is to be wrong 5 % of the time. A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule…

Problem Solutions

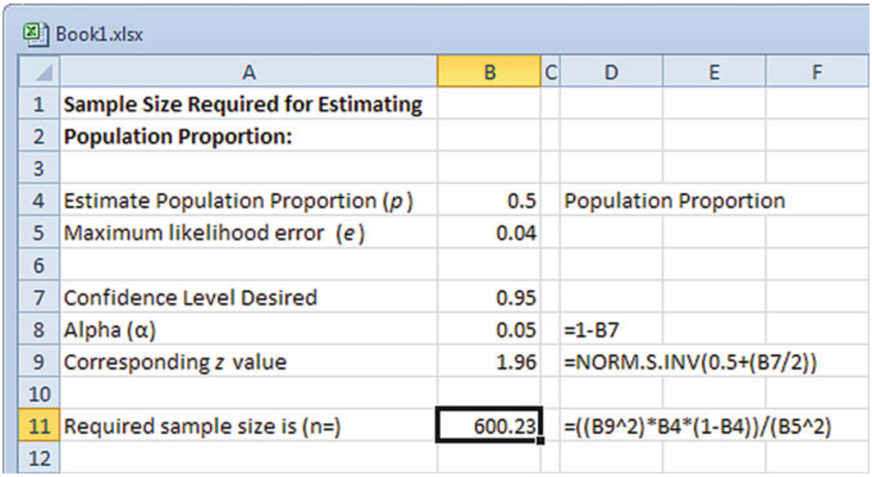
1. **Problem type: Infinite Proportion**

$$\text{n} = \frac{z^2 p(1-p)}{e^2}$$

Substitute:

p = assume 0.5 as our default when we don't know this number
e = .04
z = 1.96 (because the CI = 95.0 %)

Using our template we end up with 600.23 which after rounding is 601 for our sample size.



	A	B	C	D	E	F
1	**Sample Size Required for Estimating**					
2	**Population Proportion:**					
3						
4	Estimate Population Proportion (*p*)	0.5		Population Proportion		
5	Maximum likelihood error (*e*)	0.04				
6						
7	Confidence Level Desired	0.95				
8	Alpha (α)	0.05		=1-B7		
9	Corresponding *z* value	1.96		=NORM.S.INV(0.5+(B7/2))		
10						
11	Required sample size is (n=)	600.23		=((B9^2)*B4*(1-B4))/(B5^2)		
12						

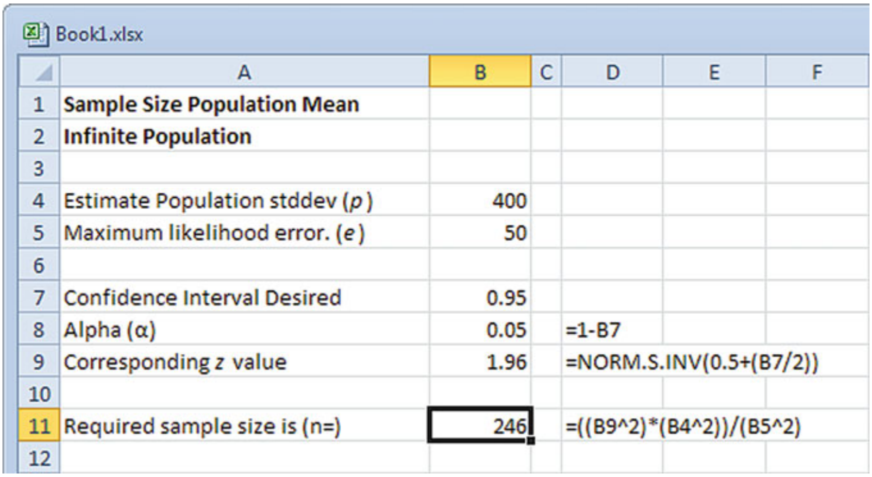
2. **Problem type: Infinite Mean**

$$n = \frac{z^2\sigma^2}{e^2}$$

Substitute:

e = \$50
z = 1.96 (because the CI = 95.0 %)
σ = \$400

Using our template we end up with 245.85, which after rounding is 246 for our sample size.



Book1.xlsx

	A	B	C	D	E	F
1	**Sample Size Population Mean**					
2	**Infinite Population**					
3						
4	Estimate Population stddev (*p*)	400				
5	Maximum likelihood error. (*e*)	50				
6						
7	Confidence Interval Desired	0.95				
8	Alpha (α)	0.05		=1-B7		
9	Corresponding *z* value	1.96		=NORM.S.INV(0.5+(B7/2))		
10						
11	Required sample size is (n=)	246		=((B9^2)*(B4^2))/(B5^2)		
12						

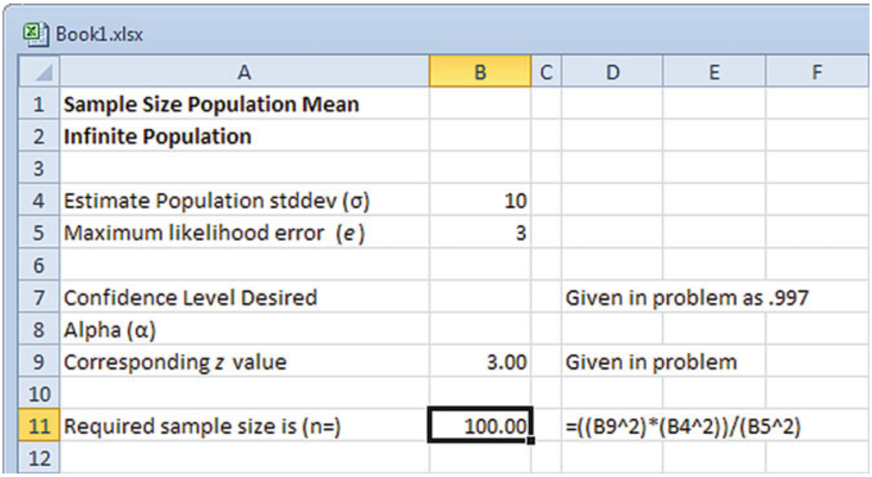
3. **Problem type: Infinite Mean**

$$n = \frac{z^2\sigma^2}{e^2}$$

Substitute:

e = \$3
z = 3.0 (because the CI = 99.7 %)
σ = \$10

Using our template we end up with 100 for our sample size.



	A	B	C	D	E	F
1	**Sample Size Population Mean**					
2	**Infinite Population**					
3						
4	Estimate Population stddev (σ)	10				
5	Maximum likelihood error (*e*)	3				
6						
7	Confidence Level Desired			Given in problem as .997		
8	Alpha (α)					
9	Corresponding *z* value	3.00		Given in problem		
10						
11	Required sample size is (n=)	100.00		=((B9^2)*(B4^2))/(B5^2)		
12						

Chapter 7
Inference

Key Concepts

Confidence interval, Estimator, Expected value, Population, Sample, and Standard error.

Discussion

This chapter introduces the concept of inferring statistical information about a population from sample data. The **sample** is the group of selected data chosen from the entire set of possible data (**population**). An **estimator** is a formula or process for using sample data to estimate a population statistic. Inference is not a preferred statistical process but becomes a required process when you are working with only part of a larger data set, in other words a sample of data. The first step is to assume an expected value and the second step is to fine tune that value using the standard error. The three basic problem types include inferring a population sum, a proportion from a sample data, and an average from the sample data. This chapter will include examples of all three of these problem types.

The **expected value, E(x),** pins down the center of the probability histogram and the **standard error (SE)** fixes its spread. Remember **standard deviation (SD)** is for a **list** of numbers, while **standard error (SE)** is for a chance **process**. The SD and SE are different. The expected value and standard error calculations vary based on the problem type.

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_7,

1. The **expected value, E(x), for the sum** of selections made at random with replacement from a box is:

$$\mathbf{E(x)\ for\ the\ sum} = (\mathbf{number\ of\ selections}) \times (\mathbf{average\ of\ box})$$
$$= (\boldsymbol{n}) \times (\bar{\boldsymbol{x}})$$

   But more generally, the **expected value** for the average or proportion of the population is the associated value calculated from the sample. The **standard error** fine tunes that value.

2. The **expected value** for the proportion of the population is inferred to be the expected value of the sample proportion and likewise.

$$\text{Proportion: } \mathrm{E(x)}_{\text{population}} = \mathrm{E(x)}_{\text{sample}}$$

3. The **expected value** for the average of the population is inferred to be the expected value of the sample average.

$$\text{Average}: \mathrm{E}(\bar{\boldsymbol{x}})_{\text{population}} = \mathrm{E}(\bar{\boldsymbol{x}})_{\text{sample}}$$

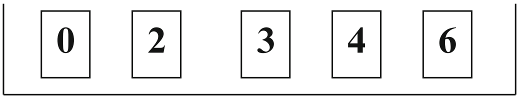
◈ Be careful to use the correct formula for the **standard error** (SE). Check if the **expected value** is for the sum, average, or proportion

The **standard error (SE)** is the likely size of a chance error and is often referred to as the “give or take” number. The idea is that the sample results are also the expected results for the population but will be off from this number by a certain amount. The SE indicates the likely size of that amount the number is off. Keep in mind the SE is an approximation. The SE is always interpreted in the same way, but can take on slightly different forms depending on what the problem requires.

1. **The SE for the sum of numbers**

$$\mathbf{SE} = \sqrt{\boldsymbol{sample\ size}} \times \boldsymbol{SD\ of\ the\ population}$$

Example: Box Model SE for the Sum



Suppose we make 25 selections from this box with replacement, and calculate the average of this box to be (0 + 2 + 3 + 4 + 6)/5 = 3, then the **expected value of the sum** of these tickets will be

$$\mathrm{E(x)} = n \times \bar{x}$$
$$\mathrm{E(x)} = 25 \times 3 = 75$$

To calculate the **standard error** we need the standard deviation of the box. To get the standard deviation you can use the STDEV function in Excel or just calculate by hand (see Chap. 3).

$$\begin{aligned}\mathbf{SD} &= 2\\ \mathbf{SE\ for\ the\ sum} &= [SQRT(n)] \times SD\\ \mathbf{SE\,for\,the\,sum} &= [SQRT(25)] \times SD\\ &= 5 \times 2\\ &= 10\end{aligned}$$

This means the expected sum of the 25 selections is likely to be 75 give or take 10. We would expect any sum of 25 numbers randomly selected from this box to fall within the range of 65–85.

2. **The SE for the average of numbers**

$$\mathbf{SE\,for\,the\ average} = \frac{\boldsymbol{SE\ for\ the\ sum}}{\boldsymbol{sample\ size}}$$

But this is more commonly written as:

$$\mathbf{SE} = \frac{\boldsymbol{SD}}{\sqrt{\boldsymbol{n}}}$$

Here is how we derive that formula, from the previous **SE for the Sum** for those who have a mathematical interest... others can skip over this section.

$$\begin{aligned}\mathbf{SE\,for\,the\ average} &= \frac{\boldsymbol{SE\ for\ the\ sum}}{\boldsymbol{n}} \times \frac{(\sqrt{\boldsymbol{n}})}{(\sqrt{\boldsymbol{n}})} = \frac{\sqrt{\boldsymbol{n}} \times \boldsymbol{SD}}{\boldsymbol{n}} \times \frac{(\sqrt{\boldsymbol{n}})}{(\sqrt{\boldsymbol{n}})}\\ &= \frac{n \times \boldsymbol{SD}}{\boldsymbol{n}(\sqrt{\boldsymbol{n}})} = \frac{\boldsymbol{SD}}{\sqrt{\boldsymbol{n}}}\end{aligned}$$

Example: SE for the Average

Find the SE for the average daily salary in the population when the sample size is 2,500, the standard deviation, as calculated from the sample data, is $25, and the expected average value is $250.00.

$$\mathbf{SE\,for\,the\ average} = \frac{\boldsymbol{SD}}{\sqrt{\boldsymbol{n}}} = \frac{25}{50} = 0.5$$

This means the expected average of the 2,500 daily salaries is likely to be $250.00 give or take $0.50.

3. **The SE for proportion percent**

   This SE needs to be reported as a percentage

$$\mathbf{SE} = \left(\frac{\boldsymbol{SD}}{\sqrt{\boldsymbol{n}}}\right) \times 100\%$$

Example: SE for the Proportion

Find the SE for the proportion of female MBA students in the population when the sample size is 100 and the standard deviation, as calculated from the sample data, is 2 people (female MBA students) and the expected proportion from the sample is 40 people or 40 %.

$$\mathbf{SE} = \frac{SD}{\sqrt{n}} \times 100\% = \frac{2}{10} \times 100\% = 20\%$$

This means the expected sum of the 100 MBA student selections is likely to be 40 % female MBA students give or take 20 %.

It is important when working with inference statistics and sample data to include confidence intervals. Confidence intervals indicate how often you will get the same results with different samples. For example a 95 % confidence interval indicates the same results will occur in 95 out of 100 samples from the same population.

A **confidence interval** is a range of values used to estimate the true value of a population statistic. The definition of degree of confidence uses **"α"** (**alpha**) to represent a probability or area. The value of **α** is the complement of the **degree of confidence.** For a 95 % (.95) degree of confidence, $\alpha = 0.05$. For a 0.99 (or 99 %) degree of confidence, $\alpha = 0.01$.

The confidence interval is determined by the **standard error.**

- The sample percentage interval of ± 1 SE corresponds to a 68 % confidence interval for the population percentage.
- The sample percentage interval of ± 2 SEs corresponds to a 95 % confidence interval for the population percentage.
- The sample percentage interval of ± 3 SEs corresponds to a 99.7 % confidence interval for the population percentage.

Confidence levels are often stated as being "about" so much. That is because the **standard errors** have been estimated from the data and the normal "approximation" has been used. We are not dealing with an exact number. We are dealing with an approximation.

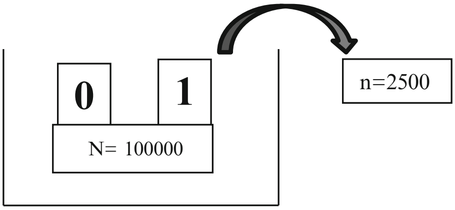
Inferring Proportions

Example Problem

There is an election for the President of the Academy of Management this year. There are 100,000 eligible voting members around the world, but you only want to enter if you have a good chance of winning. You take a simple random sample of 2,500 members through some of the regional chapters that you have access. In the sample, 1,328 members favor you. What should you expect on voting day when the entire population of 100,000 voting members participates?

In this problem, we have sample data and want to infer something about a proportion (percentage) for the entire population. In statistical terms, we want to know the **expected value** for the proportion of voters who will support the candidate, and calculate how good an estimate that is using the **standard error** for the sum of the voting members**.**

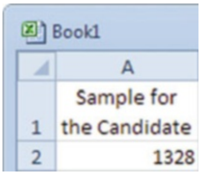
When setting up the box model, it is mandatory to assign the "1" to the variable of interest; the variable you are interested in counting (summing). In this case, it is the number of votes favoring the candidates.



Excel

➢ Input in A1 **Sample for the Candidate** and in A2 **1328**

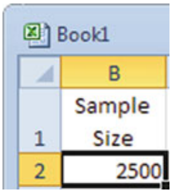
Note: This number is given from number of positive votes for the candidate in the sample.



	A
1	Sample for the Candidate
2	1328

- ➢ Input in B1 **Sample Size** and in B2 **2500**

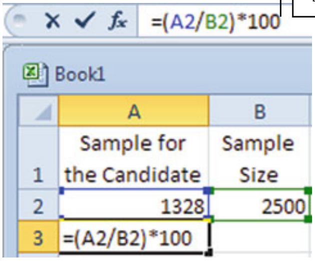
Note: given in problem as the actual sample size.



	B
1	Sample Size
2	2500

- ➢ Input in A3 = **(A2/B2) * 100**

Note: This formula calculates a percentage of positive votes in the sample.

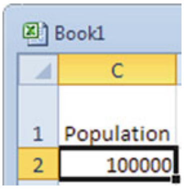


=(A2/B2)*100

	A	B
1	Sample for the Candidate	Sample Size
2	1328	2500
3	=(A2/B2)*100	

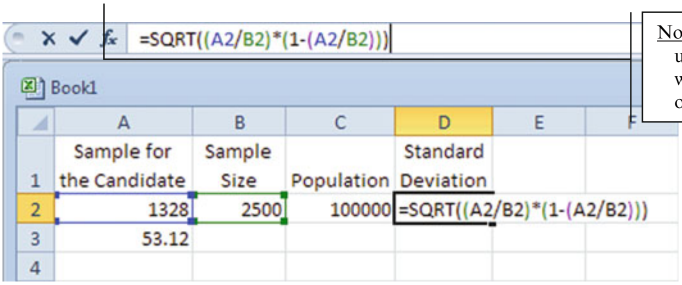
- ➢ After typing in the formula click on the green checkmark
- ➢ Input in C1 **Population** and in C2 **100000**

Note: given in problem as the entire population size.



	C
1	Population
2	100000

- ➢ Input in D1 **Standard Deviation**, and in D2 input = **SQRT((A2/B2)*(1−(A2/B2)))**



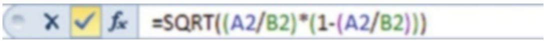
=SQRT((A2/B2)*(1-(A2/B2)))

Book1

	A	B	C	D	E	F
1	Sample for the Candidate	Sample Size	Population	Standard Deviation		
2	1328	2500	100000	=SQRT((A2/B2)*(1-(A2/B2)))		
3	53.12					
4						

Note: This formula can be used to calculate SD when you have a zero, one box.

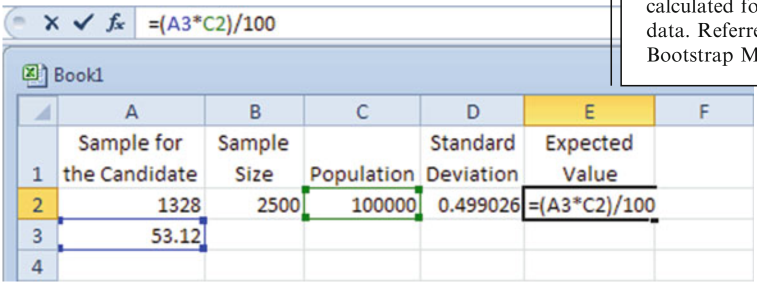
- ➢ After you have typed in this formula, click the checkmark



=SQRT((A2/B2)*(1-(A2/B2)))

- ➢ Input in E1 **Expected Value** and in E2 input = **(A3* C2)/100**

Note: You assume the same percentage of votes in the population as you calculated for the sample data. Referred to as the Bootstrap Method.

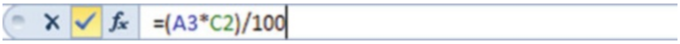


=(A3*C2)/100

Book1

	A	B	C	D	E	F
1	Sample for the Candidate	Sample Size	Population	Standard Deviation	Expected Value	
2	1328	2500	100000	0.499026	=(A3*C2)/100	
3	53.12					
4						

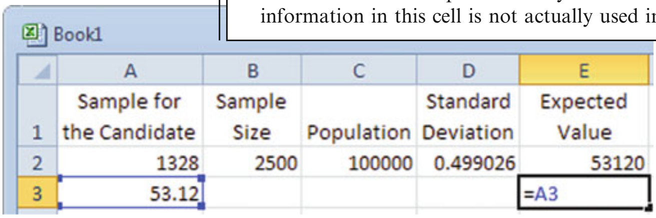
- ➢ After you have typed in the formula, click the checkmark



=(A3*C2)/100

- ➢ Input in E3 = **A3**

Note: This is just showing the same percentage in the population as you calculated in the sample. You only show here for completeness. The information in this cell is not actually used in any of the calculations

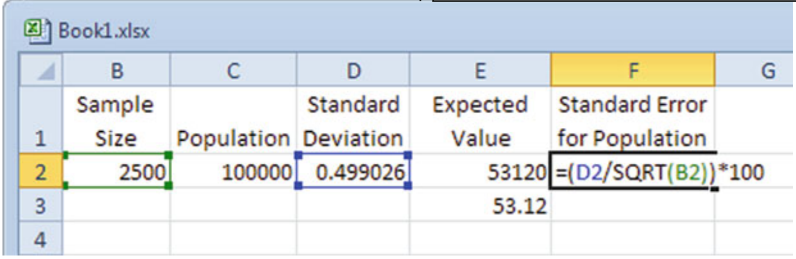


Book1

	A	B	C	D	E
1	Sample for the Candidate	Sample Size	Population	Standard Deviation	Expected Value
2	1328	2500	100000	0.499026	53120
3	53.12				=A3

- ➢ Click on the green checkmark
- ➢ Input in F1 **Standard Error for Population**
- ➢ Input in F2 = **(D2/SQRT (B2))*100**

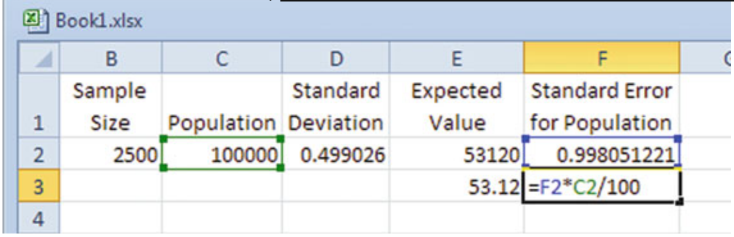
Note: This is the SE for the population as %


Book1.xlsx

	B	C	D	E	F	G
1	Sample Size	Population	Standard Deviation	Expected Value	Standard Error for Population	
2	2500	100000	0.499026	53120	=(D2/SQRT(B2))*100	
3				53.12		
4						

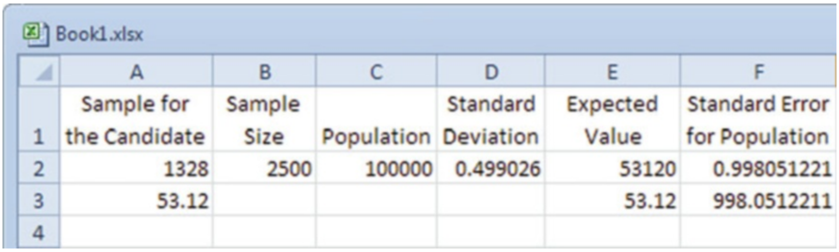
- ➢ Input in F3 = **F2*C2/100**

Note: This is the SE for the population as an actual headcount


Book1.xlsx

	B	C	D	E	F
1	Sample Size	Population	Standard Deviation	Expected Value	Standard Error for Population
2	2500	100000	0.499026	53120	0.998051221
3				53.12	=F2*C2/100
4					

For this problem, you should expect to receive 53,120 votes (E2) if all 100,000 votes were submitted, or 53 % of the vote (E3). But, this is only an estimate, and we know there is an error due to the process of sampling. That error turns out to be about 1 % (F2) or 998 votes (F3).


Book1.xlsx

	A	B	C	D	E	F
1	Sample for the Candidate	Sample Size	Population	Standard Deviation	Expected Value	Standard Error for Population
2	1328	2500	100000	0.499026	53120	0.998051221
3	53.12				53.12	998.0512211
4						

So you could expect to receive between 52,122 and 54,118 votes. Sounds like a winning proposition!

◇ Be careful how many decimal places you carry in these estimation calculations because you can't have a fraction of a vote.

Inferring Averages

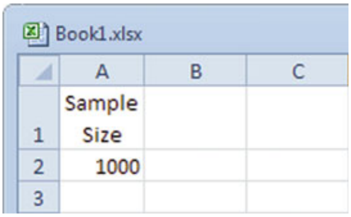
Example Problem

Suppose a city manager wants to know the average income of the 25,000 families living in his town. He hires a survey organization to take a simple random sample of 1,000 families. The total of all the sample incomes is $32,396,714. Therefore, the average sample income for the 1,000 families in the sample must be $32,396,714/1,000 = $32,397. It turns out that the standard deviation for the sum of the sample of salaries is $19,000. What is the average income for all 25,000 families likely to be, plus or minus a chance error? (This is the SE for an average.)

Excel

➢ Input in A1 **Sample Size** and in A2 **1000**

Note: given in problem as the sample size.

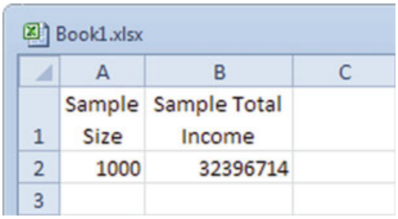


Book1.xlsx

	A	B	C
1	Sample Size		
2	1000		
3			

➢ Input in B1 **Sample Total Income** and in B2 **32396714**

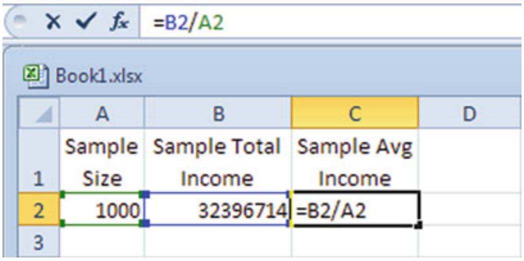
Note: given in problem as the sum of the incomes of the sample families.



Book1.xlsx

	A	B	C
1	Sample Size	Sample Total Income	
2	1000	32396714	
3			

➢ Input in C1 **Sample Avg Income** and input in C2 = **B2/A2**



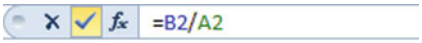
=B2/A2

Book1.xlsx

	A	B	C	D
1	Sample Size	Sample Total Income	Sample Avg Income	
2	1000	32396714	=B2/A2	
3				

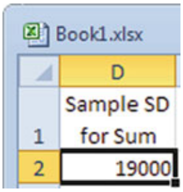
Note: Calculating the average income for the sample data. This may or may not be given in the problem. We infer this is also the average income for the population.

➢ After you have typed in the formula click the checkmark



=B2/A2

➢ Input in D1 **Sample SD** and in D2 **19000**

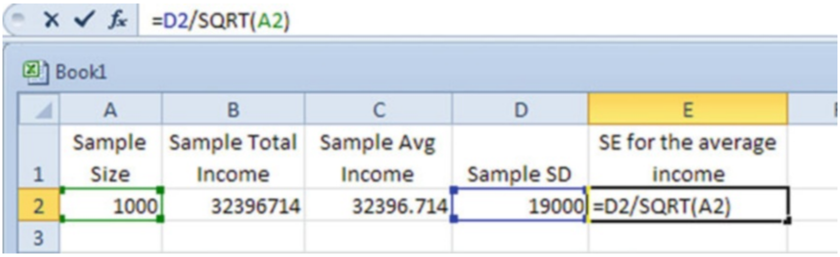


Book1.xlsx

	D
1	Sample SD for Sum
2	19000

Note: given in problem as the standard deviation for the sum of the sample salaries.

➢ Input in E1 **SE for the average income** and input in E2 = **D2/SQRT(A2)**

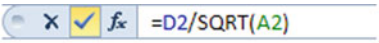


=D2/SQRT(A2)

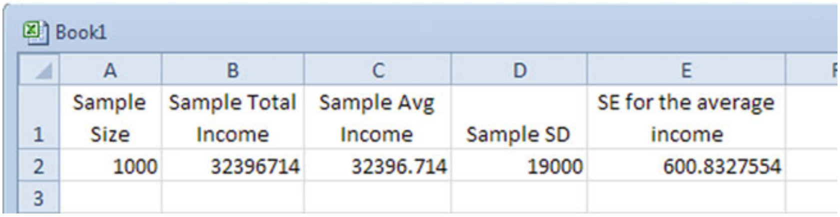
Book1

	A	B	C	D	E
1	Sample Size	Sample Total Income	Sample Avg Income	Sample SD	SE for the average income
2	1000	32396714	32396.714	19000	=D2/SQRT(A2)
3					

➢ After you have typed in the formula click the checkmark



=D2/SQRT(A2)



	A	B	C	D	E	F
1	Sample Size	Sample Total Income	Sample Avg Income	Sample SD	SE for the average income	
2	1000	32396714	32396.714	19000	600.8327554	
3						

Therefore, the average income of the population is most likely to be around $32,400 (C2), which is inferred at the beginning of the problem to be the same as the sample average income, give or take $600 (E2) which is the population SE for the average income. The $600 represents the standard error or the margin of error for this estimate.

Confidence Intervals with Proportion Inference

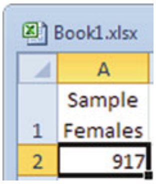
Example Problem

Your company is considering opening a new location in a nearby town with a population of 25,000 people, and you need more female workers for your plant because they are more interested in the female products produced. You hire a consultant to complete a simple random sample of 1,600 persons from the town. It turns out 917 people in the sample are females. Find a 95 % confidence interval for the proportion of females among all 25,000 residents.

Excel

➢ Input in A1 **Sample Females** and in A2 **917**

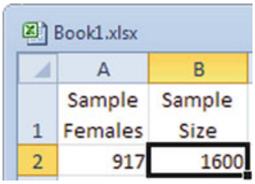
Note: given in problem as the number of females in the sample.



	A
1	Sample Females
2	917

➢ Input in B1 **Sample Size** and in B2 **1600**

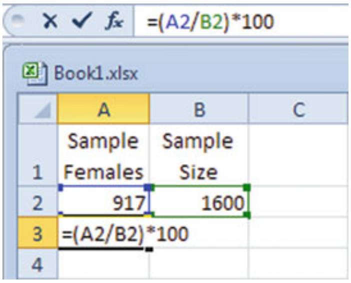
Note: given in problem as the sample size.



	A	B
1	Sample Females	Sample Size
2	917	1600

➢ Input in A3 = **(A2/B2)*100**

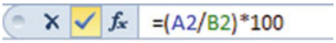
Note: Convert to a percentage of females in the sample.



=(A2/B2)*100

	A	B	C
1	Sample Females	Sample Size	
2	917	1600	
3	=(A2/B2)*100		
4			

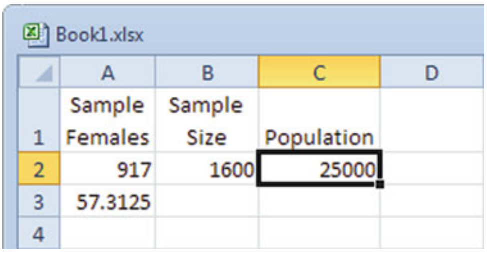
➢ After you have typed in the formula, click the checkmark



=(A2/B2)*100

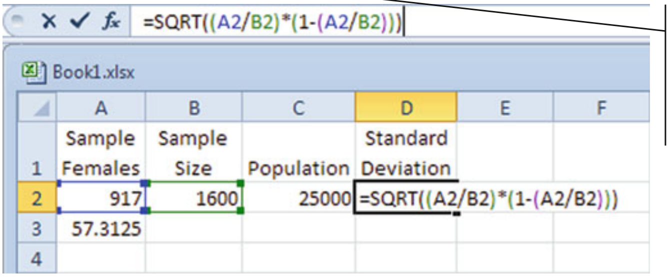
➢ Input in C1 **Population** and in C2 **25000**

Note: given in problem as the population size.



	A	B	C	D
1	Sample Females	Sample Size	Population	
2	917	1600	25000	
3	57.3125			
4				

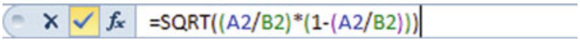
➢ Input in D1 **Standard Deviation** and in D2 input = **SQRT((A2/B2)*(1−(A2/B2)))**


=SQRT((A2/B2)*(1-(A2/B2)))

	A	B	C	D	E	F
1	Sample Females	Sample Size	Population	Standard Deviation		
2	917	1600	25000	=SQRT((A2/B2)*(1-(A2/B2)))		
3	57.3125					
4						

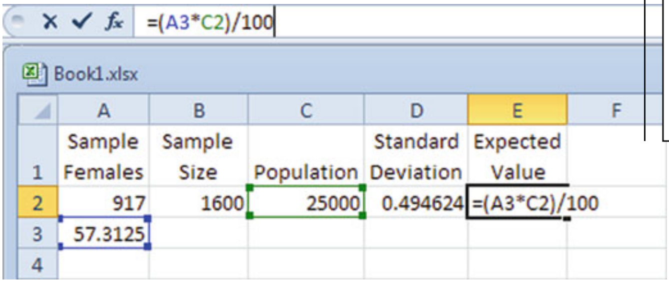
Note: This formula can be used to calculate SD when you have a zero, one box.

➢ After you have typed in the formula, click the checkmark


=SQRT((A2/B2)*(1-(A2/B2)))

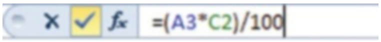
➢ Input in E1 **Expected Value** and input in E2 = **(A3*C2)/100**

Note: You assume the same percentage of females in the population as you calculated for the sample data. This is referred to as the Bootstrap Method.


=(A3*C2)/100

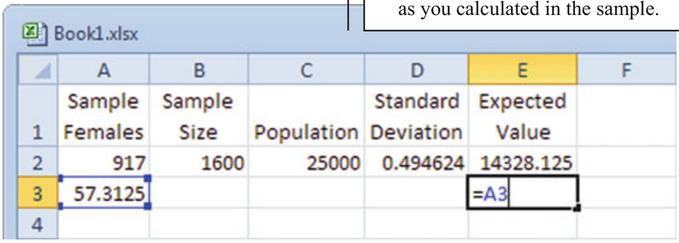
	A	B	C	D	E	F
1	Sample Females	Sample Size	Population	Standard Deviation	Expected Value	
2	917	1600	25000	0.494624	=(A3*C2)/100	
3	57.3125					
4						

➢ After you have typed in the formula, click the checkmark.


=(A3*C2)/100

➢ Input in E3 = **A3**

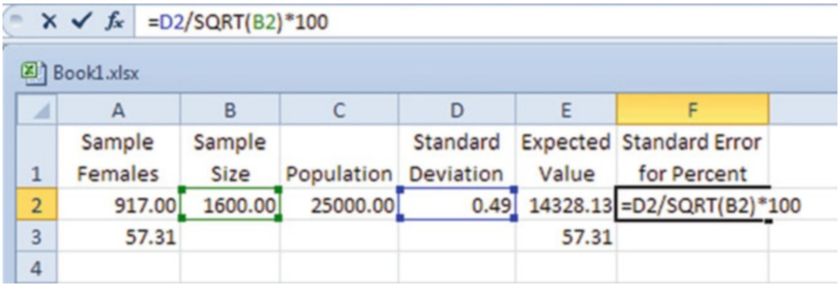
Note: This is showing the same percentage in the population as you calculated in the sample.



	A	B	C	D	E	F
1	Sample Females	Sample Size	Population	Standard Deviation	Expected Value	
2	917	1600	25000	0.494624	14328.125	
3	57.3125				=A3	
4						

➢ Input in F1 **Standard Error for Percent**

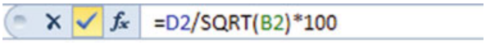
➢ Input in F2 = **D2/SQRT(B2)*100**



=D2/SQRT(B2)*100

	A	B	C	D	E	F
1	Sample Females	Sample Size	Population	Standard Deviation	Expected Value	Standard Error for Percent
2	917.00	1600.00	25000.00	0.49	14328.13	=D2/SQRT(B2)*100
3	57.31				57.31	
4						

➢ After you have typed in the formula click the checkmark.

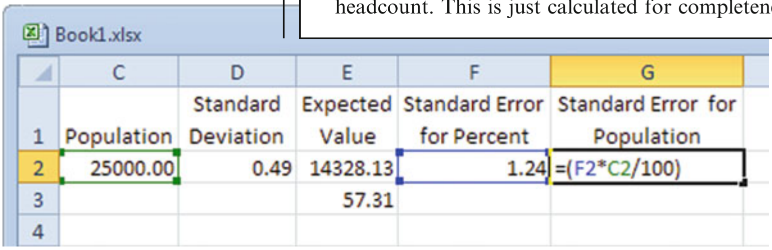


=D2/SQRT(B2)*100

➢ Input in G1 **Standard Error for Population**

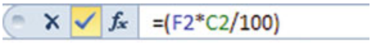
➢ Input in G2 = **(F2*C2/100)**

Note: This is the population standard error as an actual headcount. This is just calculated for completeness



	C	D	E	F	G
1	Population	Standard Deviation	Expected Value	Standard Error for Percent	Standard Error for Population
2	25000.00	0.49	14328.13	1.24	=(F2*C2/100)
3			57.31		
4					

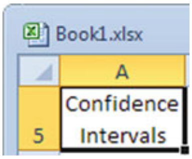
➢ After you have typed in the formula, click the checkmark



=(F2*C2/100)

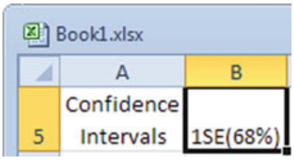
The expected value for the percentage of females in the population of workers is 57 % (E3) and is likely to be off the actual percentage of females by about 1.24 % (F2) or so. Now, we need to consider the confidence interval calculations.

➢ Input in A5 **Confidence Intervals**



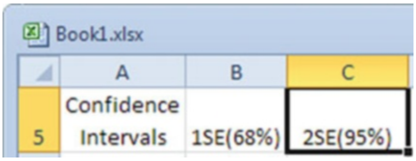
	A
5	Confidence Intervals

➢ Input in B5 **1SE (68%)**



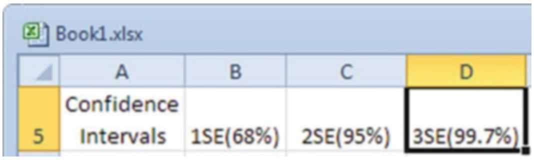
	A	B
5	Confidence Intervals	1SE(68%)

➢ Input in C5 **2SE (95%)**



	A	B	C
5	Confidence Intervals	1SE(68%)	2SE(95%)

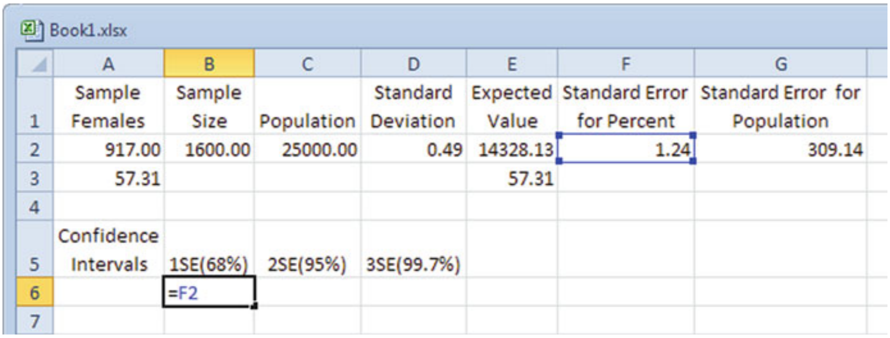
➢ Input in D5 **3SE (99.7%)**



	A	B	C	D
5	Confidence Intervals	1SE(68%)	2SE(95%)	3SE(99.7%)

➢ Input in B6 = **F2**

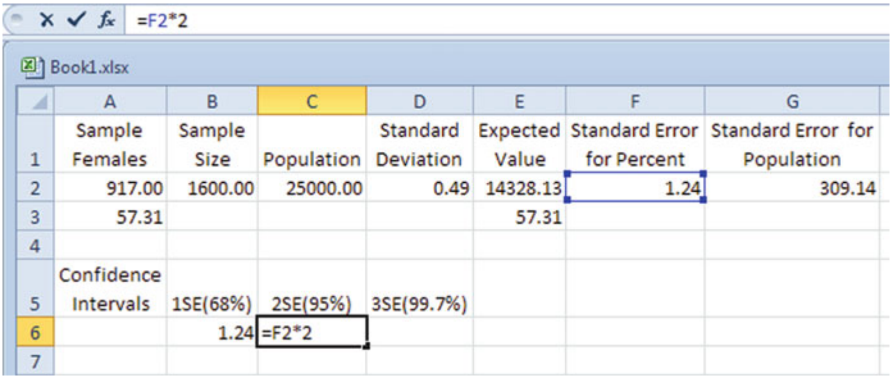
Note: This is the population Standard Error as a %



Book1.xlsx

	A	B	C	D	E	F	G
1	Sample Females	Sample Size	Population	Standard Deviation	Expected Value	Standard Error for Percent	Standard Error for Population
2	917.00	1600.00	25000.00	0.49	14328.13	1.24	309.14
3	57.31				57.31		
4							
5	Confidence Intervals	1SE(68%)	2SE(95%)	3SE(99.7%)			
6		=F2					
7							

➢ Input in C6 = **F2*2**

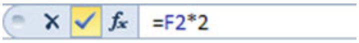


=F2*2

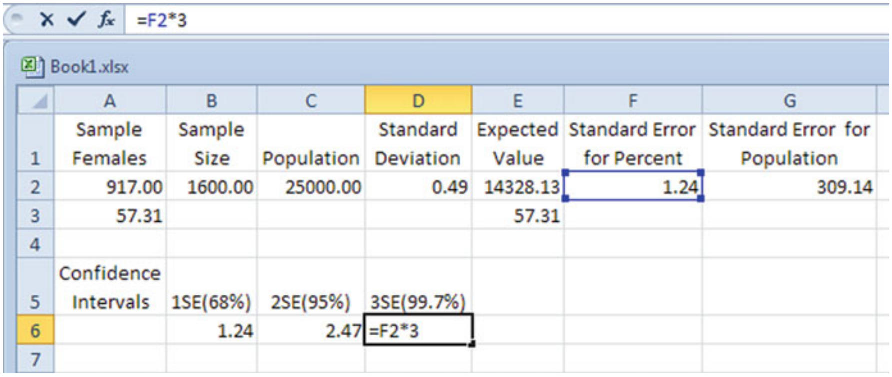
Book1.xlsx

	A	B	C	D	E	F	G
1	Sample Females	Sample Size	Population	Standard Deviation	Expected Value	Standard Error for Percent	Standard Error for Population
2	917.00	1600.00	25000.00	0.49	14328.13	1.24	309.14
3	57.31				57.31		
4							
5	Confidence Intervals	1SE(68%)	2SE(95%)	3SE(99.7%)			
6		1.24	=F2*2				
7							

➢ After you have typed in the formula, click the checkmark.



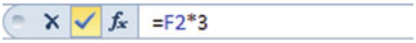
➢ Input in D6 = **F2*3**



=F2*3

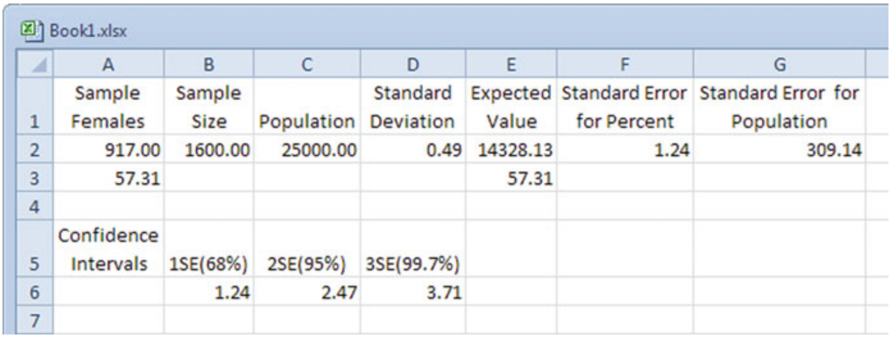
	A	B	C	D	E	F	G
1	Sample Females	Sample Size	Population	Standard Deviation	Expected Value	Standard Error for Percent	Standard Error for Population
2	917.00	1600.00	25000.00	0.49	14328.13	1.24	309.14
3	57.31				57.31		
4							
5	Confidence Intervals	1SE(68%)	2SE(95%)	3SE(99.7%)			
6		1.24	2.47	=F2*3			
7							

➢ After you have typed in the formula, click the checkmark.



=F2*3

A 95 % confidence interval for the percentage of females among the 25,000 residents corresponds to 2 SE or in other words ±2.47 % (C6).



	A	B	C	D	E	F	G
1	Sample Females	Sample Size	Population	Standard Deviation	Expected Value	Standard Error for Percent	Standard Error for Population
2	917.00	1600.00	25000.00	0.49	14328.13	1.24	309.14
3	57.31				57.31		
4							
5	Confidence Intervals	1SE(68%)	2SE(95%)	3SE(99.7%)			
6		1.24	2.47	3.71			
7							

The expected value for the percentage of females in the population of workers is 57 % (E3) and is likely to be off the actual percentage of females by about 1.24 % (F2) or so. We can be 95 % confident that between 54.53 % and 59.47 % of the residents in this town are female. That sounds like enough females for us to consider this as a viable town for opening up our new plant.

Final Thoughts and Activities

Practice Problems and Case Studies

1. Examine a current magazine and take a sample of 30 pages. Record how many of those pages have some form of advertising on them. Based on the results construct a 95 % confidence interval estimate of the percentage of pages that you would expect to have advertising for the entire magazine.
2. In Florida, members of the citrus industry make extensive use of statistical methods. One issue involves how growers are paid for oranges used to make orange juice. A sample is taken from each truckload of oranges. This sample is weighed and squeezed, and the amount of sugar in the juice is measured. Based on these sample results, an estimate is made of the total sugar in the truckload. Payment is based on the % of sugar. In a similar vein, how do customs officers decide how many boxes to inspect in a company's shipments at the docks? How about the security checks at airports? Any concerns with using samples and inference in these situations?
3. A real estate company has been provided with a portfolio of townhomes and single family homes at Tahoe Keys in California. The company surveys a nearby affluent town of 25,000 households to determine which form of residence is preferred as a vacation home, so they can target their marketing accordingly. The company took a sample of 500 of these households. It turns out only 79 households preferred single family residences as a vacation home over townhomes.

   (a) Calculate the proportion of households in the town who prefer townhomes as a vacation home.
   (b) About how far off to you expect this estimate to be off?
   (c) Calculate the 99 % confidence interval for the expected value.

4. The HR Director of a large corporation wants to study absenteeism among administrative workers at the corporate office during the year. A random sample of 25 administrative workers reveals that the average number of days absent was 12 days. Construct a 95 % confidence interval for the population average absences for the clerical workers during the year.

Discussion Boards

1. Daniel Yankelovich states "Warning labels about sampling error say nothing about whether or not the public is conflict-ridden or has given a subject much thought. This is the most serious source of opinion poll misinterpretation."
2. Studies show that if children took the 1932 IQ test today, half would score above 120 on the 1932 scale. Currently IQ scores have a mean of 100, and a SD of

15 and are normally distributed. If today's American children took the same IQ test used in 1932, the mean score would be 120, instead of 100 as it was in 1932. The net effect is that today's IQ scores are 20 points higher than in 1932. Any concerns?

3. "When you or anyone else attempts to tell me that a sample of 1,223 persons account for our opinions and tastes, here in America, I get mad as hell! This seems astounding and unfair and should be outlawed!" This is a case in which a sample poll of 1,223 people was projected to represent 120 million. Do you agree with the concern?
4. About 30 % of Americans aged 19–28 claim that they have used an illicit drug other than marijuana. If a random sample of 40 Americans aged 19–28 finds 21 who claim to have used an illicit drug other than marijuana, would you be surprised? Explain these statistics.

Group Activity

1. Sherlock Holmes often used inference. He often discussed his process of "reasoning backwards". Choose one of his cases and explain how inference was successfully employed.
2. Research how the *Current Population Survey* is used to estimate US unemployment rates.
3. Research the type of data and statistical analysis that form the foundation of Neilson Media Research.
4. Using surveys to estimate rare events typically leads to overestimating. For example, the National Rifle Association reports three million dues-paying members or about 1.5 % of American adults. In national random telephone surveys, however 4–10 % of respondents claim that they are dues-paying NRA members. Do your own web research. Discuss possible reasons for bias and methods of minimizing the bias.

Parting Thought

You know how dumb the average person is? Well, by definition, half of the population is even dumber than that...

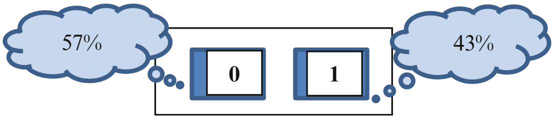
Problem Solutions

1. Examine a current magazine and take a sample of 30 pages. Record how many of those pages have some form of advertising on them. Based on the results construct a 95 % confidence interval estimate of the percentage of pages that you would expect to have advertising for the entire magazine.

Answer:
This is a proportion problem.

For this problem, we need to know the total number of pages in the magazine. These numbers will vary based on which magazine you select. Here is an example answer:

N = 86 (for the selected magazine)
n = 30
Number of pages in the sample with ads = 13 or 13/30 = .433 or 43 %
sd = we can use the short cut because either the pages have ads or they don't (zero, one model)



SD = SQRT (0.57 * 0.43) = .5

Use the formula SE for the proportion:
SE = (sd/sqrt n) × 100 %
SE = (.5/SQRT 30) × 100 % = (.5/5.47) × 100 % = 9.1 %
95 % CI = 2*9.1 = 18.2 % of the pages or in other words approximately 6 pages.

If we took more 30-page samples, we would still expect to find about 6 pages with advertising in 95 out of 100 samples.
As far as the population (all the pages), we would expect about 43 % would have ads, plus or minus 18 %.

2. In Florida, members of the citrus industry make extensive use of statistical methods. One way includes how growers are paid for oranges used to make orange juice. A sample is taken from each truckload of oranges. This sample is weighed and squeezed, and the amount of sugar in the juice is measured. Based on these sample results, an estimate is made of the total sugar in the truckload. Payment is based on the % of sugar. In a similar vein, how do customs officers decide how many boxes to inspect in a company's shipments at the docks? How about the security checks at airports? Any concerns?

Answer:

In all three situations, it is only reasonable to collect samples. The results would be more accurate if we looked at every part of the population but that would be too time intensive.

Based on the sample the farmers may be underpaid, but on the other hand, they may be overpaid. So everyone would like a high Confidence Interval (CI) to ensure the results are as accurate as possible without sampling the entire shipment. The CI and the standard error would be part of what gets negotiated between the farmers and the buyers. But, you can imagine these statistics are difficult for anyone who is not up to speed with their statistics to understand.

Consider, the truck has 1,000 oranges, and we only sample 100 with an average sugar content of 8 g. A sugar value of 12 g per medium size orange is more desired, so the farmers get paid less than the premium price. The sample size corresponds to only a 68 % CI. This would mean if we took another 100 different samples from this truck we would expect the same sugar content in only 68, but in 32 of the samples we might have a higher or lower sugar content.

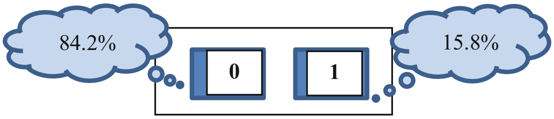
Similarly customs officers and TSA officials cannot examine every box and every person in great detail, so there is a trade off in risk versus cost. The lowest level of SE risk occurs when you survey the entire population (CI = 100 %).

3. A real estate company has been provided with a portfolio of townhomes and single family homes at Tahoe Keys in California. The company surveys a nearby affluent town of 25,000 households to determine which form of residence is preferred as a vacation home, so they can target their marketing accordingly. The company took a sample of 500 of these households. It turns out only 79 households preferred single family residences as a vacation home over townhomes.

   (a) Calculate the proportion of households in the town who prefer townhomes as a vacation home.
   (b) About how far off to you expect this estimate to be off?
   (c) Calculate the 99 % confidence interval for the expected value.

   Answer:

   N = 25,000
   n = 500
   CI = 99 %
   Number preferring single family residents = 79 or 79/500 = .158 or 15.8 %
   sd = we can use the short cut because either they like townhomes(0) or single family residents (1)



   SD = sqrt (.84 * .16) = .37
   Use the formula SE for the proportion
   SE = (sd/sqrt (n)) × 100 %
   SE = (.37/[SQRT 500]) × 100 %) = (.37/22.36) × 100 % = 1.65 %
   99 % CI = 3 × 1.65 = 4.95 % of the households prefer single family vacation homes or in other words about 25 households prefer single family vacation homes

   (A) 84.2 % prefer townhomes
   (B) ±8.27 households which equates to 8.27/500 = .0165 or approximately 1.7 %
   (C) 99 % CI = 3 * 1.7 = 5.1 %

4. The HR Director of a large cooperation wants to study absenteeism among administrative workers at the corporate office during the year. A random sample of 25 administrative workers reveals that the average number of days absent was 12 days. Construct a 95 % confidence interval for the population average absences for the clerical workers during the year

   Answer:

   n = 25
   $\bar{x} = 12$
   CI = 95 %
   The average number of days that administrative workers were absent is 12 days.
   SD = sqrt (.48 * .52) = .495
   Use the formula SE for the average:
   SE = (sd/sqrt (n))
   SE = .495/(SQRT 25) = 0.099
   95 % CI = 2* 0.099 = .198 days of absenteeism
   Workers will be absent on average 12 days plus or minus 0.20 days.

Chapter 8
Probability

Key Concepts

Addition Rule, Binomial distribution, Complement rule, Combinations, Factorial function, Conditional probability, Dependence, Multiplication Rule, Mutual exclusivity, Permutations, and Probability, Probability density function, Random variable, With and Without replacement.

Discussion

This chapter will explore the world of probabilities. It is very important to remember that probability is an estimation tool describing expected outcomes. Probability statistics explain the fraction of time an event is expected to happen in the long run, if the situation is repeated over and over again. In this area of statistics we have specific rules for how to combine probabilities and we also have special symbols to show these processes. In some cases the historical data fits a theoretical distribution allowing us to answer questions about the likelihood of an event occurring.

It is important to differentiate between discrete and continuous probability distributions. In the discrete world you have a few distinct values which are often integers. Continuous distributions describe a random process with continuous valued outcomes. Excel tools for handling both of these distributions will be explained.

Probability is the chance or likelihood that an event will happen. It is defined as the number of ways an outcome can occur, divided by the total number of outcomes that are possible for discrete events. There are some basic rules associated with probability:

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_8,

1. Probabilities are between 0 % and 100 %.
2. The probability of an event happening is 100 % minus the chance that it does not happen. If the chance of rain is 30 %, then the chance of no rain is 70 %. This is often referred to as the **Complement Rule** because the event and its opposite (complement) add up to 100 %.
3. When you draw at random, all items have the same probability of being selected.
4. When drawing at random with replacement, the draws are independent. Without replacement, the draws are dependent.

Addition Rule: This rule is used to combine probabilities one **or** the other may occur. When at least one of the events happens we combine the events by addition. In mathematical terms we often refer to this as the **union (U)** of two events. To show the probability of the union of two events it is written as:

$$\mathrm{P(AUB) = P(A) + P(B)}$$

Mutual Exclusivity: Two events are **mutually exclusive** when the occurrence of one prevents the other event from occurring. This concept is important when we add probabilities.

$$\textbf{EventA} = \text{being over 21 year old} \qquad \textbf{EventB} = \text{being 21 year old or younger}$$

Events A and B are **mutually exclusive.** It is impossible for both to occur at the same time. If Event A is true, then Event B must be false.

Multiplication Rule: This rule is used to combine probabilities one **and** the other may occur. When both of the events happen we combine the events by multiplication. In mathematical terms we often refer to this as the **intersection** (∩) of two events. To show the probability of the intersection of two events it is written as:

$$\mathrm{P(A \cap B) = P(A) \times P(B)}$$

Dependence: Events are **dependent** when the occurrence of one event affects the probability of occurrence of the other. Two events are independent if the chances for the second event stay the same, no matter how the first one turns out. This concept is important when we multiply probabilities.

$$\textbf{EventA} = \text{being under 25 year old} \qquad \textbf{EventB} = \text{being a CEO}$$

Events A and B are dependent. The probability of being a CEO may be 15 %, but if Event A is true (the person is under 25 years old), that probability may drop to less than 1 %. Event A has affected the probability of Event B.

Dependence is commonly referred to as "**without replacement**", while independence is referred to as "**with replacement**". This analogy is based on reaching in a hypothetical box and selecting a card. After the first draw if the card is placed back in the box, then the second draw is independent of the first outcome because the second draw has been completed "with replacement". This is illustrated in Fig. 8.1.

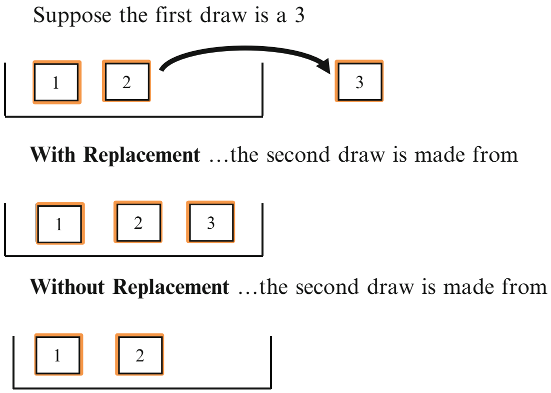

Suppose the first draw is a 3
1
2
3
With Replacement …the second draw is made from
1
2
3
Without Replacement …the second draw is made from
1
2

Fig. 8.1 The difference in drawing with and without replacement

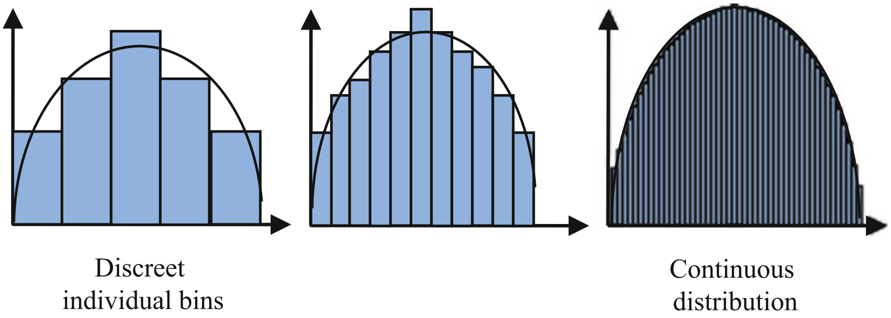


Fig. 8.2 Discreet and continuous distributions

Conditional Probability: P (A ǀ B) means the probability of A given that B has already happened. There is an implied temporal or sequential relationship: Event B has already occurred and that event is then followed by Event A.

Probability Distributions: Discrete probability distributions may be displayed with a bar chart in which the height of each bar is equal to the probability of the event as shown in Fig. 8.2. To find the probability of a group of events, we add up the individual probabilities of the events of the group.

Continuous probability distributions consider the area under the curve as the probability associated with a range of values. Continuous distributions are usually described by a **probability density function (pdf)**, where the area under the density function for a specified range of values is the probability.

Normal Distribution: One of the most common continuous probability distributions in business is the **Normal Distribution.** Excel includes functions for both the standard normal distribution (mean = 0 and standard deviation = 1) and a general normal distribution (mean other than 0 and/or standard deviation other than 1).

To obtain probability values for a standard normal distribution, the Excel function NORM.S.DIST is used. Otherwise, for the general normal distribution, the Excel function is named NORM.DIST (Note: there is not an **S** between the **M** and the **D**).

Random Variable: A **random variable** is a variable whose value is subject to variations due to chance. Conceptually it does not have a single, fixed value; it can take on a set of possible different values, each with an associated probability.

Combinations and Permutations: We can use combinations and permutations to assist in counting the number of possible outcomes which allows us to determine the probability of the event occurring. Sometimes it may be difficult to decide whether you should use a combination or a permutation. The following key words apply to situations involving combinations where order is not important: subsets, committees, work teams, work groups, and delegations. If sequencing or order is important, use permutations.

Combination: order of like items is not important	Permutation: order of like items is important
$C\binom{n}{x} = \frac{n!}{x!(n-x)!}$	$P\binom{n}{x} = \frac{n!}{(n-x)!}$

n = total number of items and x = number of items of a like kind to be chosen

The large brackets include the information that asks how many ways can we choose x items from n total items. In **combinations (C)** the order of the items is not important but in **permutations (P)**, order is important.

The **factorial function** (symbol: **!**) just means to multiply a series of sequentially descending whole numbers. Examples:

- **4!** $= 4 \times 3 \times 2 \times 1 = 24$
- **7!** $= 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 5{,}040$
- **1!** $= 1$
- **0!** $= 1$

Note: This is just one you have to remember: 0! will always equal 1.

Example 1

Consider the letters A, B, C. What is the probability of choosing the following selection **(A, B, C)** from this group without replacement? You must first calculate the total number of possible outcomes from this group when

Order is not important	Order is important
n = 3, x = 3	n = 3, x = 3
$C\binom{3}{3} = \frac{3!}{3!(3-3)!} = \frac{3!}{3!(0)!} = 1$	$P\binom{3}{3} = \frac{3!}{(3-3)!} = \frac{3*2*1}{(0)!} = 6$
ABC = BCA = CBA etc. so there is only one possible outcome, you will always have an A, B, and C no matter what	ABC, BCA, ACB, BAC, CAB, and CBA all meet the criteria. So there are six ways to get the desired selection

The next step is to calculate the actual probability. Remember probability is defined as the number of ways an outcome can occur, divided by the total number of outcomes. The number of ways an outcome can occur can be thought of as the number of ways of achieving **success** out of the total number of ways possible. We just calculated one possible outcome for the combination and six possible outcomes for the permutation.

Combination: There is only one pattern we are interested in **(A, B, C)** or there is only one way to be successful because *ABC* is the same as *BCA*, is the same as *CBA* so $x = 1$ and we just calculated the total number of outcomes as being only 1. The total number of ways of being successful divided by the total number of possible outcomes is $(1/1) = 1$ or 100 % probability you will be successful and select (A, B, C).

Permutation: There is only one pattern we are interested in **(A, B, C)** or there is only one way to be successful so $x = 1$, and we just calculated the total number of outcomes as 6. The total number of ways of being successful divided by the total number of possible outcomes is (1/6). So, there is a 16.7 % probability that you will be successful in selecting (A, B, C), in that specific order.

Example 2

Consider choosing three employees from a group of five employees without replacement: (Tom, Sue, Bob, Ann, Mary)

What is the probability that we end up with **(Ann, Mary, Tom)**? First, we need to know how many outcomes are possible.

Order is not important		Order is important
$C\binom{5}{3} = \frac{5!}{3!2!} = \frac{5*4*3*2*1}{(3*2*1)(2*1)} = 10$		Because we are filling a slate of positions, where team leader is the first slot, facilitator the second slot, and scribe the third slot. We want to know how many different slates of candidates are possible
Tom, Sue, Bob = Bob, Sue, Tom		$P\binom{5}{3} = \frac{5!}{2!} = \frac{5*4*3*2*1}{2*1} = 60$
It is the same three people, so we don't want to count this more than once. Here are the possible outcomes		Because:
Tom, Bob, Ann	Tom, Sue, Mary	Tom, Sue, Bob is a different slate from Sue, Tom, Bob etc.
Tom, Bob, Sue	Bob, Ann, Sue	
Tom, Bob, Mary	Bob, Ann, Mary	
Tom, Ann, Sue	Bob, Sue, Mary	
Tom, Ann, Mary	Ann, Sue, Mary	

The next step is to calculate the probability.

Combination: There is only one pattern we are interested in **(Ann, Mary, Tom)** because *Ann, Mary, Tom* is the same as *Tom, Ann, Mary*, etc., and we just calculated the total number of outcomes as being 10. The total number of ways of being successful divided by the total number of possible outcomes is (1/10) or a 10 % probability that you will be successful in selecting (Ann, Mary, Tom) in any order.

Permutation: There is only one pattern we are interested in **(Ann, Mary, Tom)** because we want Ann as the team leader, Mary as the facilitator, and Tom as the scribe. We just calculated the total number of outcomes as 60. The total number of ways of being successful divided by the total number of possible outcomes is (1/60) or a 1.6 % probability that you will be successful in selecting Ann as the team leader, Mary as the facilitator, and Tom as the scribe.

Binomial Distribution: This approach to calculating probabilities should be used when there are only two possible outcomes: success and failure.

Use the binomial distribution if and only if the situation satisfies the following conditions

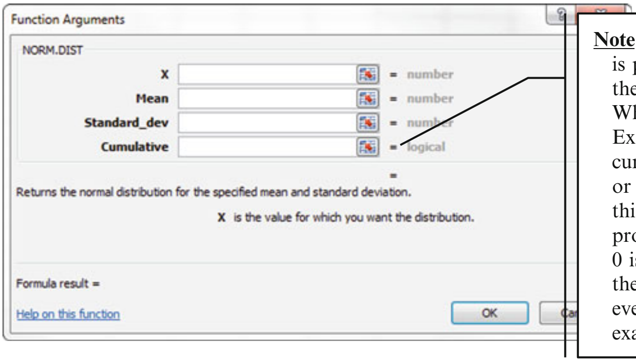
1. There is a fixed number of trials.
2. Each trial is independent of one another.
3. There are only two possible outcomes (a Success or a Failure) for each trial.
4. The probability of success, p, is the same for every trial.

An example of an experiment that has a binomial distribution would be a coin toss.

1. You would toss the coin n (a fixed number) times.
2. The result of a previous toss does not affect the present toss (trials are independent).
3. There are only two outcomes (Heads or Tails).
4. The probability of success (whether a head is considered a success or a tail is considered a success) is constant at 50 %.

Excel

NORM.DIST is the function that returns the cumulative left-tail probability that the normal random variable is less than or equal to x.



Returns the normal distribution for the specified mean and standard deviation.

X is the value for which you want the distribution.

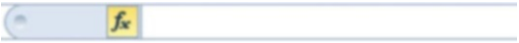
Note: When a Cumulative box is part of the required input there are two possible inputs. When TRUE or 1 is selected, Excel calculates the cumulative probability of "X or fewer" events of interest; this is a cumulative probability. When FALSE or 0 is selected, Excel calculates the probability of "exactly X" events of interest as in this example.

Finding Probabilities Using Normal Distributions

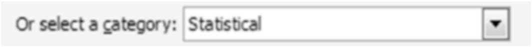
In this example, we have data on the number of cars sold each week at local dealerships over the past 6 months. Consider a normal random variable x (number of cars sold per week) with a mean (μ) of 100 cars sold per week and standard deviation (σ) of 15 cars per week.

What Is the Probability That a Dealership Will Sell 90 Cars or Less ($x \leq 90$) per Week?

➢ Highlight the cell where you want the answer to go
➢ Click on the ***fx*** button by the function input bar

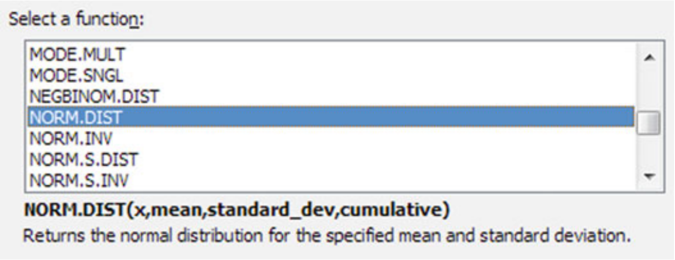


➢ Select **Statistical** under **Category**



Or select a category: Statistical

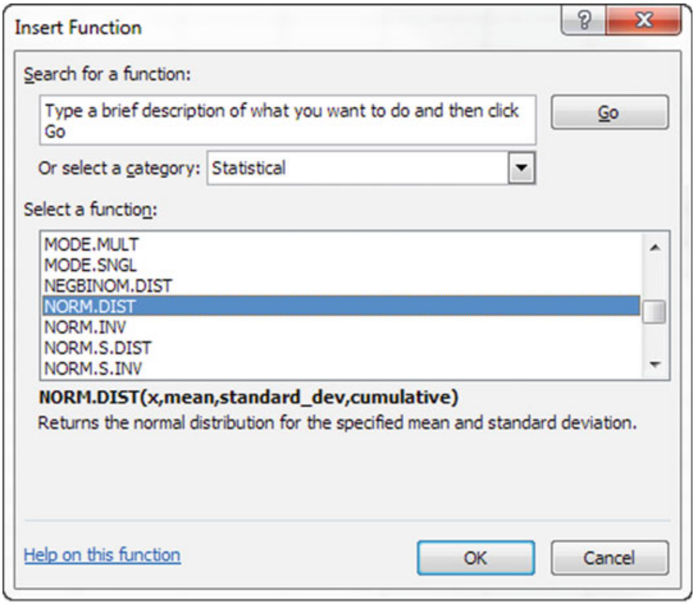
➢ Select **NORM.DIST** under **Function**



NORM.DIST(x,mean,standard_dev,cumulative)

Returns the normal distribution for the specified mean and standard deviation.

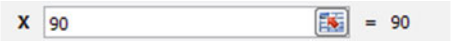
➢ Click **OK**



NORM.DIST(x,mean,standard_dev,cumulative)

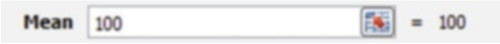
Returns the normal distribution for the specified mean and standard deviation.

➢ Input **90** for **X**



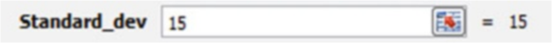
X 90 = 90

➢ Input **100** for the **Mean**



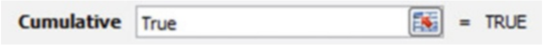
Mean 100 = 100

- ➢ Input **15** for the **Standard_dev**



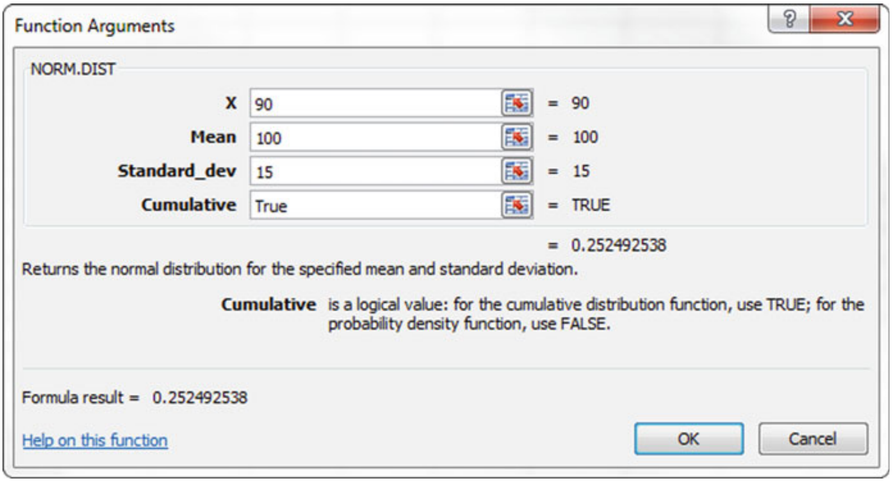
Standard_dev 15 = 15

- ➢ Input a value for **Cumulative** of **1** or **True**



Cumulative True = TRUE

- ➢ Click **OK**



Function Arguments

NORM.DIST

X	90	= 90
Mean	100	= 100
Standard_dev	15	= 15
Cumulative	True	= TRUE

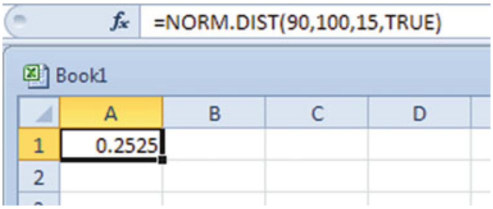
= 0.252492538

Returns the normal distribution for the specified mean and standard deviation.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability density function, use FALSE.

Formula result = 0.252492538

Excel provides an answer of 0.2525 which can be converted to 25.25 % probability that a dealership will sell 90 or less cars per week.



fx =NORM.DIST(90,100,15,TRUE)

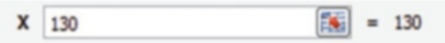
Book1

	A	B	C	D
1	0.2525			
2				

What Is the Probability That a Car Dealership Will Sell at least 130 (x ≥ 130) Cars per Week?

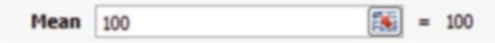
- ➢ Highlight the cell where you want the answer to go
- ➢ Click on the ***fx*** button by the function input bar

- ➢ Select **Statistical** under **Category**
- ➢ Select **NORM.DIST** under **Function**
- ➢ Click **OK**
- ➢ Input **130** for **X**



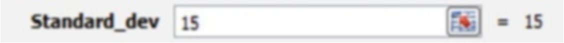
X 130 = 130

- ➢ Input **100** for the **Mean**



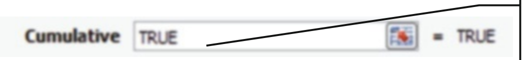
Mean 100 = 100

- ➢ Input **15** for the **Standard_dev**



Standard_dev 15 = 15

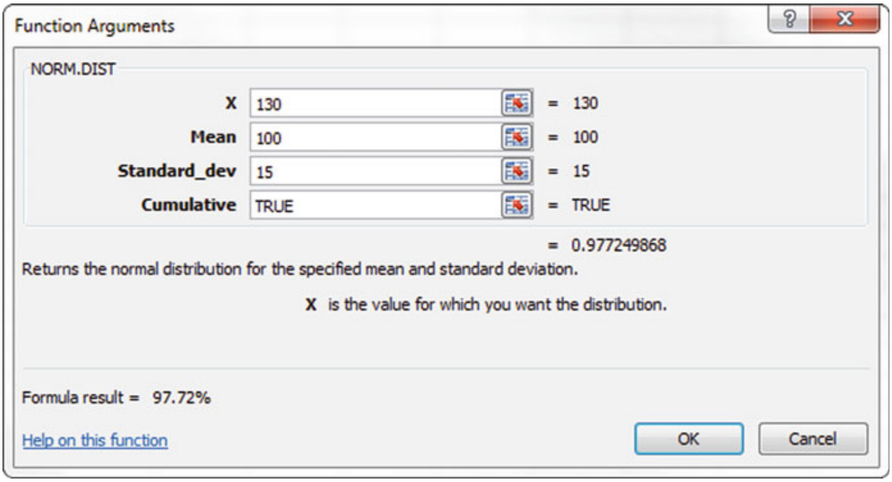
- ➢ Input a value for **Cumulative** of **1** or **True**



Cumulative TRUE = TRUE

Note: The NORM.DIST function with True (1) provides the probability the car dealership sells less than 130 cars.

- ➢ Click **OK**



Function Arguments

NORM.DIST

X 130 = 130

Mean 100 = 100

Standard_dev 15 = 15

Cumulative TRUE = TRUE

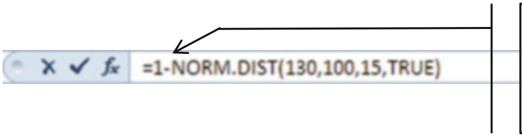
= 0.977249868

Returns the normal distribution for the specified mean and standard deviation.

X is the value for which you want the distribution.

Formula result = 97.72%

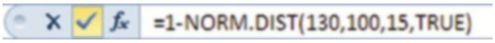
- ➢ Double click on the cell to edit the formula
- ➢ Subtract this amount from 1 in the function input box



=1-NORM.DIST(130,100,15,TRUE)

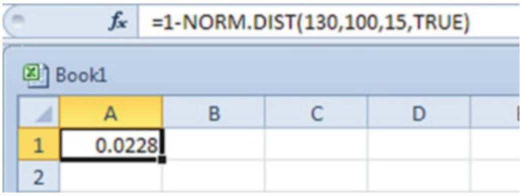
Note: To calculate the probability of at least 130 cars, you must subtract this number from 1 (as two probabilities must sum to 1).

- ➢ Click the green check



=1-NORM.DIST(130,100,15,TRUE)

Excel provides an answer of 0.0228 (2.28 %) or there is a probability of about 2.3 % that dealerships will sell 130 or more cars per week.



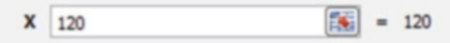
=1-NORM.DIST(130,100,15,TRUE)

Book1

	A	B	C	D
1	0.0228			
2				

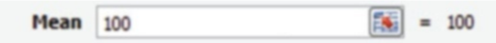
What Is the Probability That a Car Dealership Will Sell Between 90 and 120 Cars per Week?

- ➢ Highlight the cell where you want the answer to go
- ➢ Click on the ***fx*** button by the function input bar
- ➢ Select **Statistical** under **Category**
- ➢ Select **NORM.DIST** under **Function**
- ➢ Input **120** for **X**



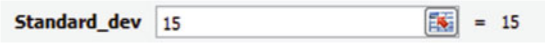
X 120 = 120

- ➢ Input **100** for the **Mean**



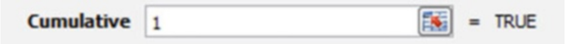
Mean 100 = 100

- ➢ Input **15** for the **Standard_dev**



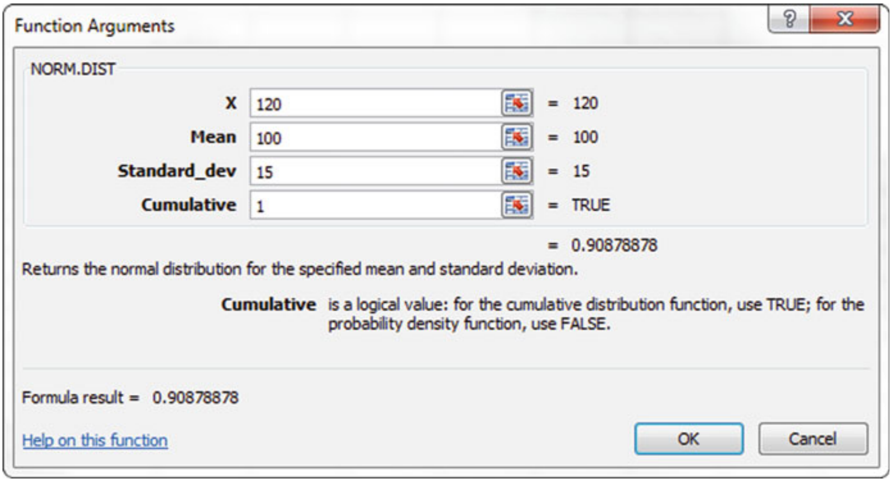
Standard_dev 15 = 15

- ➢ Input **1** or **True** for **Cumulative**



Cumulative 1 = TRUE

- ➢ Click **OK**



Function Arguments

NORM.DIST

X	120	= 120
Mean	100	= 100
Standard_dev	15	= 15
Cumulative	1	= TRUE

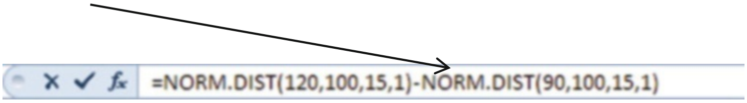
= 0.90878878

Returns the normal distribution for the specified mean and standard deviation.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability density function, use FALSE.

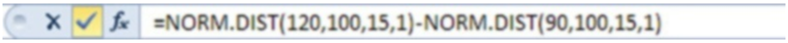
Formula result = 0.90878878

- ➢ Double click the cell to edit
- ➢ Inside the formula bar input you will need to type in the rest of the equation **- NORM.DIST (90,100,15,1)**



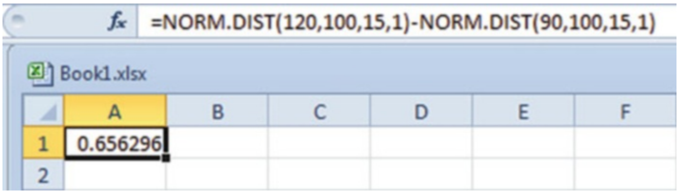
=NORM.DIST(120,100,15,1)-NORM.DIST(90,100,15,1)

- ➢ Click the green check



=NORM.DIST(120,100,15,1)-NORM.DIST(90,100,15,1)

Excel provides an answer of 0.656296 or there is a probability of about 65 % that the dealership will sell between 90 and 120 cars per week.



=NORM.DIST(120,100,15,1)-NORM.DIST(90,100,15,1)

Book1.xlsx

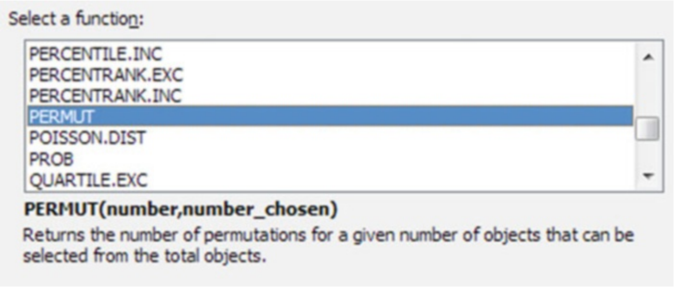
	A	B	C	D	E	F
1	0.656296					
2						

Calculating Combinations and Permutations

The main technical problems in your company involve laptops, monitors, speaker setups, software issues, and internet connectivity. It is important to forecast utilization of your IT professionals and understand the probability of the type of issues they might deal with on a day to day basis. On a typical day they can handle four of these problem areas.

Permutation

- **Order is important** as the IT staff work shift work and not all employees on each shift have the same expertise. What is the probability that the daily roster of problems will be **(Connectivity, Software, Laptops, Monitors)**? In this case x = 4 and n = 6.
  - ➢ Highlight the cell where you want the answer to go
  - ➢ Click on the ***fx*** button by the function input bar
  - ➢ Select **Statistical** under **Category**
  - ➢ Select **PERMUT** under **Select a function:**



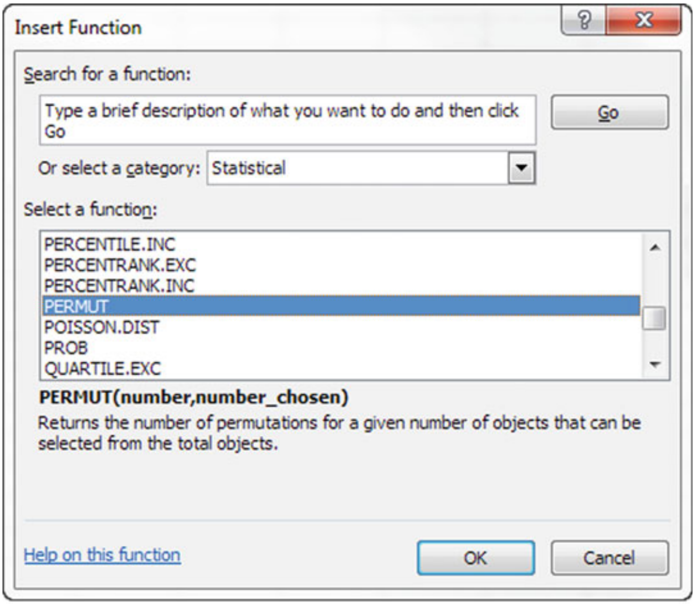
Select a function:

PERCENTILE.INC
PERCENTRANK.EXC
PERCENTRANK.INC
PERMUT
POISSON.DIST
PROB
QUARTILE.EXC

PERMUT(number,number_chosen)

Returns the number of permutations for a given number of objects that can be selected from the total objects.

- ➢ Click **OK**



Search for a function: Type a brief description of what you want to do and then click Go

Or select a category: Statistical

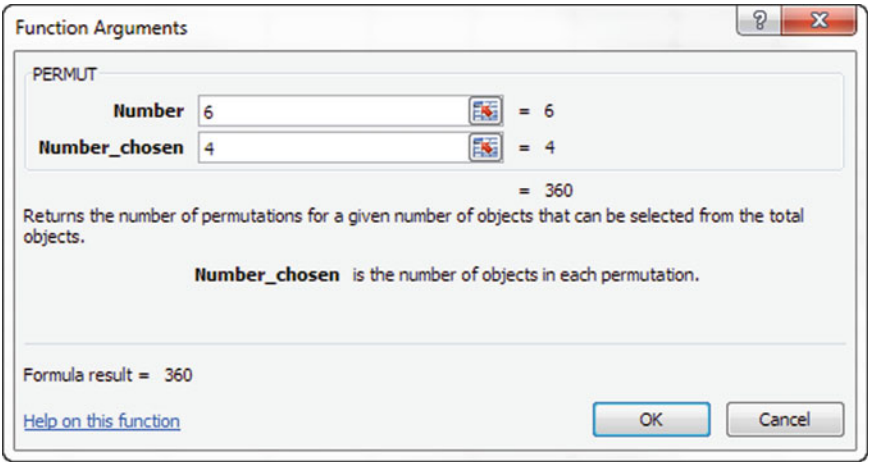
Select a function:

PERCENTILE.INC
PERCENTRANK.EXC
PERCENTRANK.INC
PERMUT
POISSON.DIST
PROB
QUARTILE.EXC

PERMUT(number,number_chosen)

Returns the number of permutations for a given number of objects that can be selected from the total objects.

- ➢ Input the **6** for **Number** (n) in the first entry box followed by **4** for **Number_chosen** (x) in the second input box, then click **OK**



PERMUT

Number 6 = 6

Number_chosen 4 = 4

= 360

Returns the number of permutations for a given number of objects that can be selected from the total objects.

Number_chosen is the number of objects in each permutation.

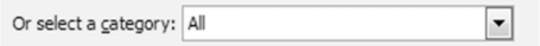
Formula result = 360

Excel will provide the answer that *360 permutations are possible*, or in other words, there are 360 possible outcomes.

The total number of ways of being successful divided by the total number of possible outcomes is (1/360) or a .3 % probability that you will be successful in having these four items in that order on your roster. There is a very low probability that your IT staff will be faced with connectivity, software, laptop, monitor problems in that specific order on any given day. You might have to look at a different kind of analysis to achieve more accurate forecasting results.

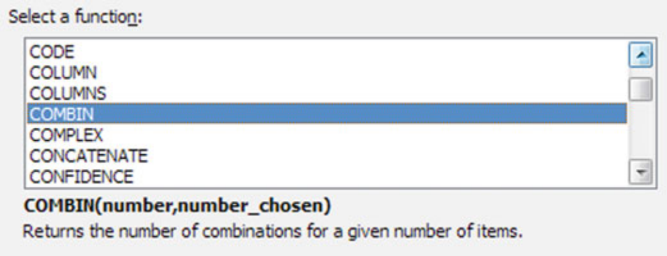
Combination

- **Order is not important.** What is the probability that the daily roster of problems will include these four areas **(Connectivity, Software, Laptops, Monitors)**? Remember in this case x = 4 and n = 6.
  - ➢ Highlight the cell where you want the answer to go
  - ➢ Click on the ***fx*** button by the function input bar
  - ➢ Select **All** under **Category**



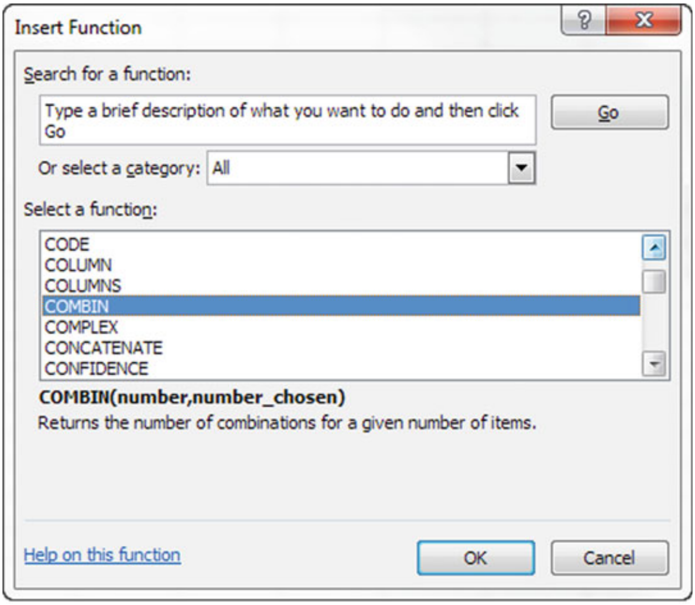
Or select a category: All

  - ➢ Select **COMBIN** under **Select a function:**



COMBIN(number,number_chosen)
Returns the number of combinations for a given number of items.

➢ Click **OK**



Insert Function

Search for a function:

Type a brief description of what you want to do and then click Go

Or select a category: All

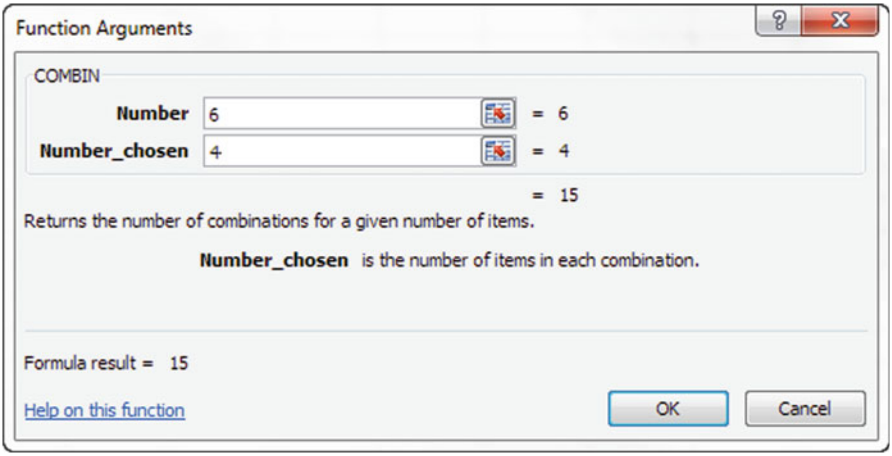
Select a function:

CODE
COLUMN
COLUMNS
COMBIN
COMPLEX
CONCATENATE
CONFIDENCE

COMBIN(number,number_chosen)

Returns the number of combinations for a given number of items.

➢ Input the **6** for **Number** (n) in the first entry box followed by **4** for **Number_chosen** (x) in the second input box, then click **OK**



Function Arguments

COMBIN

Number 6 = 6

Number_chosen 4 = 4

= 15

Returns the number of combinations for a given number of items.

Number_chosen is the number of items in each combination.

Formula result = 15

Excel will provide the answer that *15 combinations are possible*, or in other words there are 15 possible outcomes.

The total number of ways of being successful divided by the total number of possible outcomes is (1/15) or a 6.7 % probability that these IT issues will occur. There is a about a 7 % chance that your IT staff will have to deal with connectivity, software, laptop, and monitor problems on any given day.

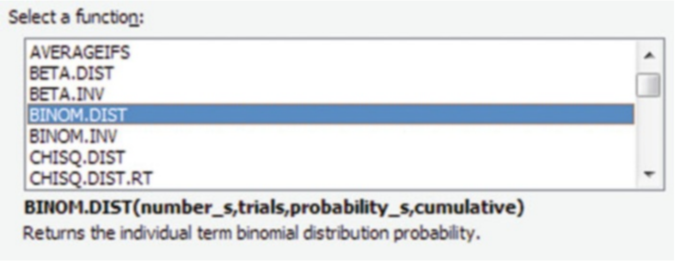
Finding Probabilities Using the Binomial Distribution

Royal Bank Retention Problem

The Royal Bank is concerned about a low retention rate for employees. On the basis of past experience, management has seen an annual employee turnover of 10 %. For any employee chosen at random, management estimates a probability of 0.1 that the person will not be with the company next year.

Choosing four hourly employees at random, what is the probability that 0 of them will leave the company next year, everyone stays? For this problem: Number of employees that leave is the number of successes ($x = 0$), the number of employees in the group is the number of trials ($n = 4$), and the probability of turnover is the probability ($p = 10$ %).

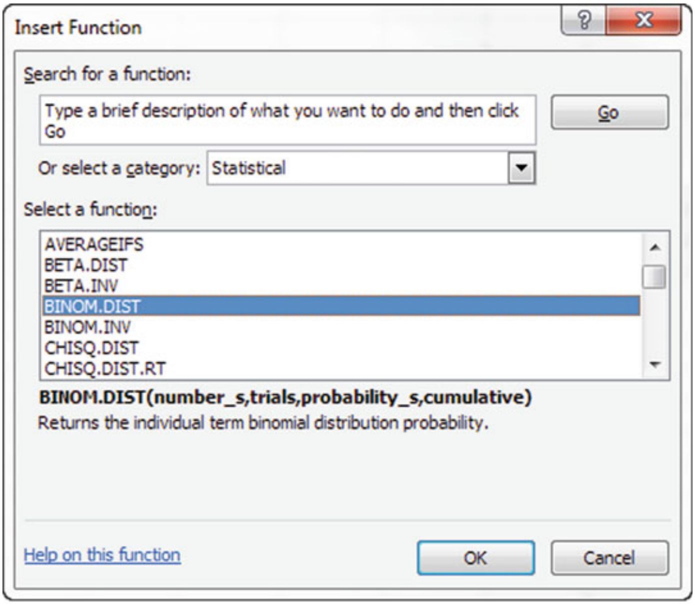
- ➢ Highlight the cell where you want the answer to go
- ➢ Click on the ***fx*** button by the function input bar
- ➢ Select **Statistical** under **Category**
- ➢ Select **BINOM.DIST** under **Select a function:**



BINOM.DIST(number_s,trials,probability_s,cumulative)

Returns the individual term binomial distribution probability.

- ➢ Click **OK**



Insert Function

Search for a function:

Type a brief description of what you want to do and then click Go

Or select a category: Statistical

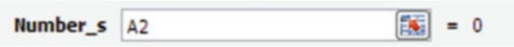
Select a function:

AVERAGEIFS
BETA.DIST
BETA.INV
BINOM.DIST
BINOM.INV
CHISQ.DIST
CHISQ.DIST.RT

BINOM.DIST(number_s,trials,probability_s,cumulative)

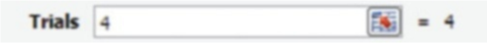
Returns the individual term binomial distribution probability.

- ➢ Input the number of successes in **Number_s**; in this case, the number of ways of being successful (x) is 0, but we have typed it into cell A2, so we can input A2



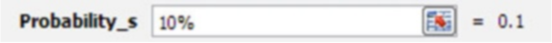
Number_s A2 = 0

- ➢ Input the number of trials in the **Trials**; in this case, the number of trials (n) is 4



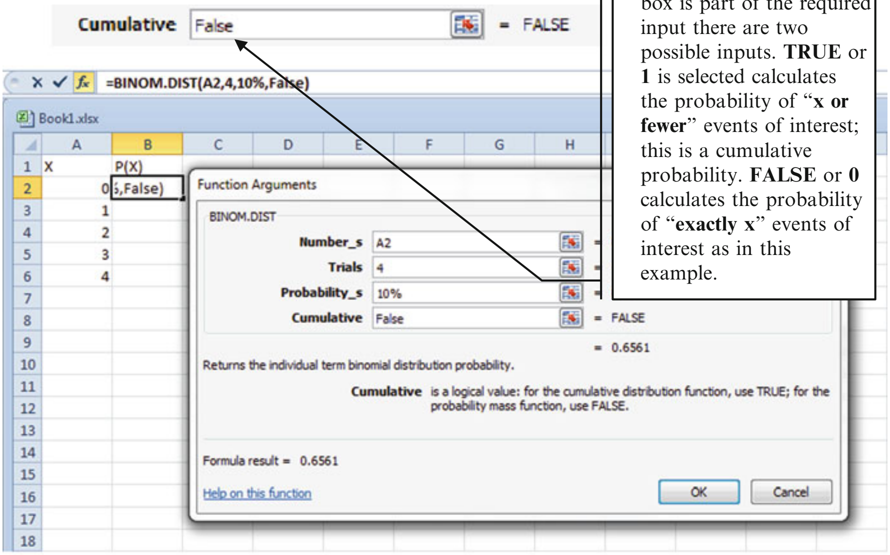
Trials 4 = 4

- ➢ Input the probability in **Probability_s**; in this case, the percentage of turnover (p) is 10 %



Probability_s 10% = 0.1

➢ Type in **FALSE** or **0** (exact probability)



Cumulative: False = FALSE

=BINOM.DIST(A2,4,10%,False)

Book1.xlsx

	A	B
1	X	P(X)
2	0	,False)
3	1	
4	2	
5	3	
6	4	

Function Arguments

BINOM.DIST

Number_s: A2

Trials: 4

Probability_s: 10%

Cumulative: False = FALSE

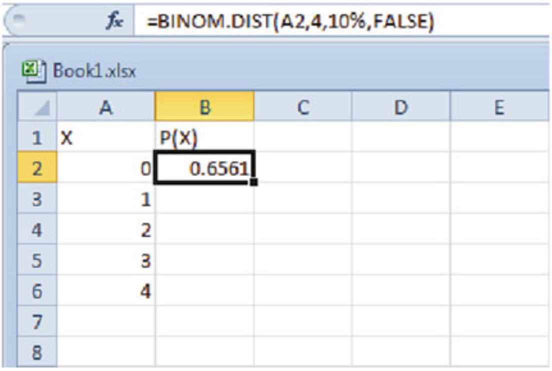
= 0.6561

Returns the individual term binomial distribution probability.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

Formula result = 0.6561

Note: When a **Cumulative** box is part of the required input there are two possible inputs. **TRUE** or **1** is selected calculates the probability of "**x or fewer**" events of interest; this is a cumulative probability. **FALSE** or **0** calculates the probability of "**exactly x**" events of interest as in this example.



=BINOM.DIST(A2,4,10%,FALSE)

Book1.xlsx

	A	B
1	X	P(X)
2	0	0.6561
3	1	
4	2	
5	3	
6	4	

Excel will provide an answer of .6561 or about a 65.6 % chance that zero out of four employees will leave the company next year. In other words, there is about a 66 % chance that everyone will stay or conversely a 34 % that someone will leave.

Common Excel Pitfalls

- ◇ Do not use probabilities that exceed 100 %.
- ◇ Be careful if you need to take into account the order of elements in counting the number of possible outcomes. If so then use a permutation not a combination.

- In earlier versions of Excel there may be a period in the statistical function name so make sure you include this to get the right function.
- If events are mutually exclusive then by definition they are dependent on one another.
- If the distribution function requires a **cumulative** box to be filled in, remember that **False (0)** calculates the probability of **exactly x** events and **True (1)** is used for **x or fewer** events.
- Do not forget to include the "%" sign when calculating Probability_s for the BINOM.DIST, otherwise, make sure to convert the percentage into decimal format.

Final Thoughts and Activities

Practice Problems

1. From four Registered Nurses (RN) and three Licensed Vocational Nurses (LVN), find the number of committees that can be formed consisting of two RN's and one LVN, if order is irrelevant.
2. How many ways can five open bank teller positions be filled with nine qualified candidates? Note that order is important because each of the teller positions has a different pay level and different title.
3. Suppose three friends decide to go to Hamburger Hut. In the last month 86.1 % of the orders were filled correctly. What is the probability that:
   (a) All three of the orders of the three friends will be filled correctly?
   (b) None of the three will be filled correctly?
   (c) At least two of the three will be filled correctly?

Discussion Boards

1. Reliability of systems can often be improved with redundancy of critical components. Airplanes have two independent electrical systems, and typically two separate radios. If one component has a 0.001 probability of failure, the probability of the two independent components both failing is only 0.000001 using the multiplication rule. Discuss where else redundancy like this is important.
2. In a Chicago study involving the use of olestra as a substitute for fat in potato chips, Proctor and Gamble researchers reported that on a nine point taste scale, olestra chips got a mean of 5.6 where regular chips got a mean of 6.4. When Proctor and Gamble were asked to provide the raw data they refused. Is this ethical?

Group Activity

1. What role does probability play in sabermetrics? Research this “Billy Bean” approach in professional baseball.
2. In the movie Twenty Four several MIT students worked with their professor in developing a method of counting cards in Las Vegas to change their odds of winning in the game of blackjack. In the end research why the system failed and the impact it had on the gaming industry.
3. We use probability to influence our daily activities and decisions. For example we may use several variables in determining which checkout lane will have the highest probability of being the quickest at our local grocery store. What are some other daily activities in which you use basic probability in making decisions?

Parting Thought

Did you hear the one about the statistician? Probably…

Problem Solutions

1. From four Registered Nurses (RN) and three Licensed Vocational Nurses (LVN), find the number of committees that can be formed consisting of two RN’s and one LVN, if order is irrelevant.

   In this case for the RN, x = 4 and n = 2. For the LVN, x = 3 and n = 1. Order doesn’t matter. We multiply because the committee is for RNs **AND** LVN.

$$\begin{aligned}\text{In Excel, use equation: } &= \text{COMBIN}(4,2) * \text{COMBIN}(3,1)\\ &= 6 * 3\\ &= 18\end{aligned}$$

$$\begin{aligned}\text{Mathematically} &= \left(\frac{4!}{2!(4-2)!}\right) * \left(\frac{3!}{1!(3-1)!}\right) = \left(\frac{4!}{2! * 2!}\right) * \left(\frac{3!}{1! * 2!}\right)\\ &= \left(\frac{4*3*2*1}{2*1*2*1}\right) * \left(\frac{3*2*1}{1*2*1}\right) = \left(\frac{24}{4}\right) * \left(\frac{6}{2}\right) = 6 * 3 = 18\end{aligned}$$

2. How many ways can five open bank teller positions be filled with the nine qualified candidates?

   Answer:
   Order **does** matter.

$$\begin{aligned}\text{In Excel, use equation: } &= \mathrm{PERMUT}(9, 5)\\ &= 15, 120\end{aligned}$$

$$= \left(\frac{9!}{(9-5)!}\right) = \left(\frac{9!}{4!}\right) = \left(\frac{9*8*7*6*5*4*3*2*1}{4*3*2*1}\right) = \left(\frac{362, 880}{24}\right)$$

$$= 15, 120$$

3. Suppose three (3) friends decide to go to Hamburger Hut. In the last month 86.1 % of the orders were filled correctly. What is the probability that:

   (a) All three orders will be filled correctly?

   Use Equation: = BINOM.DIST(3,3,86.1 %,FALSE)
   Answer: 63.83 %

   (b) None of the three will be filled correctly?

   Use Equation: = BINOM.DIST(0,3,86.1 %,FALSE)
   Answer: 0.27 %

   (c) At least two of the three will be filled correctly?

   Use Equation: = BINOM.DIST(2,3,86.1 %,FALSE) + BINOM.DIST (3,3,86.1 %,FALSE)
   Answer: 94.74 %

You add the two probabilities because the question asks for the probability of at least two of the three will be filled correctly; you add the chance of having two correct and the chance of having three correct.

Chapter 9
Correlation

Key Concepts

Association, Correlation coefficient, Correlation matrix, Causation, Direct relationship, Inverse relationship, Pearson correlation coefficient, and Positive relationship.

Discussion

This chapter explores the linear relationship between variables. The correlation statistics quantify the magnitude and direction of the relationship. Correlational analysis can be applied between x's as well as between the x's and y. This chapter will describe several ways to generate these statistics in Excel.

Correlation: Correlation is a powerful measure of the **association** between two variables. This is not the same as causation. **Causation** infers that one variable causes an effect in the other. A change in one variable creates a change in the other.

Statistic: The **Pearson correlation coefficient (r)** is commonly used to measure the degree of correlation between data sets.

This coefficient can range from +1 to −1. The correlation coefficient provides two important pieces of information: the strength of the relationship between variables and the type of relationship between the data sets. If both data sets increase or both decrease together we refer to that as a **direct relationship**, and the r statistic will be **positive**. If as one data set values increase, while the other data set values decrease, we refer to this as an **inverse relationship**, and the r statistic will be **negative**.

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_9,

Positive r → x and y change in the same direction
ex. The higher employee satisfaction, the higher productivity.
Negative r → x and y change in opposite directions
ex. The more children I have the fewer expensive vacations I will take.

Figure 9.1 provides some graphical examples. The lower left graph shows no association between the x data set and the y data set; as x increases, y shows no tendency to increase or decrease. Therefore the r value is zero with no sign. The lower right graph shows a slight linear pattern to the data sloping upwards to the right. This slight association between the variables results in an r value of .4. As the variable on the x axis increases, the variable on the y axis also increases resulting in a positive sign for the correlation coefficient. The two upper graphs show strong linear patterns with an r of .9. When data slopes downward to the right note the sign of r becomes negative (upper right graph). In other words as

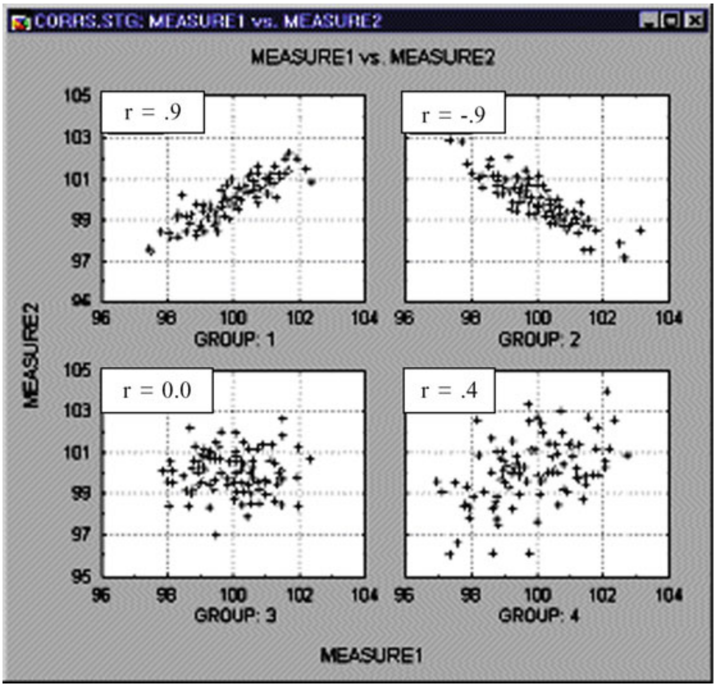
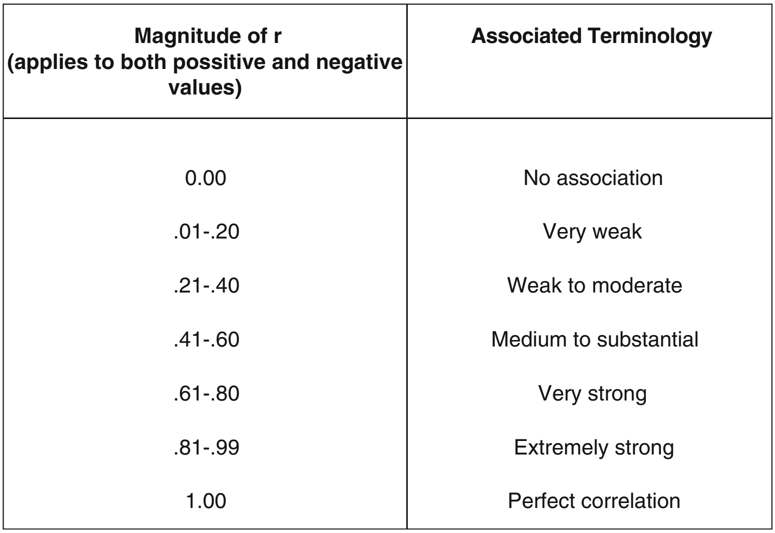


Fig. 9.1 Graphical examples of r

the x data increase, the y data decrease indicating an inverse relationship between the two data sets.

Weaker associations are common in social science data: for example, consider a university entrance application that associates test scores to college success. The behavioral relationships of people are not always strongly correlated, so we often see r values between 0.3 and 0.7. In physics and engineering we might seek higher r values. In designing the space shuttle we would want a stronger r value, perhaps as high as .99 in analyzing the relationship between heat transference and the amount of a particular alloy used in designing the outer shell of the shuttle. A "rule of thumb" translation guide is provided in Fig. 9.2.



Magnitude of r (applies to both possitive and negative values)	**Associated Terminology**
0.00	No association
.01-.20	Very weak
.21-.40	Weak to moderate
.41-.60	Medium to substantial
.61-.80	Very strong
.81-.99	Extremely strong
1.00	Perfect correlation

Fig. 9.2 General terminology in the magnitude of r

Note that r = 0.40 does not mean that 40 % of the data is tightly clustered around a line. It also does not mean that the data display twice as much linearity as data with an r = 0.20.

◈ The correlation **co-efficient** can only measure linear associations.

Because r is a pure number, meaning it is computed using standard units, it is not affected by a change in scale of the measured units. This means r is not affected when you multiply each value of one variable by the same number. It is not affected if you add the same number to all values of one variable. Finally it is not affected by interchanging the two variables (x ↔ y). These properties are shown in Figs. 9.3 and 9.4.

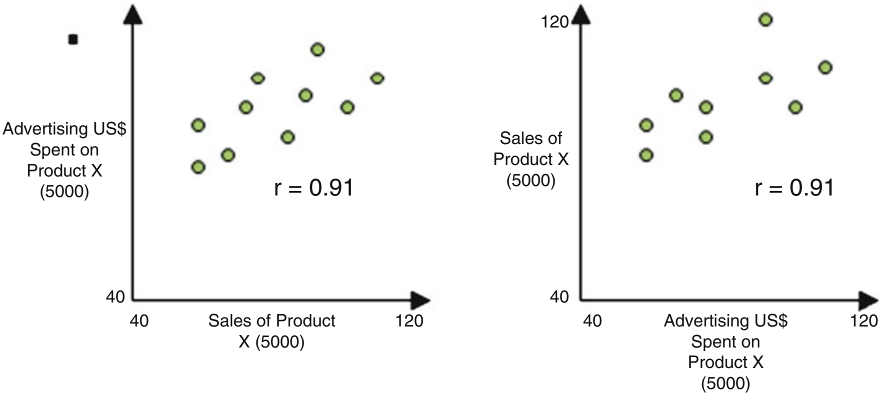


Fig. 9.3 Interchanging the variables (x ↔ y)

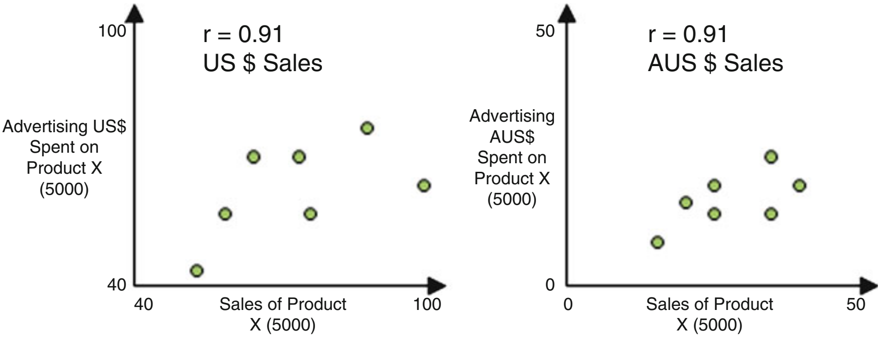


Fig. 9.4 Adding and multiplying the same number(s) to all variables. Converting from U.S $ to Australian $ with 1 month's worth of recorded daily average shows the same correlation coefficient r = .91

The correlation coefficient is sensitive to **outliers**. Remember the general definition of outliers is data that exceeds ±3 standard deviations. When outliers are present the r value can be misleading. Consider Fig. 9.5.

Since outliers tend to reduce the correlation coefficient, careful consideration should be given in determining whether they remain in the calculation of r.

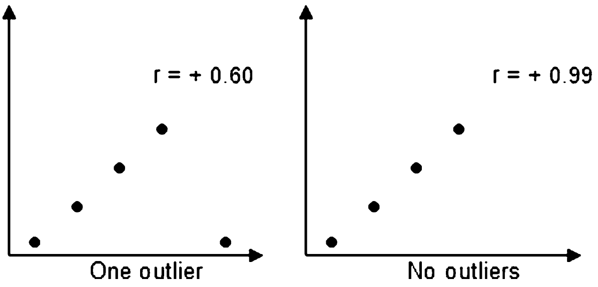


Fig. 9.5 The effect of outliers on r

Nonlinear data caution

There are several cautions in working with correlation co-efficients. The first is that the correlation coefficient can only measure "linear" associations. Consider the nonlinear data presented in Fig. 9.6.

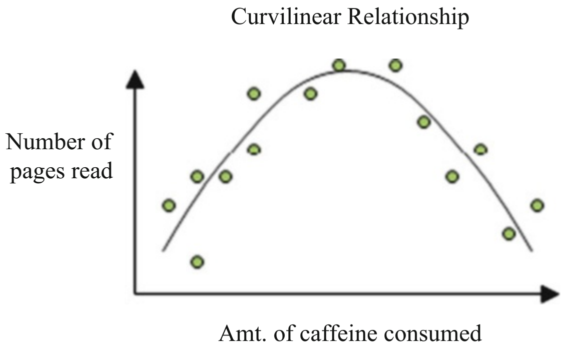


Fig. 9.6 Nonlinear data: the relationship between the amount of caffeine consumed and number of pages read

Figure 9.6 suggests there is a strong positive relationship between the amount of caffeine consumed and how many pages someone can read up to a certain point. After that point, no matter how much more caffeine you consume, there is little to no change in the number of pages you can read. Then you reach a certain level of caffeine consumption that starts reducing the number of pages you can read. (Perhaps you are too jittery to hold the book still.)

Using a single value of r cannot clearly define this nonlinear relationship. In this case you could calculate three separate r values in the three different (x, y) relationships; this is shown in Fig. 9.7.

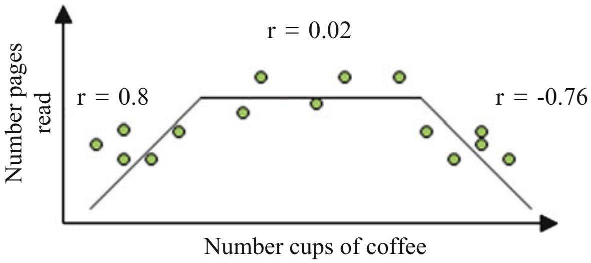


Fig. 9.7 Multiple r values for multiple relationships

Average data caution

The second issue involves working with average data. Aggregated or averaged data is data that has been collected at the individual level and combined in some way into group data. Correlations based on averages can be misleading.

In individual level data, there may be a lot of spread around the averages. If you average that data into group data, the spreads are reduced and you generate an inflated r value. Consider the example in Fig. 9.8.

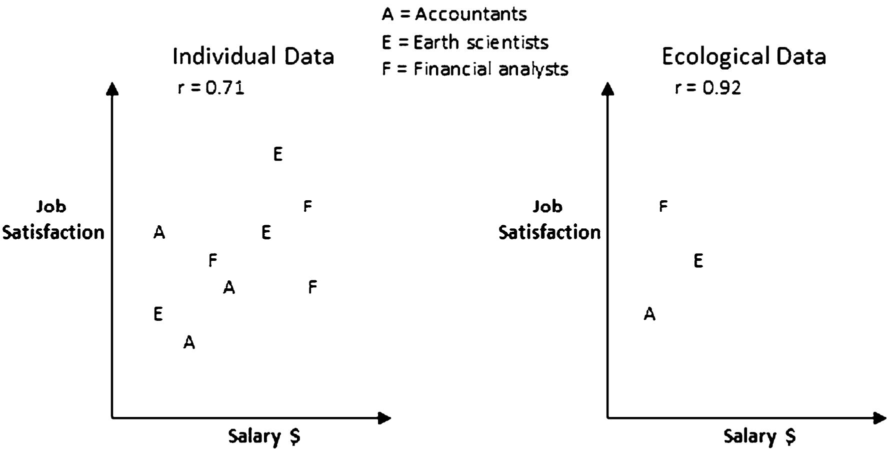


Fig. 9.8 Correlations of averaged or ecological job satisfaction data produces inflated r values

In Fig. 9.8 the panel on the left indicates the individual level data generated by measuring annual salary and job satisfaction among accountants, earth scientists and financial analysts within one company. The resulting co-efficient is r = 0.71.

The second panel shows the average salary and job satisfaction values for each of the three professional groups. By averaging the data we have reduced the spread and produced a stronger linear association. The r value in this case has increased to 0.92. It is less accurate to use the correlation co-efficient based on average data to predict relationships at the individual level. These correlations tend to overstate the strength of association, so use caution.

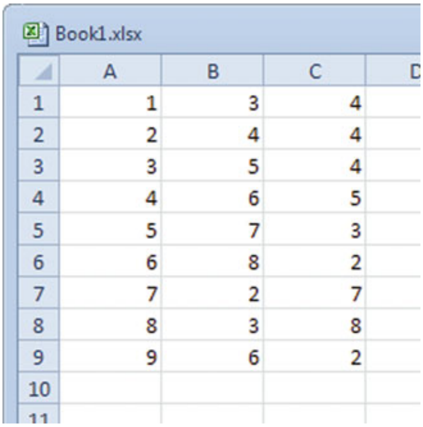
Excel

Correlation: One r Value or Correlation Matrix

There are two common ways to access the correlation functionality inside of Excel. They are both equally easy and vary by how many data sets (columns of data) you want to correlate simultaneously.

Method 1: Two or More Data Sets (Matrix)

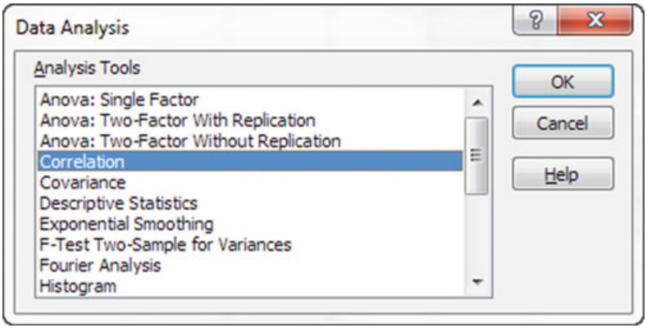
Enter the data below:



Book1.xlsx

	A	B	C	D
1	1	3	4	
2	2	4	4	
3	3	5	4	
4	4	6	5	
5	5	7	3	
6	6	8	2	
7	7	2	7	
8	8	3	8	
9	9	6	2	
10				
11				

- ➢ Select the **Data** tab
- ➢ Select the **Data Analysis** function
- ➢ Choose **Correlation** from the list of Analysis Tools



Data Analysis

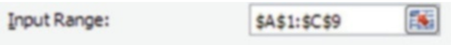
Analysis Tools

Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram

- ➢ Click **OK**

➢ Input the data you want to correlate in the **Input Range**

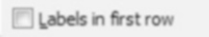
   ◇ You need a minimum of 2 columns of data but you can input as many columns as you are interested in correlating.



Input Range: $A$1:$C$9

➢ Click the box next to **Labels** in first row, if you have labels; however, in this example, no labels are used so we leave the box unchecked

   ◇ It is highly recommended you input labels to help keep track of the variables.

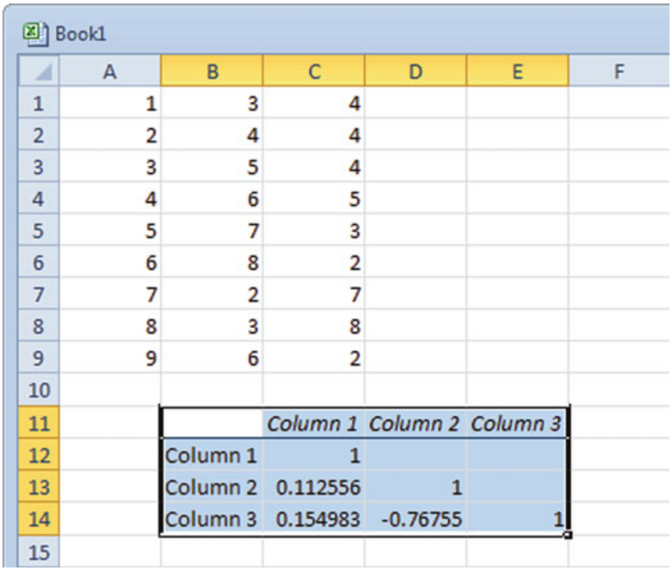


□ Labels in first row

➢ Click on the **Output Range** and input **Cell B11**

➢ Click **OK**

The output should appear as a matrix starting from cell B11 as shown in the screenshot below. The output matrix shows just the lower half as the upper half is a mirror image of the lower half. This is because the correlation between x and y is the same value as the correlation between y and x. The diagonal of 1's indicates the perfect correlation each variable has with itself. The diagonal divides the upper half correlation values from the lower half correlation values.

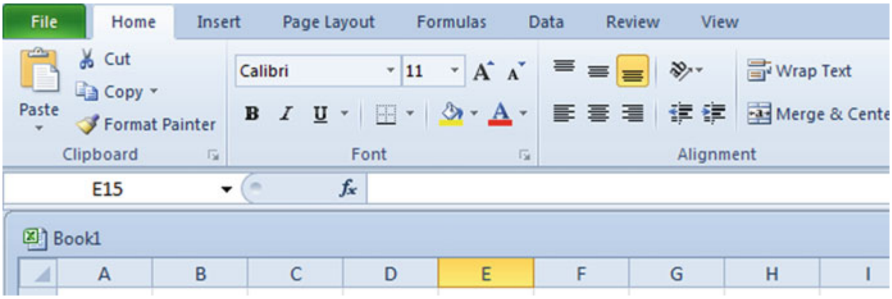


Book1

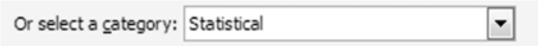
	A	B	C	D	E	F
1	1	3	4			
2	2	4	4			
3	3	5	4			
4	4	6	5			
5	5	7	3			
6	6	8	2			
7	7	2	7			
8	8	3	8			
9	9	6	2			
10						
11			*Column 1*	*Column 2*	*Column 3*	
12		Column 1	1			
13		Column 2	0.112556	1		
14		Column 3	0.154983	-0.76755	1	
15						

Method 2: Only 2 Data Sets

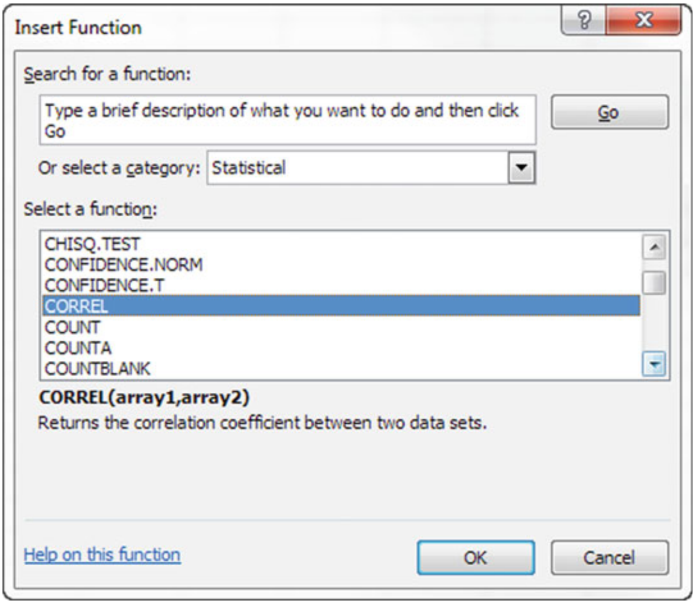
For this method you don’t even have to be within the data analysis tab, just start from your Excel Home tab.



- ➢ Click the ***fx*** button on Formula Bar
- ➢ Select **Statistical** from the **Or select a category:** drop down menu


Or select a category: Statistical

- ➢ Select **CORREL** or **PEARSON** (these two functions produce the same answer) from the **Select a function:** list


Insert Function

Search for a function:

Type a brief description of what you want to do and then click Go

Or select a category: Statistical

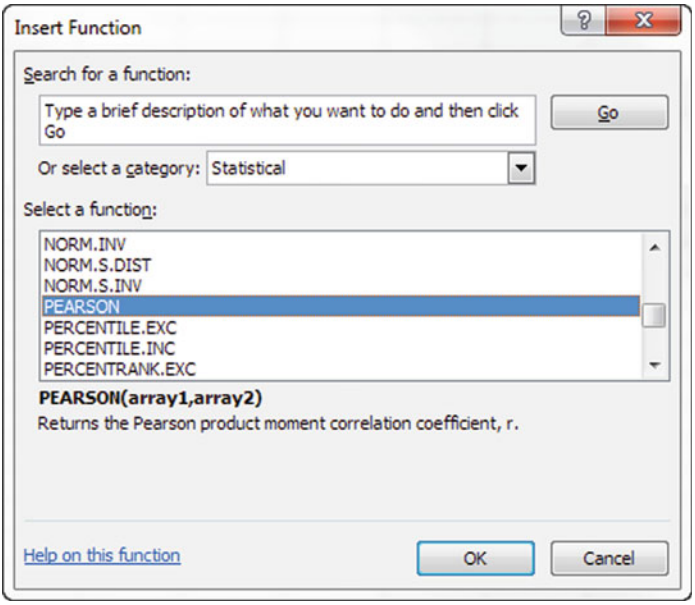
Select a function:

CHISQ.TEST
CONFIDENCE.NORM
CONFIDENCE.T
CORREL
COUNT
COUNTA
COUNTBLANK

CORREL(array1,array2)

Returns the correlation coefficient between two data sets.

OR



Insert Function

Search for a function:

Type a brief description of what you want to do and then click Go

Or select a category: Statistical

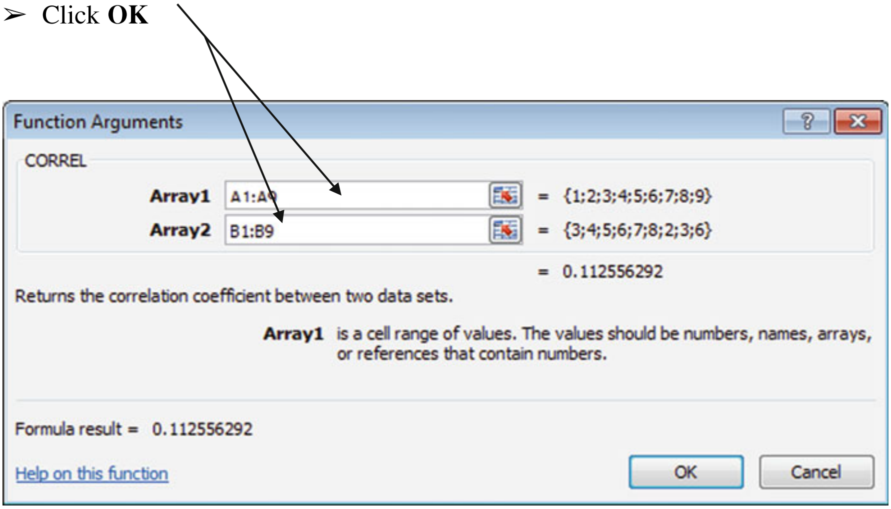
Select a function:

NORM.INV
NORM.S.DIST
NORM.S.INV
PEARSON
PERCENTILE.EXC
PERCENTILE.INC
PERCENTRANK.EXC

PEARSON(array1,array2)

Returns the Pearson product moment correlation coefficient, r.

- ➢ Click **OK**
- ➢ Input the range for the 2 data sets you are correlating
- ➢ Click **OK**



Function Arguments

CORREL

Array1 A1:A9 = {1;2;3;4;5;6;7;8;9}

Array2 B1:B9 = {3;4;5;6;7;8;2;3;6}

= 0.112556292

Returns the correlation coefficient between two data sets.

Array1 is a cell range of values. The values should be numbers, names, arrays, or references that contain numbers.

Formula result = 0.112556292

The results will be the Pearson Correlation Co-efficient (r).

Common Excel Pitfalls

- ◇ Make sure to click the box indicating you are including variable labels, if your data has variable labels.
- ◇ Make sure to click inside the Array1 input box before you highlight the input data field. Then click inside the Array2 data input box before highlighting the next input data field. If you don't click inside Array 2 data input box, you will just overwrite the input you selected for Array 1.
- ◇ By default, Excel will display as many decimals as it can to fill a cell. To reduce the number of digits displayed after the decimal, format the cell to display no more than 4 decimal places.

Final Thoughts and Activities

Practice Problems

1. The following is a set of data from a sample of n = 12 items:

X	3	7	5	15	12	4	9	8	10	11	21	2
Y	6	14	10	28	24	8	18	16	20	22	40	4

   (a) Create a scatter plot of the variables X and Y. What relationship, if any, appears to exist between the two variables?
   (b) Compute the correlation coefficient.
   (c) How strong is the relationship between the two variables? Explain.

2. Customers who made online purchases from an Internet retailer were randomly sampled from the retailer's database. The dollar value of each customer's purchases along with the time the customer spent shopping the company's online catalog were recorded. The sample results are contained in the file **Catalog**.

   (a) Create a scatter plot of the variables ***Time (x)*** and ***Purchases (y).*** What relationship, if any, appears to exist between the two variables?
   (b) Compute the correlation coefficient for these sample data.
   (c) What inferences can you make about the relationship between time spent on the company's online catalog and the dollars spent shopping at the company's site?

3. InTech has created a new technological widget. In the first few months, InTech produces very few widgets, but as time progresses more widgets are available on the market. The data in the **InTech** file provides the number of items in stock at the local warehouse and the cost per item.

(a) Create a scatter plot of the variables ***# of Items in Stock (x)*** and ***Price per Item (y)***. What relationship, if any, appears to exist between the two variables?
(b) Compute the correlation coefficient for this sample data.
(c) What inferences can you make about the supply of widgets and the cost per widget?

4. Different countries sell a variety of drugs at different prices. For each of the 10 popular prescription drugs, the file contains a list of the retail price (US$) for the drug in different countries. Use the data in **DrugPrices** to calculate the answers to the problem.

   (a) Compute the correlation coefficient for these sample data.
   (b) What inferences can you make about the 10 popular prescription drugs and the different costs per country?

Discussion Boards

1. Unemployment is related to violent crime.
2. Baseball players with higher batting averages receive higher salaries. If I am a professional baseball player, raising my batting average will get me a higher salary.
3. Many studies have found an association between smoking and heart disease. One study found an association between coffee drinking and heart disease. As part of your company's wellness program, should you conclude coffee drinking causes heart disease and ban coffee in the workplace?
4. Think of some relationships and associations that you might find at your organization. For example: how are wellness programs related to job satisfaction, or company child care services associated with attendance?

Group Activity

In designing a profitable airline, the number of seats in a plane is an important variable. Likewise customer comfort is also important. Leonardo da Vinci provided some of the earliest data on human measurements. He developed correlational relationships between various measurements. Choose one of his correlational relationships that might still be valid in designing a profitable aircraft.

Parting Thought

I have had my results for a long time; but I do not yet know how I am to arrive at them…

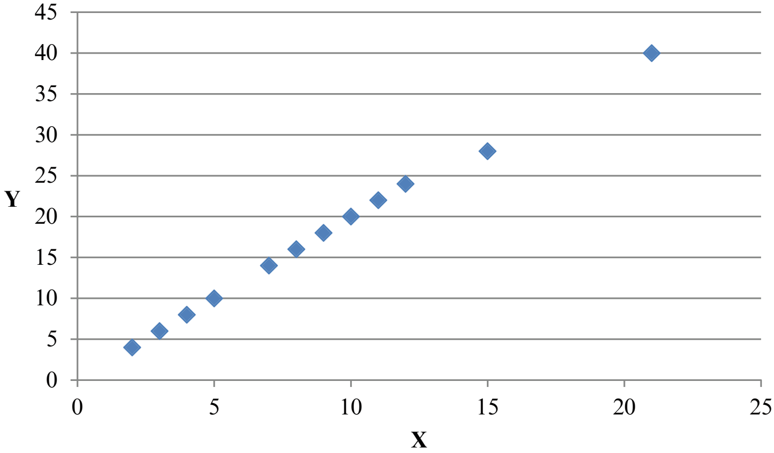
Problem Solutions

1. The following is a set of data from a sample of n = 12 items:

X	3	7	5	15	12	4	9	8	10	11	21	2
Y	6	14	10	28	24	8	18	16	20	22	40	4

   (a) Create a scatter plot of the variables X and Y. What relationship, if any, appears to exist between the two variables?

   Answer:



   (b) Compute the correlation coefficient.

   Answer: r = 0.9989

   (c) How strong is the relationship between the two variables? Explain.

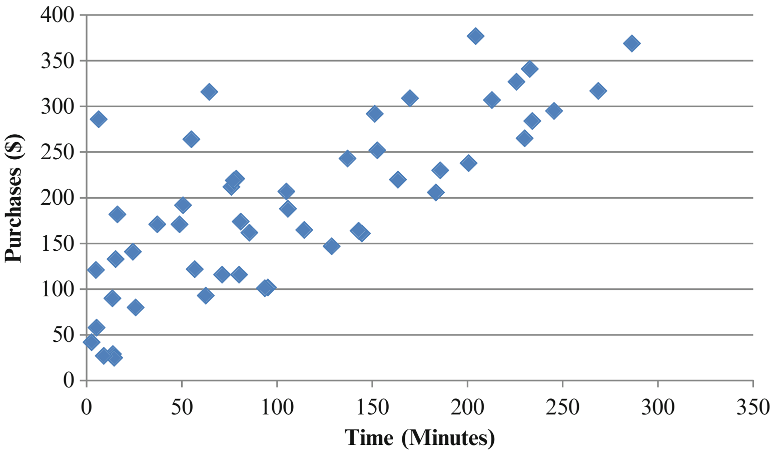
   Answer: There is an almost perfect positive relationship between X and Y. Almost all of the points fall into a straight line with a positive slope.

2. Customers who made online purchases from an Internet retailer were randomly sampled from the retailer's database. The dollar value of each customer's purchases along with the time the customer spent shopping in the company's

online catalog that was recorded. The sample results are contained in the file **Online**.

(a) Create a scatter plot of the variables Time (x) and Purchases (y). What relationship, if any, appears to exist between the two variables? Answer: A very strong positive correlation between online purchases and time spent shopping online.

Answer:



(b) Compute the correlation coefficient for these sample data.

Answer: r = 0.7564

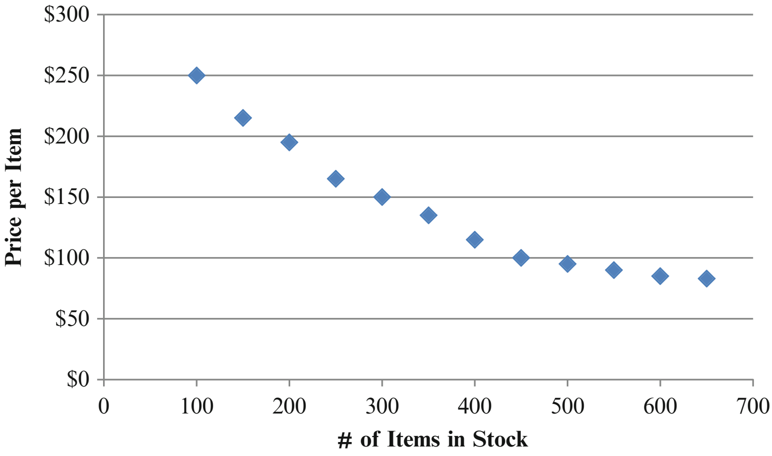
(c) What inferences can you make about the relationship between time spent on the company's online catalog and the dollars spent shopping at the company's site?

Answer: The correlation of 0.7564 implies a strong positive relationship between time spent looking at the company's catalog and the amount of dollars spent shopping at the company's site. Customers who spend a longer period of time looking at the company's online catalog are more likely to spend more money shopping at the company's site.

3. InTech has created a new technological widget. In the first few months, InTech produces very few widgets, but as time progresses more widgets are available on the market. The data in **InTech** provides the number of items in stock at the local warehouse and the cost per item.

(a) Create a scatter plot of the variables # of Items in Stock (x) and Price per Item (y). What relationship, if any, appears to exist between the two variables?

Answer:



(b) Compute the correlation coefficient for these sample data.

Answer: −0.9591

(c) What inferences can you make about the supply of widgets and the cost per widget?

Answer: The correlation of −0.9591 implies a strong negative relationship between the supply of widgets in stock and the cost per widget. The more widgets that are in stock the lower the price per item.

4. Different countries sell a variety of drugs at different prices. For each of the 10 popular prescription drugs, the file contains a list of the retail price (US$) for the drug in different countries. Use the data in **DrugPrices** to calculate the answers to the problem.

(a) Compute the correlation coefficient for these sample data.

Answer:

	US Price	*Canada Price*	*Great Britain Price*	*Australia Price*
US Price	1			
Canada Price	0.9983	1		
Great Britain Price	0.9953	0.9980	1	
Australia Price	0.9959	0.9985	0.9994	1

(b) What inferences can you make about the 10 popular prescription drugs and the different costs per country?

Answer: The correlations imply a strong positive relationship between each of the countries prescription drug costs. Prescription drugs that tend to cost more in the US tend to cost more in Canada, Great Britain, and Australia. Prescription drugs that tend to cost less in the US tend to cost less in Canada, Great Britain, and Australia.

Chapter 10
Simple Linear Regression

Key Concepts

Coefficient of determination, Method of least squares, Market model, Regression, Residual, Residual plots, Slope, Standardized residuals, y- intercept, $\bar{y}$, y_i, and $\hat{y}$.

Discussion

This chapter will introduce the simple linear regression technique that involves only one x and one y. This is a technique that creates a linear fit to the data set. In the simple linear regression technique the (x, y) data generate a line of the form y = mx + b. Figure 10.1 illustrates the linear relationship between temperature in degrees C (x) and the number of flower buds (y). Excel creates the linear trendline that best fits the data. To understand this statistical technique several important concepts are defined throughout this chapter. A step by step process flow is developed in Excel to complete the creation and testing of a simple linear regression model.

Simple Linear Regression: This is a technique that models the linear relationship in your data set. The *simple* refers to using only one x to predict y. The *linear* reminds you that we are fitting a straight line through the data in this approach. The *regression* means pulling the data points back (regressing) to a line that tries to go through the "middle" of the data distribution. The (x,y) data generate a line of the form **y = mx + b** that best describes how changes in y can be predicted from changes in x. This will allow us to estimate the y-value for any given x-value.

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_10,

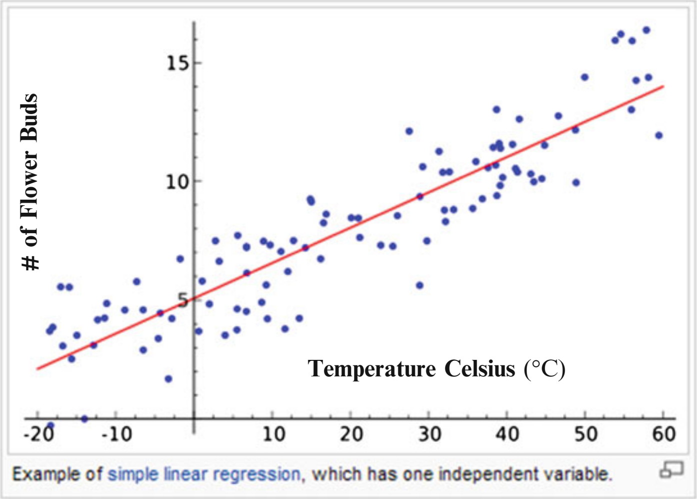


Fig. 10.1 Simple linear regression

Statistics: The **coefficient of determination** (R^2or r^2) indicates how well the line fits your data. It tells you how much of your y can be explained by the x you used in the model. It is most easily understood when expressed as a %.

The regression line does not usually describe the data perfectly. It is a useful summary of the main trend, but it does not capture the random variation of the data points about the line. But how useful is the regression line? The answer is based on the **coefficient of determination** or r^2. The "r squared," tells you how much of the variability in y is explained by x. Larger values of r^2 are considered better because they indicate a stronger relationship between x and y. Often the **coefficient of determination** appears in its capitalized form R^2. It ranges in value from 0 to 1. An r^2 of 1 would indicate that the data all fall perfectly along a straight line.

Although the **correlation coefficient** is restricted to a linear relationship between two variables, the **coefficient of determination** can be used for nonlinear relationships and for relationships that have two or more independent variables. In that sense, the **coefficient of determination** has a wider range of applicability.

The **y-intercept (b)** is the y value when x equals zero. In other words it is where the regression line cuts the y-axis. If the line stops before it hits the y axis then you can extrapolate the line back to where it would cut the axis. If this is not a real value on the line, it may not have any real meaning. For example, graphing the number of employees (x) versus number of units produced (y) when no employees are working may not make sense. Let's assume the y intercept is 100 when x equals zero. This would not make any sense. You would not expect to make 100 units with no employees. So be careful if you extrapolate the line to create a y-intercept, it may not make any sense in the real world! In Fig. 10.2, the y-intercept is 2.0. When there is no wage inflation there is still a price inflation of 2 %. That intercept makes sense.

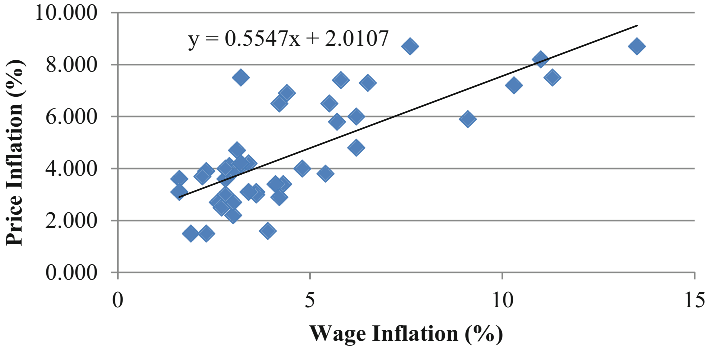


Fig. 10.2 Price versus wage inflation

The **slope (m)** of the regression line is just the increase in y for a unit increase in x. It will be positive if x and y move in the same direction, i.e. both increase or both decrease. It will be negative if they move in opposite directions, i.e. as x increases then y decreases or vice versa. In Fig. 10.2 the slope is 0.55. The equation from this data set is **y = 2 + 0.55 x**. For each increase of 1(%) of x this means y will increase about 0.55(%). In other words, a 1 % increase in wage inflation will result in a 0.55 % increase in price inflation.

In the basic model the (x, y) data generates a line of the form y = mx + b that best describes how changes in y can be predicted from changes in x. This allows you to estimate the y-value for any given x-value.

The **market model** is an important application of simple linear regression. In this model we assume that the rate of return on a stock is linearly related to the rate of return on the overall market. This can be defined as:

$$\begin{array}{lccccc} & R = & \beta_0 & + & \beta_1 & R_m \\ \textbf{compare} & \updownarrow & \updownarrow & & \updownarrow & \updownarrow \\ & y = & b & + & m & x \end{array}$$

Where

R = return on a particular stock (the dependent variable)
β_0 = y intercept
β_1 = slope or the beta coefficient which measures how sensitive the stock's rate of return is to change in the level of the overall market
R_m = return on some major stock index

When β is greater than 1, the stock's rate of return is more sensitive to changes in the level of the "overall" market than is the average stock. In other words, a stock with β greater than 1 will be more volatile than the overall market.

In this model, the slope is a measure of the stock's systematic risk, because it measures the volatility of the stock's price in comparison to the volatility across the system (overall market).

However, there is also non-systematic risk that is associated with the listed stock company. Advertising effectiveness, a recent merger or if the company has filed Chap. 11 (bankruptcy) can be variables that also affect the volatility of the stock. This type of risk is minimized by creating a stock portfolio. However, the systematic or market-related risk remains.

The portfolio's β_1 is estimated by calculating the mean of the entire individual stock betas that are in the portfolio. Depending on an investor's risk aversion different β values will be desirable. Risk-averse investors, who predict a fall in the market, will seek portfolios with β_1 less than 1. Likewise those who believe the market is about to increase, will seek a portfolio with β_1 greater than 1, which represents higher volatility.

A common approach in determining the best fitting line utilizes the **method of least squares**. The following symbols are important in understanding this part of the discussion:

Symbols

x: independent variable
y: dependent variable
y_i: ith value of the measured or observed y
$\hat{y}$: predicted or estimated value of y
$\bar{y}$: average or mean value of y
ε: error or residual term
b: y- intercept
m: slope of the line

The **Method of Least Squares** uses the data to provide those values of the y-intercept and the slope that minimize the sum of the squares of the differences between the observed values (y_i) of the dependent variable and the estimated (predicted) values of the dependent value ($\hat{y}$). This difference is referred to as the residual or error term (ε). The **residual** is simply the difference between the measured value of y and the predicted value of y. The predicted value of y is generated from the least squares regression line. The real value of y is the data you collected.

Figure 10.3 includes data from a model in which we want to predict quarterly sales for restaurants by knowing the student populations nearby. Residuals are shown in the last column. The residual is the difference between the actual observed sales data in the third column, minus the sales data predicted by the regression line in the fourth column.

Restaurant	**Student Pop** **(1000s)**	**Quarterly Sales** **($US 1000s)**	**Excel Predicted Quarterly Sales ($US 1000s)**	**Residuals** **($US 1000s)**
1	2	58	70	-12
2	6	105	90	15
3	8	88	100	-12
4	8	118	100	18
5	12	117	120	-3
6	16	137	140	-3
7	20	157	160	-3
8	20	169	160	9
9	22	149	170	-21
10	26	202	190	12

Fig. 10.3 Residuals

In Fig. 10.4 the values of the independent variable (x) are plotted along the horizontal axis. The vertical axis is used to plot both the observed values of y and the Excel estimated (predicted) values of y for each and every value of x. This graph shows the difference between the observed values (y_i) of Quarterly Sales and the Excel predicted values $(\hat{y}_i)$ of Quarterly Sales as vertical lines.

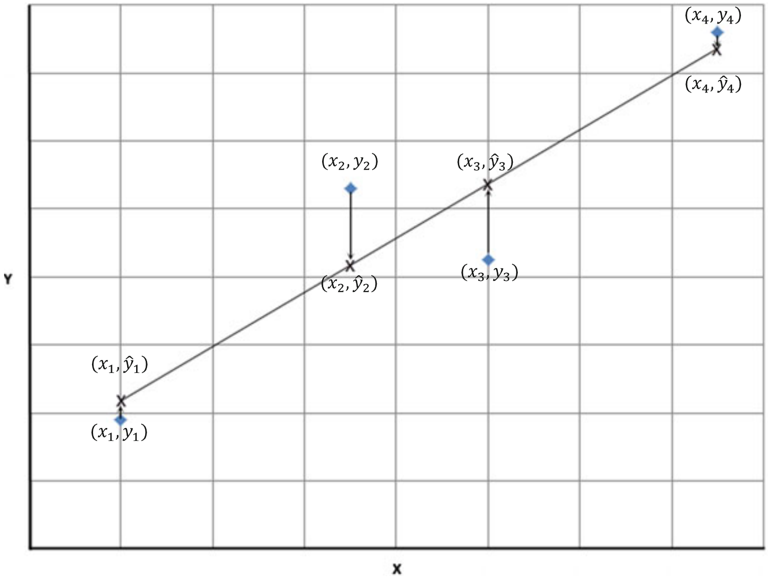


Fig. 10.4 Observed values (y_i) and estimated values $(\hat{y}_i)$ for the dependent variable

Residuals and Tests for Linearity

To accurately apply the linear regression model, it is important to verify that our data really do follow a linear relationship. Assume we have the same data set where the independent variable is student population and the dependent variable is quarterly sales at local pizza restaurants. Remember in this model we are interested in predicting sales from the student population data.

Figure 10.5 depicts a residual plot, with the residuals $(y - \hat{y})$ on the vertical axis and the independent variable (x) along the horizontal axis. The symbol for residuals is the lower case Greek letter epsilon (ε).

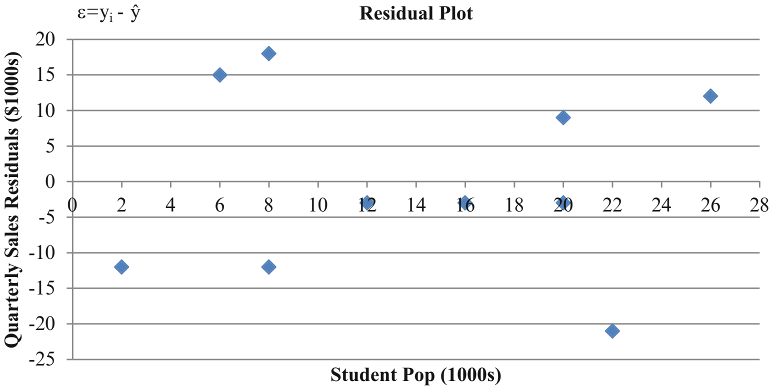


Fig. 10.5 Residual plots

The plot in Fig. 10.5 shows a random distribution of residuals which is what we want if the data is truly linear in nature. In other words, we do not want to see a pattern that would allow us to feel comfortable in predicting the next residual value. A pattern raises a red flag and suggests that we really do not have a distribution of data that supports using a linear model. Figures 10.6 and 10.7 are examples of residual plots that raise a red flag.

Residual Plot Against x

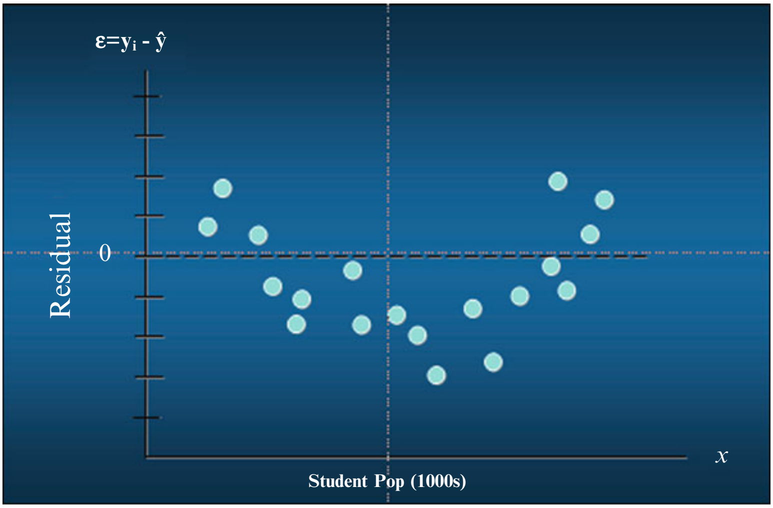


Fig. 10.6 Residual plots

Residual Plot Against x

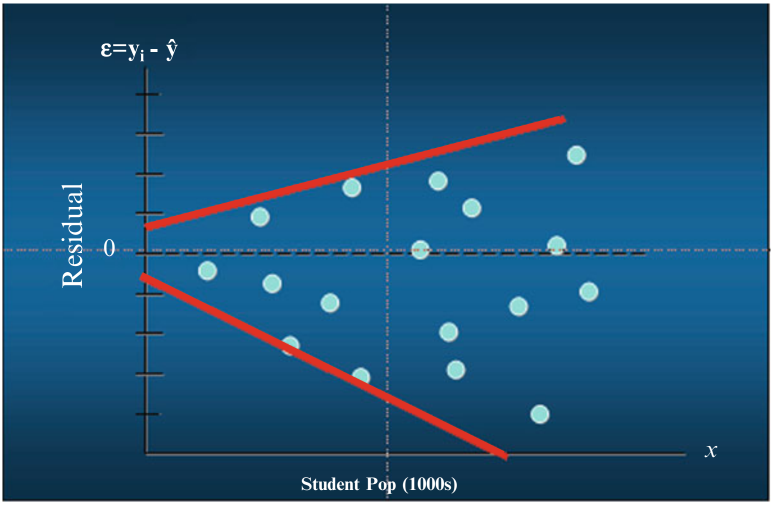
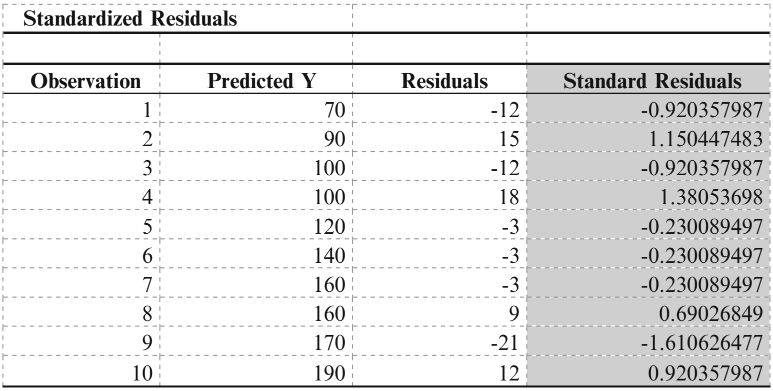


Fig. 10.7 Residual plots

Figure 10.6 shows a curvilinear distribution of residuals; not good. Figure 10.7 shows small values or residuals for small values of x and progressively larger residuals for larger values of x in a funnel distribution; also, not good. Neither of these data sets suggests that a simple linear model would make sense to use for analysis and prediction.

Standardized Residuals and Outliers

Standardized residuals are residuals that have been converted into standard units. Residual analysis in standard units can help us determine whether outliers are present in the data. For normally distributed **standardized residuals**, we would expect about 99 % of our residuals to fall between ±3 standard units. Check if this is true by eyeballing the data. If you have values that are less than −3 or greater than +3 these may be outliers or they may be extreme values that should still be included. The first step is always to check if the measurement was recorded accurately. If after investigation the value is valid, you could run the model with and without the outlier to determine the impact and then decide if it is ok to exclude the value from your model (Fig. 10.8).



Standardized Residuals

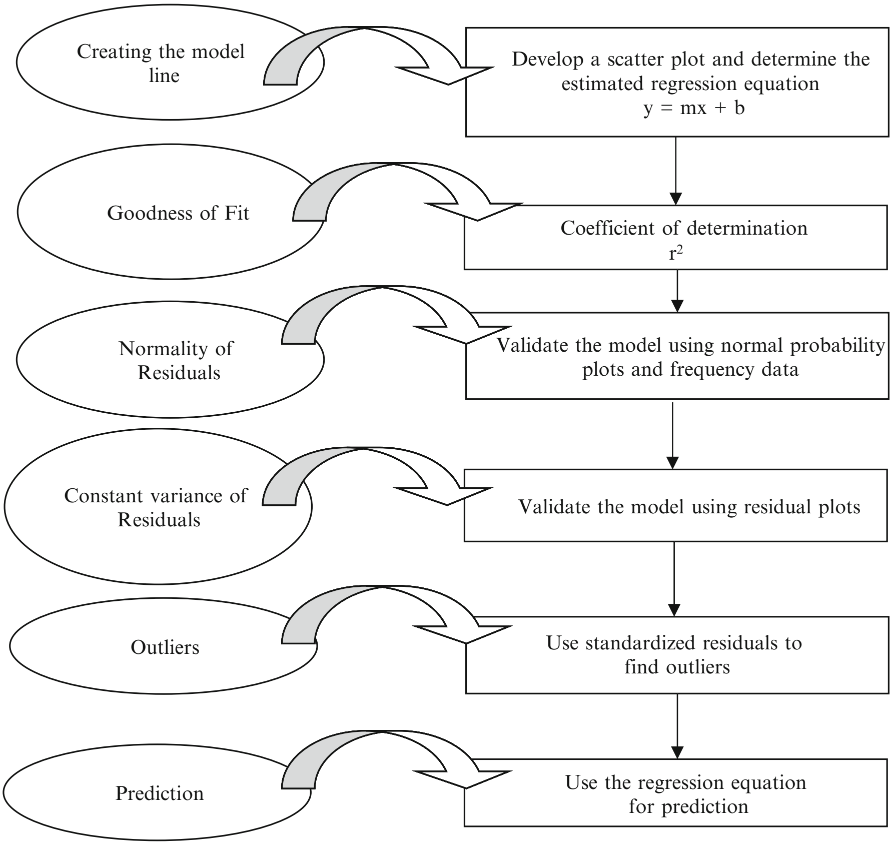
Observation	Predicted Y	Residuals	Standard Residuals
1	70	-12	-0.920357987
2	90	15	1.150447483
3	100	-12	-0.920357987
4	100	18	1.38053698
5	120	-3	-0.230089497
6	140	-3	-0.230089497
7	160	-3	-0.230089497
8	160	9	0.69026849
9	170	-21	-1.610626477
10	190	12	0.920357987

Fig. 10.8 Standardized residuals

None of these standardized residuals violate our rule, so we will assume we do not have any outliers in this data set.

Regression Summary

The next section will provide detailed Excel instructions on completing each of the steps shown below. A thorough discussion on significance tests associated with regression is described in the next two chapters rather than inserting it here. But all model analysis should also complete significance testing in addition to the 6 steps indicated.


Creating the model line
Develop a scatter plot and determine the estimated regression equation
y = mx + b
Goodness of Fit
Coefficient of determination
r²
Normality of Residuals
Validate the model using normal probability plots and frequency data
Constant variance of Residuals
Validate the model using residual plots
Outliers
Use standardized residuals to find outliers
Prediction
Use the regression equation for prediction

◇ Remember to complete significance testing as described in Chaps. 11 and 12.

Excel

Regression tools include analysis completed through use of the Scatterplot and through use of the Regression function. Both of these techniques will be described in detail in this next section. The topics covered in this section will include:

1. Computing the regression line and the coefficient of determination using the **Scatterplot function** (**Create model, R^2 Goodness of fit**)
2. Developing a regression model using the **Regression function** (**Create model**)
3. Compute residual plots using the Regression function
4. Testing for normality of the distribution of the residuals (**Normality of residuals**)

5. Testing for constant variance of the residuals (**Outliers, Constant variance of residuals**)

After completing the analyses using these EXCEL tools you are now in a position to use the model for prediction.

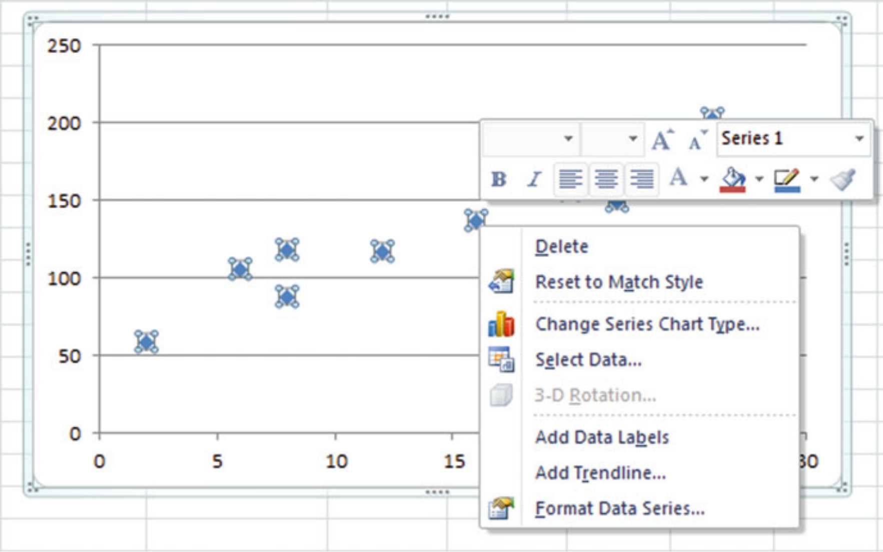
Scatterplot: Compute the Regression Line and the Coefficient of Determination

A prospective buyer of a pizza restaurant is interested in knowing if the student population near the restaurant may help him estimate what the expected sales should be. Data were collected such that the independent variable is student population at universities near 10 pizza restaurants and the dependent variable is quarterly sales at those local pizza restaurants.

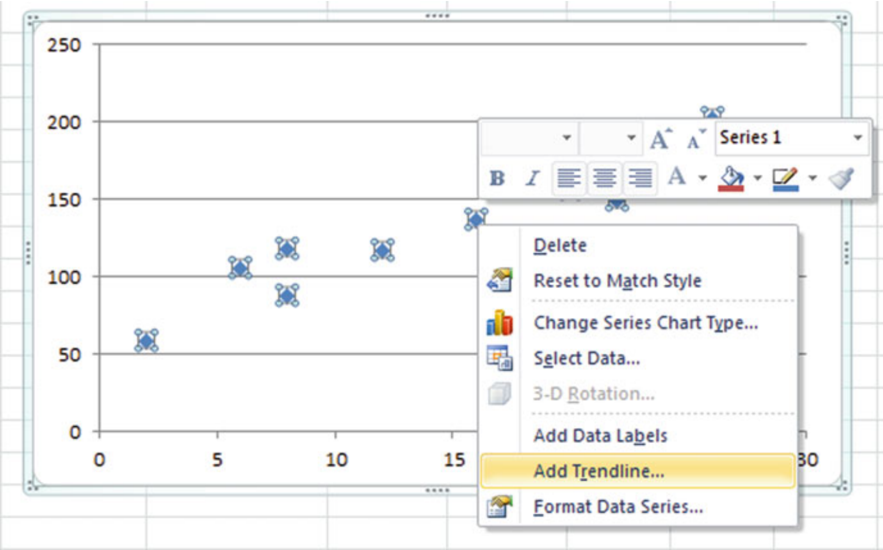
Input Data for Example Problem:

Restaurant	Student population (1,000s)	Quarterly sales ($US 1,000s)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

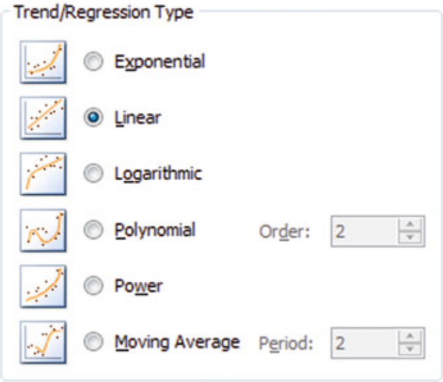
- ➢ Input the data from columns two and three. Include labels
- ➢ Create a scatter plot using the **Insert Chart** function
- ➢ Position the mouse over any data point in on the scatter graph and right click to display a list of options



➢ Select the **Add Trendline** option



➢ Select **Linear**



Trend/Regression Type

○ Exponential

● Linear

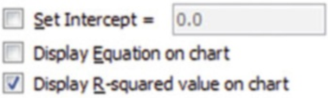
○ Logarithmic

○ Polynomial Order: 2

○ Power

○ Moving Average Period: 2

➢ Choose **Display R-squared value on chart** option

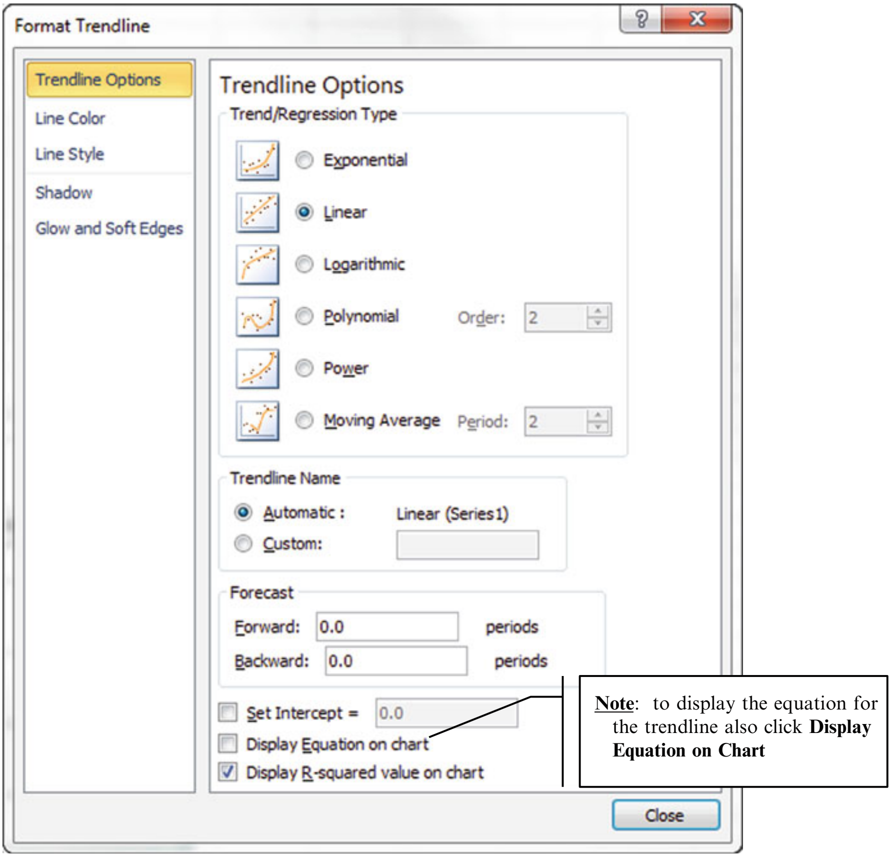


□ Set Intercept = 0.0

□ Display Equation on chart

☑ Display R-squared value on chart

➢ Click **Close**



Format Trendline

- Trendline Options
- Line Color
- Line Style
- Shadow
- Glow and Soft Edges

Trendline Options

Trend/Regression Type

○ Exponential
● Linear
○ Logarithmic
○ Polynomial Order: 2
○ Power
○ Moving Average Period: 2

Trendline Name

● Automatic : Linear (Series1)
○ Custom:

Forecast

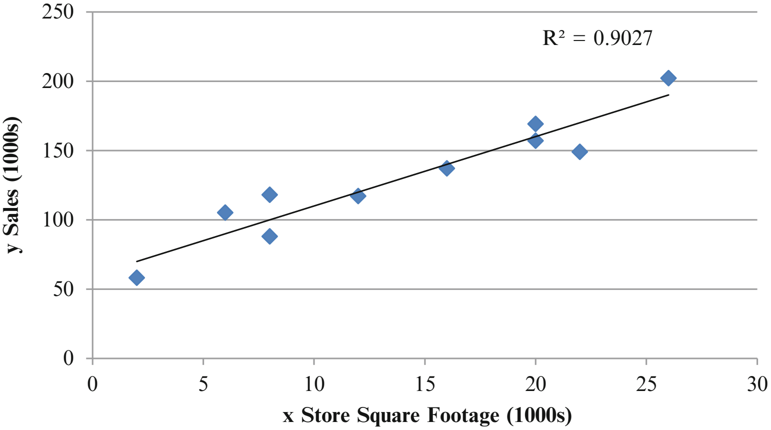
Forward: 0.0 periods
Backward: 0.0 periods

□ Set Intercept = 0.0
□ Display Equation on chart
☑ Display R-squared value on chart

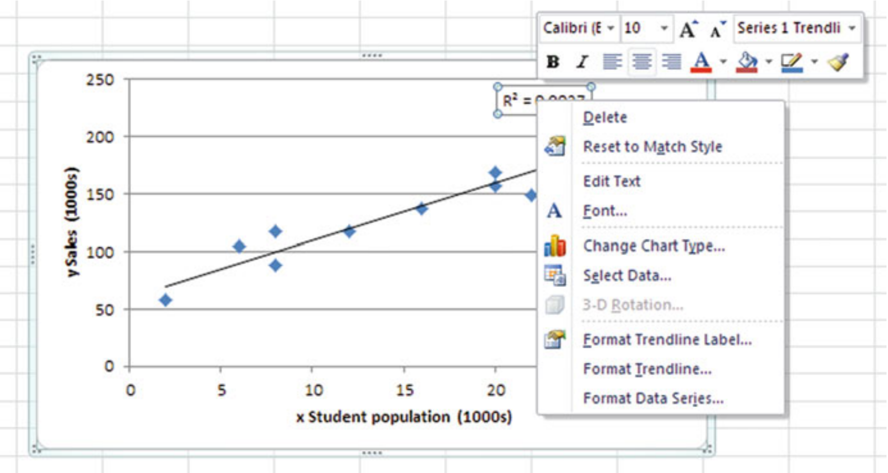
Close

Note: to display the equation for the trendline also click **Display Equation on Chart**

Resulting Trendline and R² display



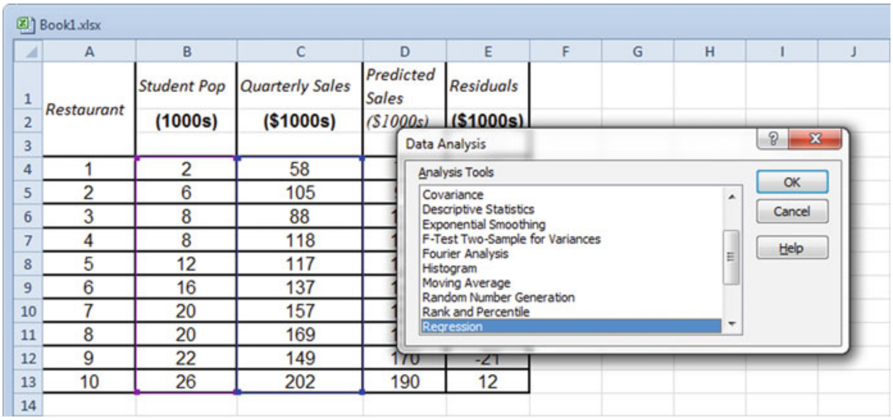
If you want to change the font or size of the labels for the trendline, just right click on them. You can select **Edit Text** or **Font** to allow you to edit as desired.



If you want to move the labels to a different location on the graph, just left click on one of the corners of the text box and drag it to another location.

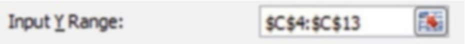
Regression Function: Compute the Regression Model

- ➢ Select the **Data** tab
- ➢ Choose the **Data Analysis** function
- ➢ Choose the **Regression** from the list of Analysis Tools



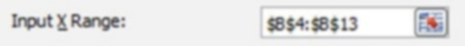
Restaurant	Student Pop (1000s)	Quarterly Sales ($1000s)	Predicted Sales ($1000s)	Residuals ($1000s)
1	2	58		
2	6	105		
3	8	88	1	
4	8	118	1	
5	12	117	1	
6	16	137	1	
7	20	157	1	
8	20	169	1	
9	22	149	170	-21
10	26	202	190	12

- ➢ Click **OK**
- ➢ Enter the y data in the **Input Y Range** box



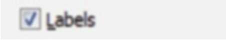
Input Y Range: $C$4:$C$13

- ➢ Enter the x data in the **Input X Range** box



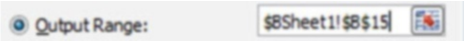
Input X Range: $B$4:$B$13

- ➢ Select **Labels** if you have them (a good idea!)



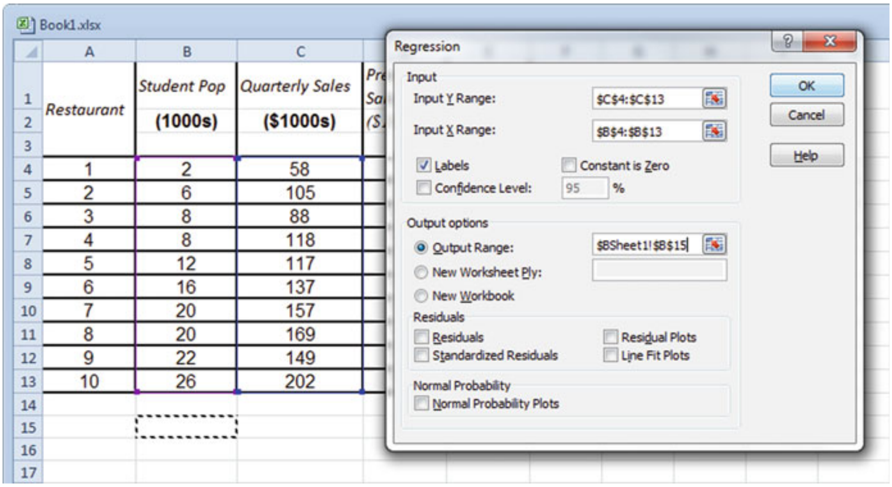
☑ Labels

- ➢ Select **Output Range**. Enter the cell to identify the upper left corner of the section of the worksheet where the output will appear



● Output Range: $BSheet1!$B$15

- ➢ Click **OK**



Book1.xlsx

	Restaurant	Student Pop (1000s)	Quarterly Sales ($1000s)
4	1	2	58
5	2	6	105
6	3	8	88
7	4	8	118
8	5	12	117
9	6	16	137
10	7	20	157
11	8	20	169
12	9	22	149
13	10	26	202

Regression

Input
Input Y Range: $C$4:$C$13
Input X Range: $B$4:$B$13
☑ Labels
☐ Constant is Zero
☐ Confidence Level: 95 %

Output options
● Output Range: $BSheet1!$B$15
○ New Worksheet Ply:
○ New Workbook

Residuals
☐ Residuals
☐ Residual Plots
☐ Standardized Residuals
☐ Line Fit Plots

Normal Probability
☐ Normal Probability Plots

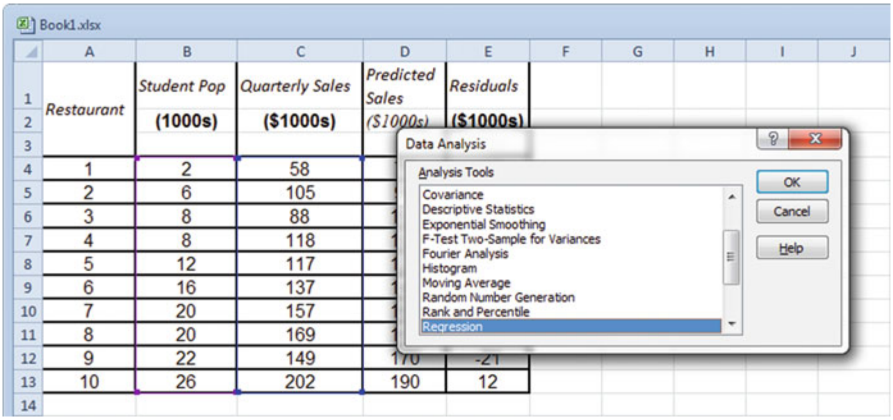
See Fig. 10.9. For the actual output.

Compute Residual Plots Using the Regression Function

➢ Select the **Data** tab

➢ Choose the **Data Analysis** function

➢ Choose the **Regression** function from the list of Analysis Tools



Book1.xlsx

	A	B	C	D	E
1–2	*Restaurant*	*Student Pop* **(1000s)**	*Quarterly Sales* **($1000s)**	*Predicted Sales* *($1000s)*	*Residuals* **($1000s)**
4	1	2	58		
5	2	6	105		
6	3	8	88		
7	4	8	118		
8	5	12	117		
9	6	16	137		
10	7	20	157		
11	8	20	169		
12	9	22	149	170	-21
13	10	26	202	190	12

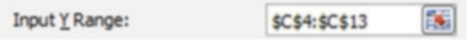
Data Analysis

Analysis Tools

Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression

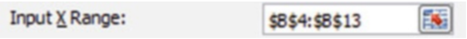
When the Regression dialog box appears:

➢ Enter your **Input Y Range** for you data



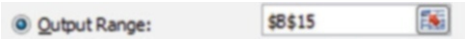
Input Y Range: $C$4:$C$13

➢ Enter your **Input X Range** for you data



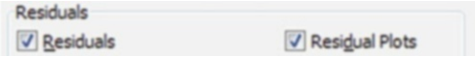
Input X Range: $B$4:$B$13

➢ Select **Output Range** and include upper left corner of the section of the worksheet where the output will be printed



● Output Range: $B$15

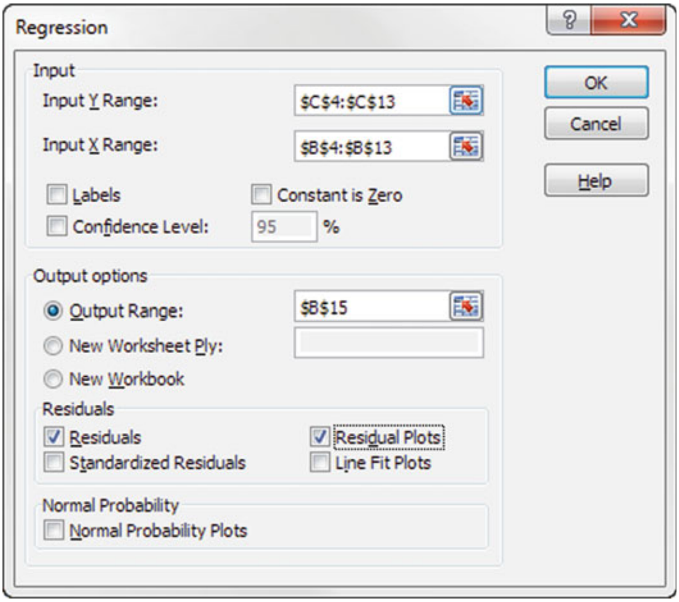
➢ Select the **Residuals** and **Residual Plots** option



Residuals

☑ Residuals ☑ Residual Plots

- ➢ Click **OK**



Regression

Input

Input Y Range: $C$4:$C$13

Input X Range: $B$4:$B$13

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

● Output Range: $B$15

○ New Worksheet Ply:

○ New Workbook

Residuals

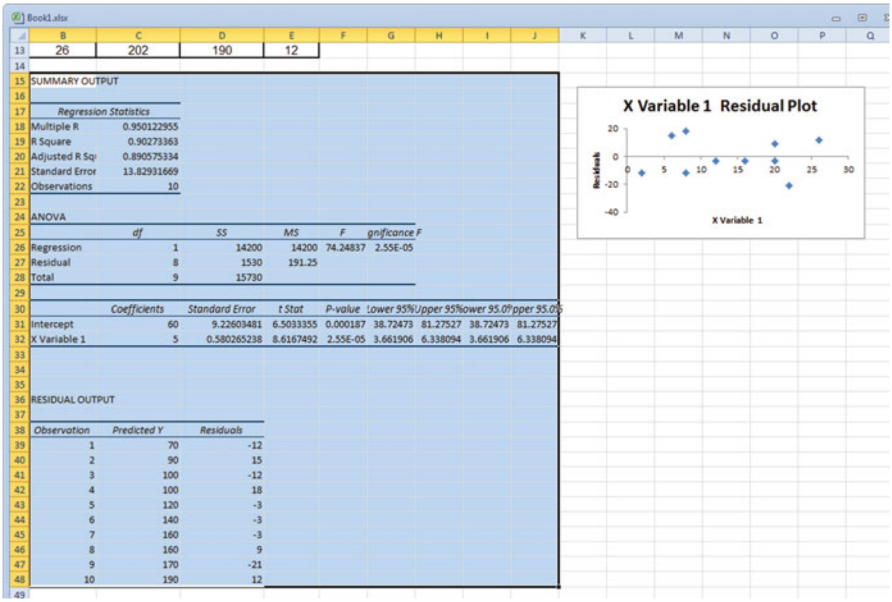
☑ Residuals ☑ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

The resulting output will show a plot of the residuals ($y_i - \hat{y}$) against the independent variable (x). If you also selected **Residuals** as well as **Residual Plots,** Excel will provide a table of all the observed y values and the associated predicted y values. Remember the numerical difference between these two y-values is defined as the residual.



Book1.xlsx

26	202	190	12

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950122955
R Square	0.90273363
Adjusted R Sq	0.890575334
Standard Error	13.82931669
Observations	10

ANOVA

	df	SS	MS	F	gnificance F
Regression	1	14200	14200	74.24837	2.55E-05
Residual	8	1530	191.25		
Total	9	15730			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	60	9.22603481	6.5033355	0.000187	38.72473	81.27527	38.72473	81.27527
X Variable 1	5	0.580265238	8.6167492	2.55E-05	3.661906	6.338094	3.661906	6.338094

RESIDUAL OUTPUT

Observation	Predicted Y	Residuals
1	70	-12
2	90	15
3	100	-12
4	100	18
5	120	-3
6	140	-3
7	160	-3
8	160	9
9	170	-21
10	190	12

Fig. 10.9 Output chart and residual plot

Using Excel's Regression Tool to Test for Normality of the Distribution of Residuals

A robust model will include a normal distribution of the residuals.

Method 1: Normal Probability Plot

You follow the same steps as you used to construct a residual plot but now you should also select **Normal Probability Plot**. Excel will provide you with a normal probability plot. The normalized values of the residuals are plotted on the y-axis and the independent variable remains on the x-axis. If the points approximate a straight line the residuals could have come from a normal distribution which is what you want.

Method 2: Normal Distribution of Residuals

In addition to running a normal probability plot you could simply create a frequency plot of the residuals as shown in Fig. 10.10. As long as the frequency plot approximates a normal distribution you are good to go.

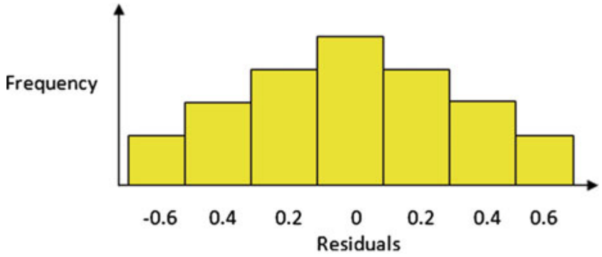
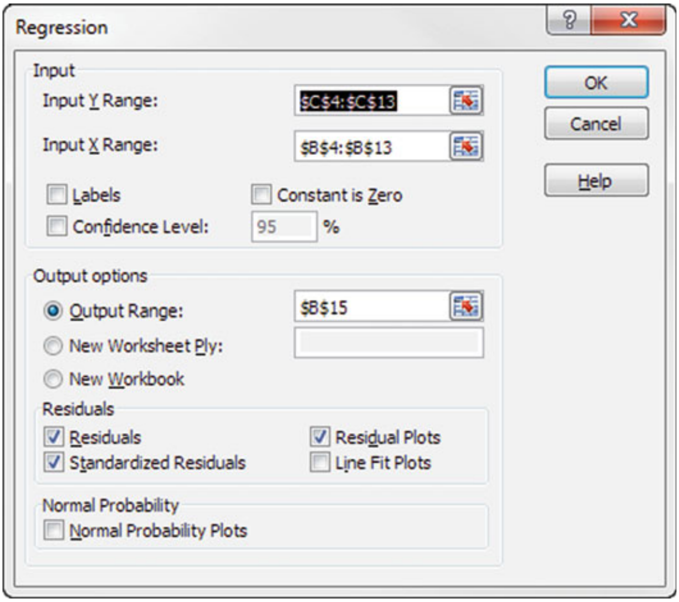


Fig. 10.10 Frequency plot of residuals

In this case, both tests for normality of the residual distribution are acceptable.

Using Excel's Regression Tool to Test for Constant Variance of Residuals

Follow the same steps as you used to construct a residual plot, but now, you should also select **Standardized Residuals**. Excel will provide you with a table of standardized residuals for each y value. Any residuals beyond ±3 standard units should be treated as outliers. Too many outliers may raise a red flag with your data set.



Input

Input Y Range: $C$4:$C$13

Input X Range: $B$4:$B$13

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

● Output Range: $B$15

○ New Worksheet Ply:

○ New Workbook

Residuals

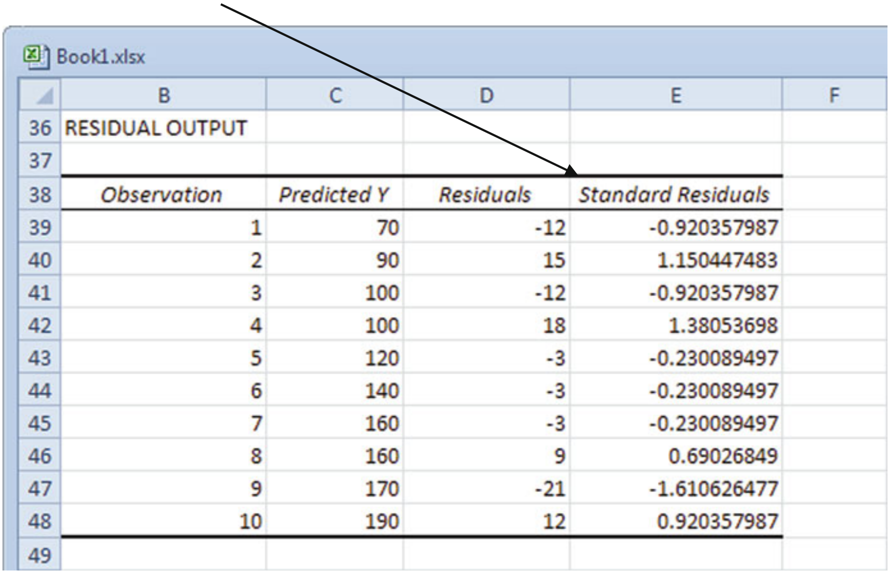
☑ Residuals ☑ Residual Plots

☑ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

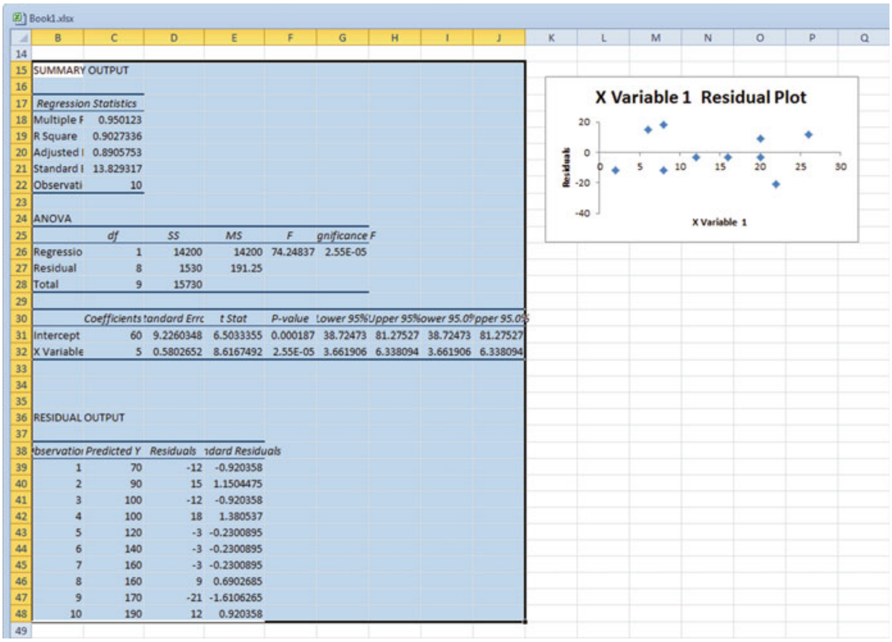
Standard Residuals Output:



RESIDUAL OUTPUT

Observation	Predicted Y	Residuals	Standard Residuals
1	70	-12	-0.920357987
2	90	15	1.150447483
3	100	-12	-0.920357987
4	100	18	1.38053698
5	120	-3	-0.230089497
6	140	-3	-0.230089497
7	160	-3	-0.230089497
8	160	9	0.69026849
9	170	-21	-1.610626477
10	190	12	0.920357987

Standard residual plots are often quite small and need to be enlarged to identify any patterns. To do this just grab a corner and pull to the size you desire.



SUMMARY OUTPUT

Regression Statistics	
Multiple F	0.950123
R Square	0.9027336
Adjusted I	0.8905753
Standard I	13.829317
Observati	10

ANOVA

	df	SS	MS	F	gnificance F
Regressio	1	14200	14200	74.24837	2.55E-05
Residual	8	1530	191.25		
Total	9	15730			

	Coefficients	tandard Errc	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	60	9.2260348	6.5033355	0.000187	38.72473	81.27527	38.72473	81.27527
X Variable	5	0.5802652	8.6167492	2.55E-05	3.661906	6.338094	3.661906	6.338094

RESIDUAL OUTPUT

bservatio	Predicted Y	Residuals	ıdard Residuals
1	70	-12	-0.920358
2	90	15	1.1504475
3	100	-12	-0.920358
4	100	18	1.380537
5	120	-3	-0.2300895
6	140	-3	-0.2300895
7	160	-3	-0.2300895
8	160	9	0.6902685
9	170	-21	-1.6106265
10	190	12	0.920358

Summary of Regression Analysis Process

Step 1: Develop a model that is based on realistic relationships. For the dependent variable, find an independent variable that makes sense as a variable which can generate a linear relationship. For example, using how much caffeine you have consumed to predict how many pages in a managerial report you can read, doesn't make sense as a linear relationship. However, how many times your office assistant interrupts you, may have a linear relationship with number of pages in a report that you can read. We know that most likely the relationship won't be perfect, so we need to recognize the presence of an error term (residuals). (Create model)

Step 2: Gather x,y data pairs. (Create model)

Step 3: Create a scatter plot to decide if a linear model is appropriate. At this point, outliers may already be apparent; you need to investigate these before going forward. Remember an outlier can greatly influence all your regression statistics. (Outliers)

Step 4: Use Excel to calculate the regression equation. It is usually helpful to have the equation printed on the scatter plot. (Create model)

Step 5: Use Excel to generate the coefficient of determination (r^2). This gives you an idea of how much of the variance in y can be explained by the variance in x. This value can range from 0 to 1. The closer the r^2 is to zero the weaker the fit. (Goodness of Fit)

Step 6: Validate the linear regression model can be used to represent this data. The validation tests have to do with the residuals, because our entire modeling technique is based on minimizing the residual. You can also think of it as running checks on "ε" in our model equation.

> The **first test** is to check that the distribution of frequency data of the residuals is normal. We have two ways to do this: a histogram of the residuals, a normal probability plot. Remember, in a normal distribution of residuals, we would expect 99 % of our residuals to fall between +/−3 standard units. (Normality of residuals)
>
> The **second test** is a residual plot that compares residuals to the x values. Residuals should be randomly distributed around zero; otherwise a clearly visible pattern in this plot indicates a red flag. Without relatively constant variance (change in the residuals), we have a problem and another model should be used. In other words, the residual plots should not indicate any strong pattern that would allow you to predict the next residual value. (Constant variance of residuals)

Step 7: If all tests were good in Step 6, it is now appropriate to use the regression equation to predict values of the dependent variable (y) for given values of (x). (Prediction)

Common Excel Pitfalls

- Make sure to click the box indicating you are including variable labels.
- Make sure to click inside the **Input Y** box before you highlight the input data field. Then, click inside the **Input X** box before highlighting the next input data field. If you don't click inside the **Input X** box you will just overwrite the input you selected for **Input Y**.
- Look at the data for a linear model before you run a linear regression; don't just cram data in a model for the sake of generating a model.
- Be consistent with your units throughout the analysis including on the graphs.
- Pay attention to outliers.
- Make sure to label your independent and dependent variables, x being your independent variable and y being your dependent variable.

Final Thoughts and Activities

Practice Problems

1. The marketing manager of a supermarket chain is trying to see if shelf space can predict sales for breakfast cereal. A random sample of 12 equal-sized stores is selected with the following results.

Number of shelves	Sales ($U.S. 1,000s)	Aisle location
5	160	0
5	220	1
5	140	0
10	190	0
10	240	0
10	260	1
15	230	0
15	270	0
15	280	1
20	260	0
20	290	0
20	310	1

   (a) Construct a scatter plot.
   (b) Determine the coefficient of determination, R^2, and interpret its meaning.
   (c) How useful do you think this regression model is for predicting sales?

2. The Director of Human Resources at Jaunk-It moving company wants to predict the number of cubic feet of goods moved, based on labor hours (stored in **Jaunkit** file). In other words if he pays his workers for 12 h of work, about

how many cubic feet of goods are they moving into the storage unit? Use the data to help the Director understand this relationship.

(a) Construct a scatter plot.
(b) Determine the coefficient of determination, R^2, and interpret its meaning.
(c) How would you explain this relationship to the Director?

3. Nine's Wine Vineyard makes a selection of wines which goes through independent taste testing every year. In the file **Wine**, the different varieties of wine are ranked by taste and number of bottles sold. Use the data to help the Nine's Wine Vineyard understand the relationship between its wine and sales.

   (a) Construct a scatter plot.
   (b) Determine the coefficient of determination, R^2, and interpret its meaning.
   (c) How would you use this regression model for predicting sales?

Discussion Boards

1. Research at the University of Texas indicates a negative correlation between economic growth and the percentage of lawyers in the work force. Predatory litigation lowered GNP by 10 % below its potential level during the 1980s; discuss this relative to the past 10 years.
2. Monthly charges for cell phones have decreased as worldwide subscribers have increased. Is it reasonable to use worldwide subscriptions to predict your monthly budget for your company cell phone?
3. CEOs have been under serious criticism from organized labor for the fat paychecks they receive. Many CEOs claim they put in longer hours, endure more stress, etc. In fact, there is a linear relationship between CEO salaries and financial success of a company. Discuss this using current data.

Group Activities

1. Provide data from the web to develop a regression relationship between calories and grams of fat for pizza. As a conscientious employee, you want to be able to predict the calorie content of the new pizza you are featuring in the cafeteria that has 10.5 g of fat.
2. Discuss the beta values of your favorite stock and favorite mutual fund. Have they changed much over the last 5 years?
3. Your company is getting ready to move into a new office space, but painting will be required. The toughness of paint is usually a good indication of how long it will last. Consumer Reports rates paints on their toughness and also provides an overall rating for each paint. The paint that has been selected for your office has

an overall rating of 65, what toughness would you expect? Build a linear model from the data provided on the Consumer Reports website.

4. Moving people into cubicles increases our profit. Choose a position (pro or con) on this statement. Collect data from the web to support your position. Include some regression analysis.

Parting Thought

Statisticians must stay away from children's toys because they regress so easily. . .and. . .the last few available graves in the cemetery are called residual plots. . .

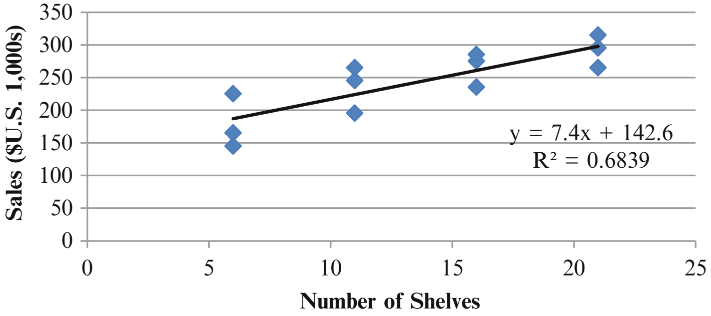
Problem Solutions

1. The marketing manager of a supermarket chain is trying to see if the number of shelves can predict sales for breakfast cereal. A random sample of 12 equal-sized stores is selected with the following results.

Number of shelves	Sales ($U.S. 1,000s)	Aisle location
5	160	0
5	220	1
5	140	0
10	190	0
10	240	0
10	260	1
15	230	0
15	270	0
15	280	1
20	260	0
20	290	0
20	310	1

Note in the solution for this problem we don't actually need to use the aisle location data. Be careful to identify the key variables for each regression model. The independent variable is the number of shelves and our variable of interest is the dependent variable, sales.

(a) Construct a scatter plot



(b) Determine the coefficient of determination, R^2, and interpret its meaning.

Answer: $R^2 = 0.6839$
Therefore, 68 % of sales can be explained by number of shelves.

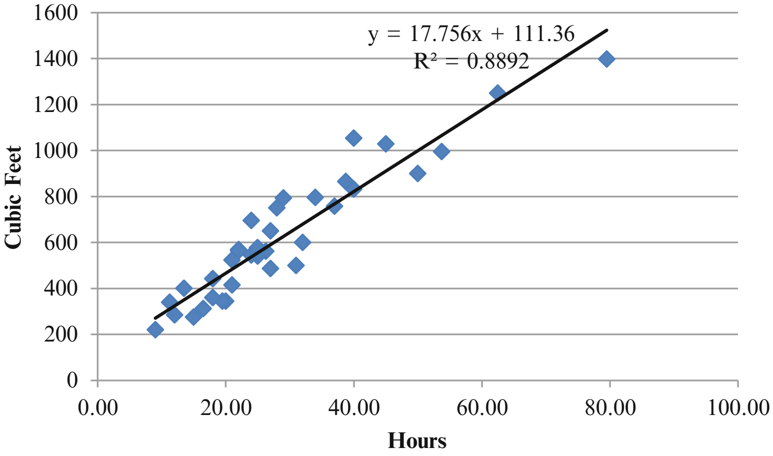
(c) How useful do you think this regression model is for predicting sales?

This regression is helpful in predicting sales because it helps the marketing manager understand that for each increase in shelf space of an additional foot, weekly sales are estimated to increase by $7.40 (from the slope).

2. The Director of Human Resources at Jaunk-It moving company wants to predict the number of cubic feet of goods moved based on labor hours (stored in **Jaunkit** file). In other words if he pays his workers for 12 h of work, about how many cubic feet of goods are they moving into the storage unit? Use the data to help the Director understand this relationship.

The independent variable is the number of hours and our variable of interest is the dependent variable, cubic feet of goods being moved.

(a) Construct a scatter plot



(b) Determine the coefficient of determination, R^2, and interpret its meaning.

Answer: $R^2 = 0.8892$
Therefore, 89 % of cubic feet moved can be explained by the labor hours worked.

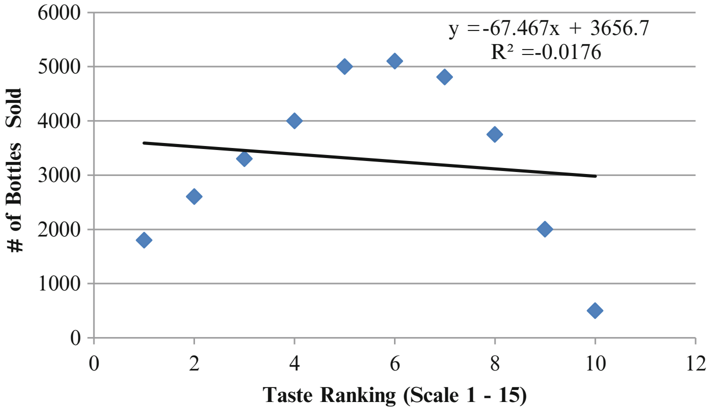
(c) How would you explain this relationship to the Director?

There is a strong linear relationship between number of hours and cubic feet moved. This means that the more cubic feet on a property that requires moving, the more hours it will take for workers to move these things. Knowing how many hours he has paid his employees should help him predict how many cubic feet of stored goods have been moved. If he pays his employees for 30 h of moving time, he should expect **17.76 (30) + 111.36 = 644.16,** or approximately 644 of cubic feet of storage space to be filled with moved goods. Once this relationship is established, he can flag lower performing employees. Higher performing employees may warrant bonuses.

3. Nine's Wine Vineyard makes a selection of wines which goes through independent taste testing every year. In the file **Wine**, the different varieties of wine are ranked by taste and number of bottles sold. Use the data to help the Nine's Wine Vineyard understand the relationship between its wine and sales.

The independent variable is the taste ranking and our variable of interest is the dependent variable, number of bottles sold.

(a) Construct a scatter plot.



(b) Determine the coefficient of determination, R^2, and interpret its meaning.

Answer: $R^2 = -0.017$.

Therefore, only about 2 % of bottles of wine sales can be explained by the taste ranking.

(c) How would you use this regression model is for predicting sales?

This regression model is a poor predictor of sales because it is not a linear relationship, but a curvilinear relationship. You should not use this model for any type of prediction.

Chapter 11
Significance Tests Part 1

Key Concepts

Alternate hypothesis, Chi-square test, Confidence level, Critical values, Cronbach's alpha, Degrees of freedom, F-test, Hypothesis, Null hypothesis, Precision, Reliability, Statistical significance, t-test, Validity, X^2-test, and z-test.

Discussion

The next two chapters help you test your sample results to ensure they are real and not due to chance. Remember none of these tests are necessary if you have access to the entire population. In Part 1 Chap. 11, we will discuss the **F-test** and the **t-test**. In Part 2 Chap. 12 we will discuss the **X^2-test** and the **z-test**. Note there are many more exotic significance tests, beyond the basic four tests presented here.

Every test of significance examines whether the observed difference is real or just due to chance variation. We use **hypotheses** or statements of expected outcome. There are **null hypotheses** and **alternative hypotheses**. Each hypothesis represents one side of the discussion. The **null hypothesis (H_0)** expresses the idea that an observed difference is due to chance. The **alternative hypothesis (H_α)** expresses the idea that an observed difference is real and not due to chance. The **null hypothesis** is the one that is always tested. The **alternative hypothesis** is set up as the opposite of the **null hypothesis**. So if the **null hypothesis** is not true, we can accept the **alternative hypothesis**.

Remember **statistical significance** does not mean "important". **Statistical significance** is measuring the probability our results are real and not due to chance error. The significance level is a critical probability in choosing between the null hypothesis and the alternative hypothesis.

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_11,

Basic Concepts

Degrees of freedom is the number of values in the final calculation of a statistic that are free to vary; calculated by the [number of terms – one]. In other words if you have five variables in a relationship and you select values for four of them, there is no freedom in choosing the fifth. The fifth can only be one value. In this case the number of degrees of freedom would be four.

Validity indicates the extent to which our test is measuring what it is supposed to be measuring. For example many critics claim that although IQ tests provide consistent results, they don't really measure intelligence. They argue IQ tests have a low **validity**.

Reliability refers to the consistency with which results occur. Since classical IQ tests provide consistent results we say they have a high **reliability**.

The following analogy depicts both of these concepts. A customer in a pub is given a set of six darts to throw at a dartboard on the nearby wall. However the dartboard has a white sheet covering it and the area around it. The customer has a general idea where the target is and proceeds to throw the six darts. Figure 11.1 shows the pattern of dart hits on the sheet from the first customer.

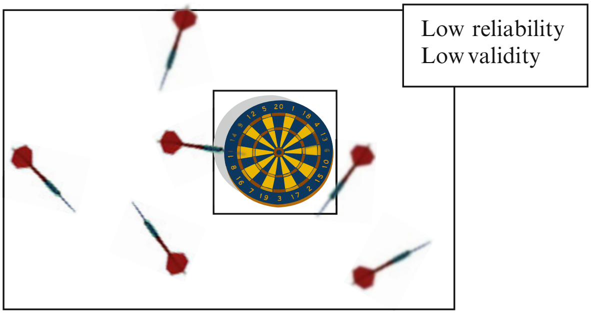


Fig. 11.1 Customer 1 dart hits

Second and third customers also participate. Their results are shown in Figs. 11.2 and 11.3.

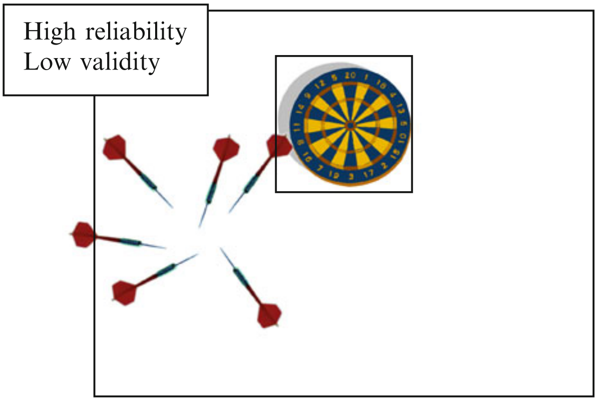


Fig. 11.2 Customer 2 dart hits

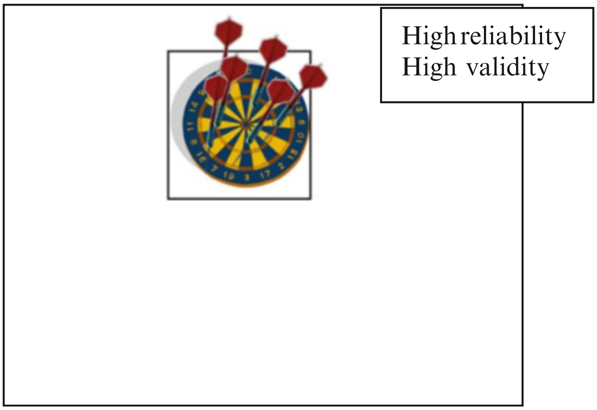


Fig. 11.3 Customer 3 dart hits

Customer 2 and Customer 3 demonstrate high **reliability**. Dart after dart lands in nearly the same place, in other words you can count on these results being generated no matter how many darts are thrown, or in significance testing terms you can count on these results being generated no matter how many samples are chosen.

Proximity to the target is the analogy of **validity** as indicated by Customer 3's results. Close grouping (high reliability) of the dart hits doesn't prove too useful, unless they hit the desired target. In significance testing the desired target is "hitting" the real population statistic. We can also think about **reliability** in terms of **precision**. Therefore, any **standard error** implies an estimation of **reliability**.

To summarize these relationships see Fig. 11.4.

Fig. 11.4 Validity and reliability

If you have...	You can have...
High validity	High reliability only
Low validity	High or low reliability
Low reliability	Low validity only
High reliability	High or low validity

One of the most common **reliability** measures is **Cronbach's alpha (**$\boldsymbol{\alpha}$ or $\mathbf{R}_\alpha$ or $\mathbf{r}_\alpha$**)**. This **reliability** co-efficient can hold values from −1 to +1. The generally agreed upon lower limit is ±0.70, although it may be much less in exploratory research. We typically think of R_α is absolute terms.

Validity can be measured empirically by the correlation between defined sets of variables. The three most widely accepted forms of validity are **convergent**, **discriminant**, and **nomological validity**. **Convergent validity** assesses the degree to which two measures of the same concept are correlated. **Discriminant validity** is the degree to which two conceptually similar ideas are distinct. The empirical test remains correlation among measures, but this time the summated scale (combining several individual variables into a single composite measure) is correlated with a similar but conceptually distinct scale. The correlation should be low demonstrating the two scales are sufficiently different. Finally, the **nomological validity** refers to the degree that the summated scale makes accurate predictions of other concepts.

These types of validity can be assessed using multi-trial, multi-method matrices, structural equation modeling etc. which is beyond the scope of this book.

Significance testing requires you to clearly define what you are trying to prove is in fact real (statistically significant) versus just due to chance. So, for example, suppose you are trying to prove employee training is effective by comparing pre and post training test scores using a sample of employees. If you are using the population data rather than a sample, your observed results are obvious. However, whenever you use sample data, you must test if your results are real or simply due to chance. In other words, if you took another sample, would you get similar results?

In this process you must define the level of risk you are willing to take in assuming your results are real. The only way to be 100 % sure is to use the entire population of data. We typically define the risk using a **Cronbach's alpha (α) value**. The most common α value is .05, which corresponds to a 95 % confidence level. This means that if you selected 100 different samples from this population you would get similar results in 95 of them. Graphically a 95 % confidence level ($\alpha = 0.05$) is depicted in Fig. 11.5.

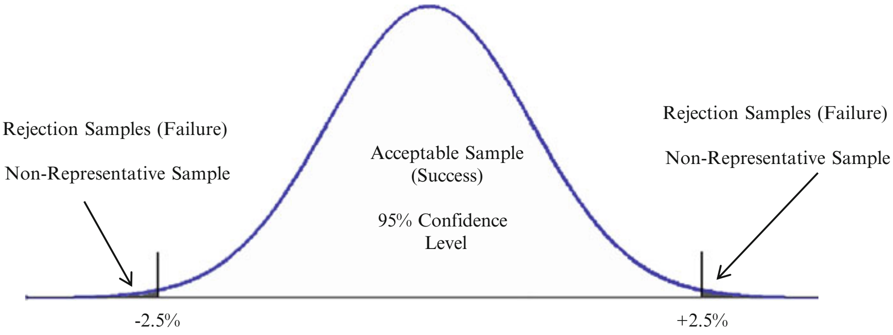


Fig. 11.5 95 % confidence level

But let's say that is too much risk and you want to be 99 % sure, in 99 out of 100 samples from the population you want to get similar results. You would then use an α value of 0.01.

One way to increase the confidence level and thereby decrease the risk is to use larger and larger samples. Remember to have no risk, 100 % confidence level, then your sample size would need to be the entire population of data. As sample size increases, costs (time and money) may also increase; access to the entire population may be difficult with large sample sizes.

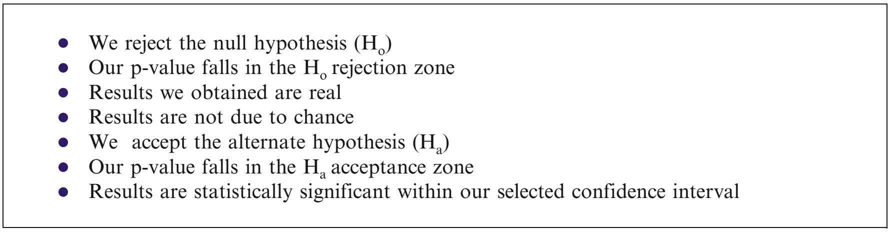
The next step is to establish the relationship between **α** and the **hypothesis**. We always test for the **null hypothesis**. The significance level is a critical probability in choosing between the **null hypothesis** and the **alternative hypothesis**. The level of significance determines the probability level that is considered too low to warrant support of the **null hypothesis**. Here is an example:

H_0: The training has no effect. The average test score before training ($\bar{x}_1$) is really the same as the average score after training ($\bar{x}_2$).

H_a: The test scores before ($\bar{x}_1$) and after training ($\bar{x}_2$) are statistically significantly different within our defined confidence level.

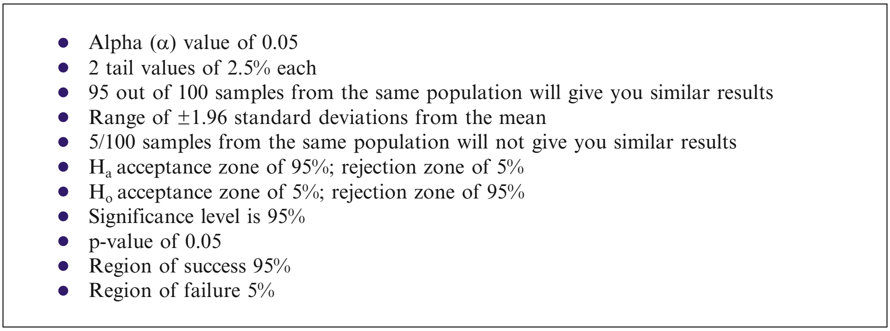
The alpha (α) value defines the tail of the probability distribution. Remember that the model is set up around the null hypothesis. Let's relate the **p-value** Excel calculates from the sample to the model we set up in Fig. 11.5. After calculating the **p-value** based on the sample data, we compare that with the pre-selected α value (risk level). If the calculated **p-value** is less than our selected α, we accept our results are real and not due to chance; we reject the null hypothesis.

There are many different ways to phrase rejecting the null hypothesis as summarized in Fig. 11.6. Likewise Fig. 11.7 summarizes the results for the terminology associated with a 95 % confidence level.



- We reject the null hypothesis (H_o)
- Our p-value falls in the H_o rejection zone
- Results we obtained are real
- Results are not due to chance
- We accept the alternate hypothesis (H_a)
- Our p-value falls in the H_a acceptance zone
- Results are statistically significant within our selected confidence interval

Fig. 11.6 Summary of the terminology for **statistically significant** results



- Alpha (α) value of 0.05
- 2 tail values of 2.5% each
- 95 out of 100 samples from the same population will give you similar results
- Range of ±1.96 standard deviations from the mean
- 5/100 samples from the same population will not give you similar results
- H_a acceptance zone of 95%; rejection zone of 5%
- H_o acceptance zone of 5%; rejection zone of 95%
- Significance level is 95%
- p-value of 0.05
- Region of success 95%
- Region of failure 5%

Fig. 11.7 Summary of terminology for **95 % confidence level**

The **p-value** is the fractional area of the right tail of the **F-distribution** above the $\mathbf{F_{calculated}}$. $\mathbf{F_{calculated}}$ is indicated as simply an **F** in the Excel output. When the **p-value** $<$ α you are fundamentally decreasing the area in the tail and therefore increasing the acceptance region of H_a. On the other hand, when the **p-value** $>$ α you are increasing the area in the tail, therefore decreasing the acceptance area of H_a. To prove our results are real we want to minimize the rejection (tail) regions of the probability distribution.

The other test for statistical significance uses the same concepts as with the **p-value** comparison, but compares a **test statistic *critical*** with a **test statistic *calculated*.**

These test statistics are selected based on the type of problem you are trying to solve. Since our significance tests using the test statistics are based on relative positions between the critical value and the calculated value, it is recommended you work with the absolute values to avoid confusion. The rule to reject H_0 and establish statistical significance is given when the Excel calculated statistic is greater than the critical value of the statistic. So for example with the F-test

$$|\text{Fcalculated}| > |\text{Fcritical}|$$

Many reports include both types of tests since Excel often provides data to calculate both in the same output. However, the completion of only one of the analyses is sufficient to establish statistical significance.

◇ **Caution:** If you do run both types of tests and your results are not consistent, go back and check your input is correct and that you have run the Excel tools properly. With the same input data, mathematically the results should always give similar results.

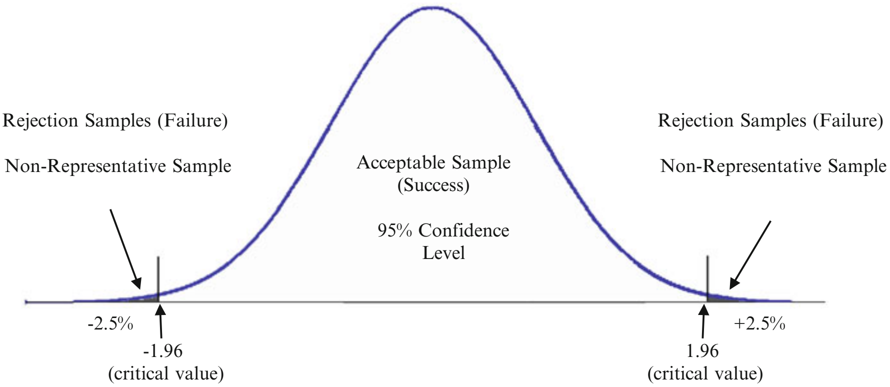


Fig. 11.8 Acceptable/rejection areas as a % and as a z-score in a normal distribution

The central area in Fig. 11.8 indicates the acceptable range of data for a 95 % confidence interval. The acceptable region corresponds to a z-score of ±1.96.

Remember if 95 % is in the central area, 5 % lies in the tails, but you have two tails so each tail must only have 2.5 % of the data. The corresponding confidence level is 95 % (0.95), with an alpha (α) level of 5 % (0.05); meaning, that in the long run, the probability of making an erroneous decision when H_0 is true will be fewer than five times in 100. You can think of the alpha areas or the tails as the failure region or the rejection region. The values that lie on the boundary of this rejection region are referred to as **critical values**.

Choosing the Appropriate Significance Test

Just because something is statistically significant doesn't mean that action should be taken in the direction indicated by the test. Other factors, such as the need of the application, economics, safety, etc. should be taken into account. Remember, significance tests only say that it's actually true.

⟐ Never allow statistics to overrule common sense.

The choice of the best technique depends on the type of questions to be answered, the number of variables and the measurement scale, sample size and how much is known about the population. It is important to determine if the solution requires a one-tailed or two-tailed analysis.

One-Tailed Tests

In this situation, we test whether one mean is higher than the other. We are only interested in one side of the probability distribution, which is illustrated in Fig. 11.9.

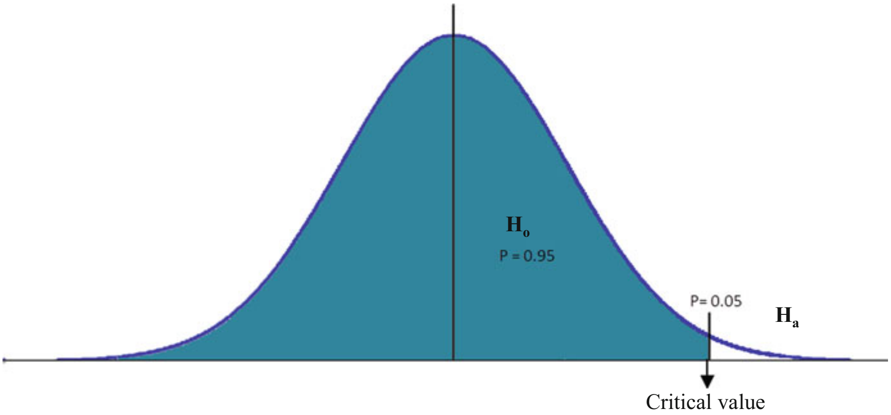


Fig. 11.9 One-tailed significance test

In this distribution, the shaded region shows the "success region" or where one mean is statistically significantly greater than the other (where the alternative hypothesis is true). The unshaded area represents the "failure region" where the first mean is not greater than the second mean or in other words where the alternative hypothesis holds true. Because we are only interested in one side of the distribution, or one "tail", this type of test is called a **one-sided** or a **one-tailed test**.

Two-Tailed Tests

A two-tailed test is used to determine if two means are different from one another. As an example, let's assume that we want to check if the pH of a stream has changed significantly in the past year. A useful, although not technically accurate, way to think about p-values is to think of them as the probability that the null hypothesis is true. Using this perspective, we generally believe the null hypothesis is true unless there is compelling evidence to think otherwise. When the probability that the null hypothesis is true is very low (the *p*-value is less than .05), we conclude that the null hypothesis is probably not true and therefore the alternate hypothesis is probably true. Figure 11.10 illustrates a region of acceptance or success as 95 % for a two-tailed test.

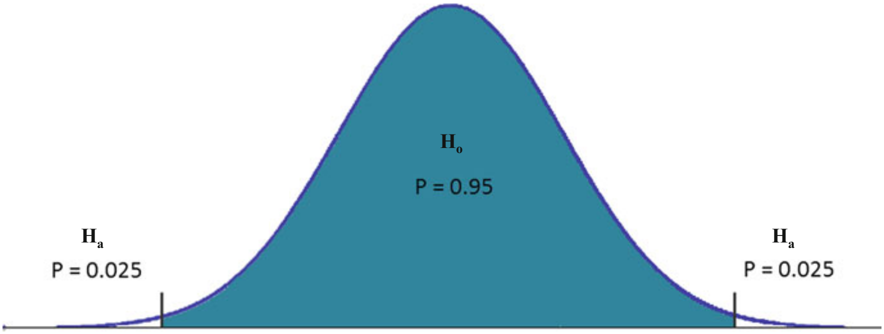


Fig. 11.10 Two-tailed test for 95 % confidence level

The rejection or failure region is 0.025 × 2 = 0.05. In other words this could represent a 95 % confidence interval. When we run our two tailed significance test, we need to have a calculated p value of less than that of 0.025 to be able to achieve statistical significance. Because we are interested in both sides of the distribution, or two “tails”, this type of test is called a **two-sided** or a **two-tailed test**.

In summary the problem will dictate how many tails are appropriate to include in the significance testing. Figure 11.11 compares the key difference in the design of the two versus one tailed tests.

Figure 11.12 summarizes the four basic significance tests that are explained in this book. The F-test and t-test are presented in this chapter, while the next two tests are discussed in Chap. 12.

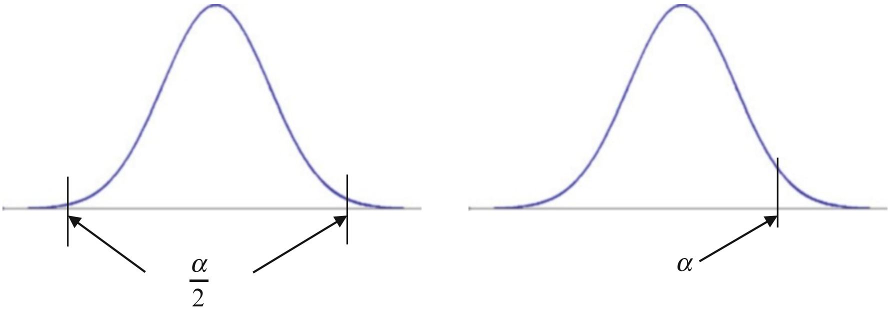
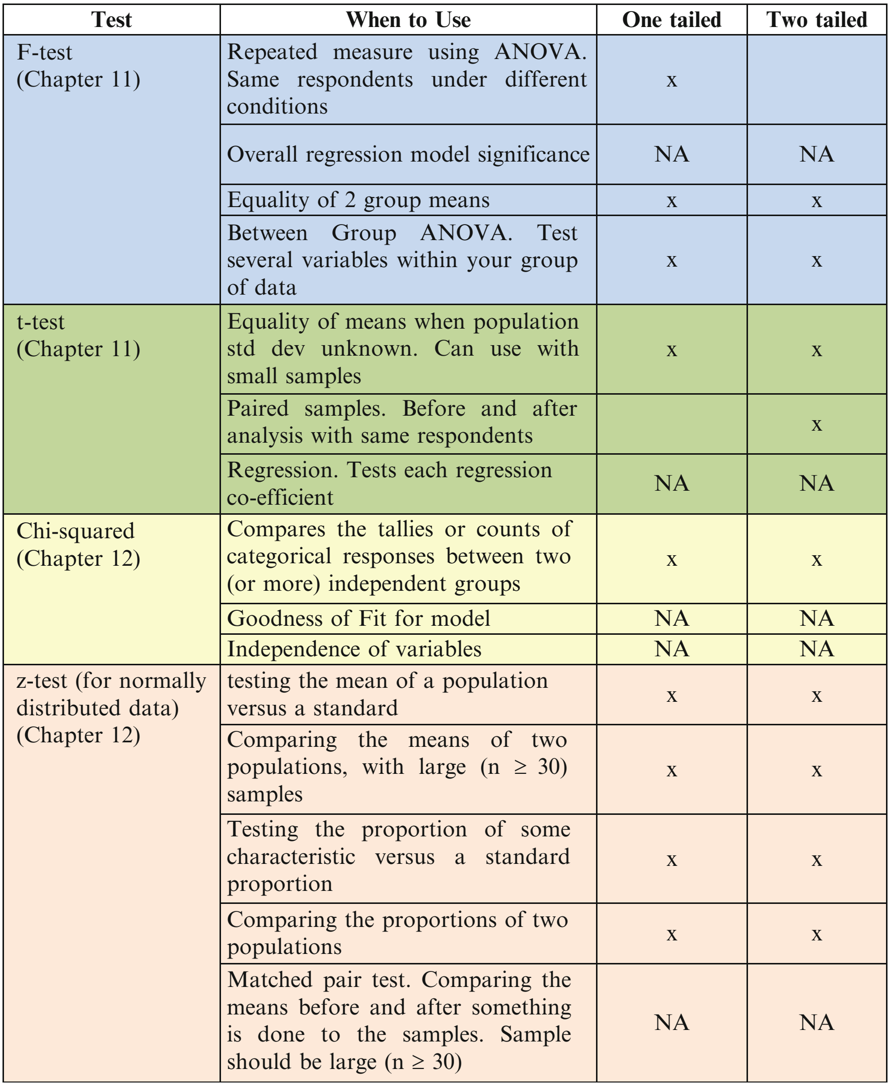


Fig. 11.11 One-tailed versus two-tailed test significance tests



Test	When to Use	One tailed	Two tailed
F-test (Chapter 11)	Repeated measure using ANOVA. Same respondents under different conditions	x	
	Overall regression model significance	NA	NA
	Equality of 2 group means	x	x
	Between Group ANOVA. Test several variables within your group of data	x	x
t-test (Chapter 11)	Equality of means when population std dev unknown. Can use with small samples	x	x
	Paired samples. Before and after analysis with same respondents		x
	Regression. Tests each regression co-efficient	NA	NA
Chi-squared (Chapter 12)	Compares the tallies or counts of categorical responses between two (or more) independent groups	x	x
	Goodness of Fit for model	NA	NA
	Independence of variables	NA	NA
z-test (for normally distributed data) (Chapter 12)	testing the mean of a population versus a standard	x	x
	Comparing the means of two populations, with large (n ≥ 30) samples	x	x
	Testing the proportion of some characteristic versus a standard proportion	x	x
	Comparing the proportions of two populations	x	x
	Matched pair test. Comparing the means before and after something is done to the samples. Sample should be large (n ≥ 30)	NA	NA

Fig. 11.12 Selecting the correct significance test

Significance Tests

The next two sections will define the **F-test** and the **t-test** with examples. The Excel tools will then be discussed for each of these significance tests.

F-test

The F-test Can Be Used Effectively Under Several Conditions

- **Example 1 (One-way Repeated Measures Using ANOVA):** In this case, each subject is measured several times under different conditions, and the results are measured. The repeated measure ANOVA (Analysis of Variance) **F-test** can be used to assess whether any of the conditions are on average different versus the null hypothesis, that all conditions yield the same mean response.
- **Example 2 (Regression Problems):** This is an important test when we run regression models. It is most often used when comparing statistical models that have been fit to a data set in order to identify the model that best fits the population from which the data were sampled.
- **Example 3 (F-test for Equality of Two Variances):** The **F-test** can also be used to test whether two population variances are equal. It does this by comparing the ratio of the two variances.
- **Example 4 (Between Group ANOVA):** Within the ANOVA tool in Excel, **F-tests** can be used to test for the significance of group means.

Excel provides F-test analysis within the **Regression** and **ANOVA** tools. Additional Excel tools include the **F-test** and the **F-test Two Sample for Variance**.

Rejection Rule

$$\text{Reject } H_0 \text{ if } F_{\text{calculated}} > F_{\text{critical}} \text{ or p-value}_{\text{calculated}} > \text{p-value}_{\alpha}$$

The **critical value of F (F_{critical})** can be thought of as the largest value to occur by chance for the given degrees of freedom. If the $\mathbf{F_{calculated}}$ is larger than $\mathbf{F_{critical}}$, then the null hypothesis of equivalent group means is rejected. In other words your results are not simply due to chance and are statistically significant. These statistics are depicted in Fig. 11.13.

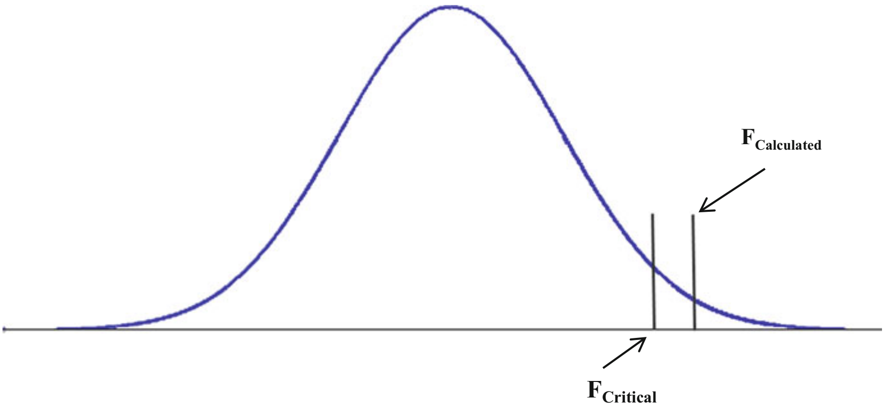


Fig. 11.13 Acceptance of H_a when $F_{\text{calculated}}$ is greater F_{critical}

The **p-value** is the probability that a value of **F** greater than or equal to the $\mathbf{F_{calculated}}$ could have occurred by chance if there were no difference in the means. The **p-value** is the fractional area of the right tail of the **F-distribution** above the $\mathbf{F_{calculated}}$. $\mathbf{F_{calculated}}$ is indicated as simply an **F** in the Excel output. When the **p-value** is less than your chosen confidence level, you have statistically significant results so keep going!

Basic Descriptions of F-Test Applications

This first section will describe the general application of each problem type by way of an example and the second section will walk through the step by step Excel applications.

Example 1: One-Way Repeated Measures Using ANOVA

The **F-test**, in one-way analysis, is used to assess whether the expected values of a quantitative variable within a sample differs from the other sample. In this case each subject is measured several times under different conditions. For example, suppose that a medical trial compares four treatments. The subjects are given each of the four treatments and the results are measured. The ANOVA **F-test** can be used to assess whether any of the treatments is on average superior, or inferior, to the others versus the null hypothesis that all four treatments yield the same mean response. This is an example of an "omnibus" test, meaning that a single test is performed to detect if there are any of several possible differences that are real differences, not just due to chance. Alternatively, we could carry out pairwise tests among the treatments. For instance, in the medical trial example with four treatments we could carry out six tests among pairs of treatments.

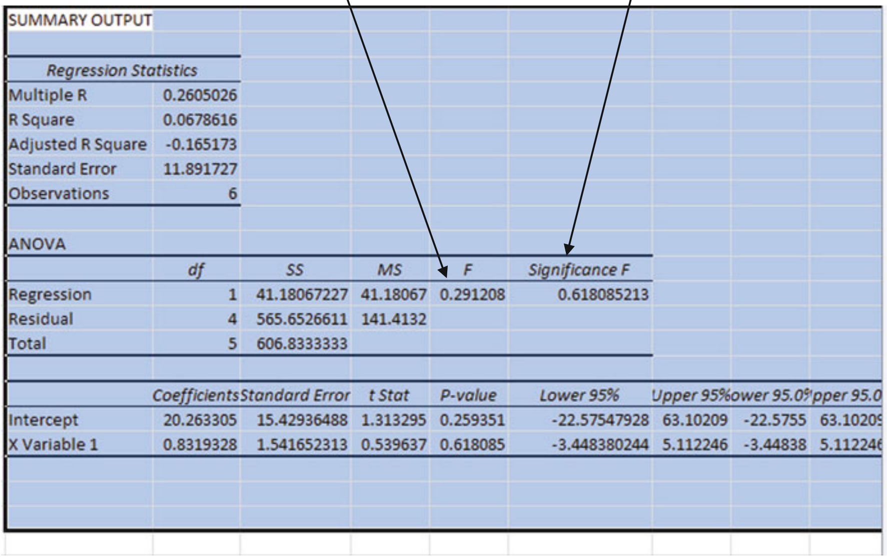
The advantage of the ANOVA **F-test** is that we do not need to pre-specify which treatments are to be compared, and we do not need to adjust for making multiple comparisons. The disadvantage of the ANOVA **F-test** is that if we reject the null hypothesis, we do not know which treatments can be said to be significantly different from the others; if the **F-test** is performed at level α we cannot state that the treatment pair with the greatest mean difference is significantly different at level α. We just know there is some real difference in the samples.

Example 2: Regression Problems

This tests the significance of the overall regression model. The overall model is defined by

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_0$$

In Fig. 11.14, we have included only one independent variable for simplicity in the explanation of significance. Under the ANOVA heading (Analysis of Variance), Excel does not provide the **F$_{\textbf{critical}}$,** so we can only use the **Significance F** value. The **Significance F** value is really just the **p-value**, the probability associated with the **F$_{\textbf{calculated}}$** of 0.291208. In other words there is a 0.618 probability of getting an F-value greater than 0.29. This is where the confidence level or the alpha value comes in. Select the desired level for risk or alpha value. Let's select a 95 % confidence level, so we have an alpha of 0.05. Now, compare the p-value with the selected alpha value; since the **p-value** (**Significance F)** of 0.618 is greater than 0.05, we do not have statistically significant results at the 95 % confidence level. Our regression model will not yield useful results the way it is currently set up. So back to the drawing board and rethink the model.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.2605026
R Square	0.0678616
Adjusted R Square	-0.165173
Standard Error	11.891727
Observations	6

ANOVA

	df	SS	MS	F	Significance F
Regression	1	41.18067227	41.18067	0.291208	0.618085213
Residual	4	565.6526611	141.4132		
Total	5	606.8333333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0
Intercept	20.263305	15.42936488	1.313295	0.259351	-22.57547928	63.10209	-22.5755	63.10209
X Variable 1	0.8319328	1.541652313	0.539637	0.618085	-3.448380244	5.112246	-3.44838	5.112246

Fig. 11.14 Summary output

Example 3: F-Test for Equality of Two Variances

The associated Excel tool is the **F-test Two-Sample for Variances** located in the **Data Analysis** tab. For example, you can use this **F-test** tool on samples of productivity for two corporate departments. The tool provides the results of a test of the null hypothesis that these two samples come from distributions with equal variances versus the alternative hypothesis that the variances are not equal in the underlying distributions. In simple terms, we confirm or reject the idea that the difference in the results of productivity of these two departments is indeed real and not just due to chance. A value of F close to 1 provides evidence that the underlying population variances are equal, usually not what we want to hear.

This test can be a two-tailed or a one-tailed F-test. The two-tailed version returns the two-tailed probability that the variances in the two samples are not significantly different. If the probability calculated in **F.Test** is greater than your alpha value then the difference in the variances is not statistically significant. The one-tailed version only tests in one direction that the variance from the first population is either greater than or less than (but not both) the second population variance. The choice is determined by the problem. For example, if we are testing a new process, we may only be interested in knowing if the new process is less variable than the old process. For example the analysis can be used to answer the following questions:

- Does a new process, treatment, or test reduce the variability of the current process? (one-tailed)
- Given satisfaction scores from accountants and engineers, do these professions have different levels of satisfaction? (two-tailed)

The Excel **F.TEST** function also returns two-tailed results. The syntax of the Excel **F.Test** function is:

$$\mathbf{F.TEST(array1, array2)}$$

Where the supplied **array1** and **array2** are two samples of data values.

Example 4: Between Group ANOVA

You may find the need to analyze more than one independent variable, or independent variables that have more than two levels. The ANOVA tools which use the **F- statistic** allow for analysis in the differences between the group means in this situation.

- **One-way Between Group ANOVA**- the subject is only measured once. For example you might have data on tire performance for many tire types. You have an overall performance rating and you have data on performance in foul weather conditions. You are interested in seeing if foul weather performance is related to the overall performance rating. ANOVA allows you to break up the group according to the overall performance and then see if foul weather performance is indeed significantly different across overall performance numbers.
- **Two-way Between Group ANOVA**- the subject is only measured once but you can analyze two factors simultaneously. For example: you are interested in the role of gender and age on consumer perception rating of truck commercials during the Super Bowl.

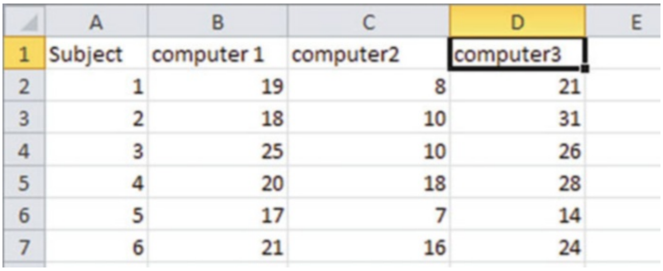
Excel

Excel examples will be provided for each of the four F-tests problem types:

- Example 1: F-Test One-way Repeated Measures using ANOVA
- Example 2: Regression
- Example 3: F-test for Equality of two variances
- Example 4: Between Group ANOVA

Example 1: One-Way Repeated Measures Using ANOVA

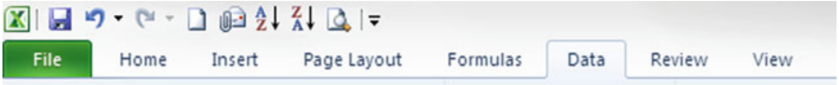
Six office assistants are selected to help choose the best replacement computers for the support staff. The computers have various size screens and different styles of keyboards. We need to select the computer that allows the subjects the highest level of accuracy in word documents, so their proficiency scores are recorded for each of the three computers. We need to find the highest scoring computer but we also need to make sure these results are significant at the 0.05 confidence level. Enter the data from Fig. 11.15.



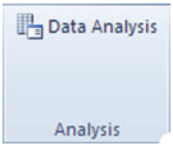
	A	B	C	D	E
1	Subject	computer 1	computer2	computer3	
2	1	19	8	21	
3	2	18	10	31	
4	3	25	10	26	
5	4	20	18	28	
6	5	17	7	14	
7	6	21	16	24	

Fig. 11.15 Input

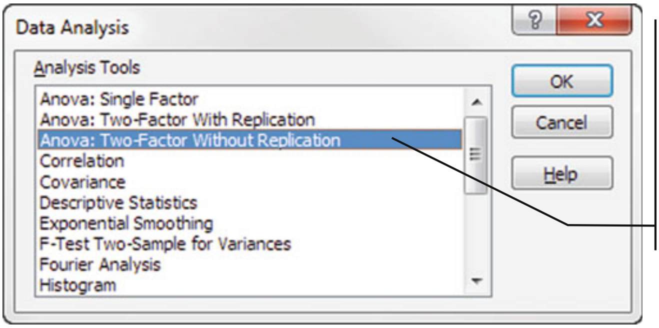
➢ Click on the **Data** tab



➢ Click on the **Data Analysis** function

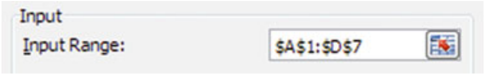


- From the menu select **ANOVA: Two-Factor Without Replication**.



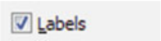
Note: "Without Replication" refers to n=1 subject per cell. "Two-Factor" refers to the subjects and computers.

- Click **OK**
- Click inside the **Input Range** box and highlight the data input area
  - Make sure you include the labels on the rows and the columns to keep track of the results



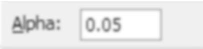
Input
Input Range: $A$1:$D$7

- Check the **Labels** box



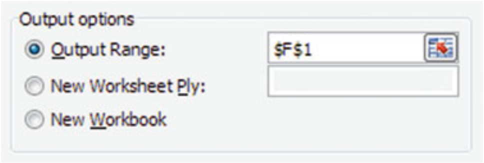
☑ Labels

- Input an alpha value of **0.05**



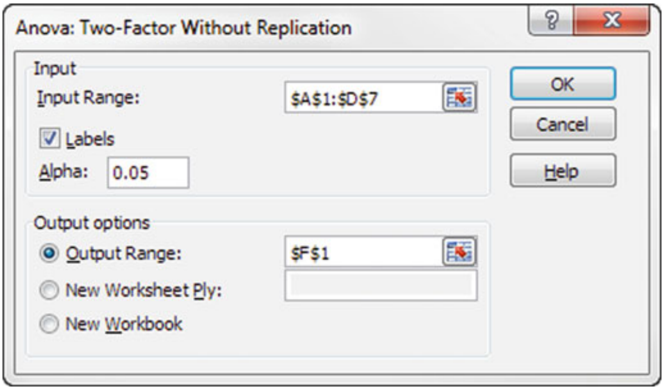
Alpha: 0.05

- Select the **Output Range**, click inside box and highlight the cell where you want the output printed



Output options
● Output Range: $F$1
○ New Worksheet Ply:
○ New Workbook

- Click **OK**



Anova: Two-Factor Without Replication

Input

Input Range: $A$1:$D$7

☑ Labels

Alpha: 0.05

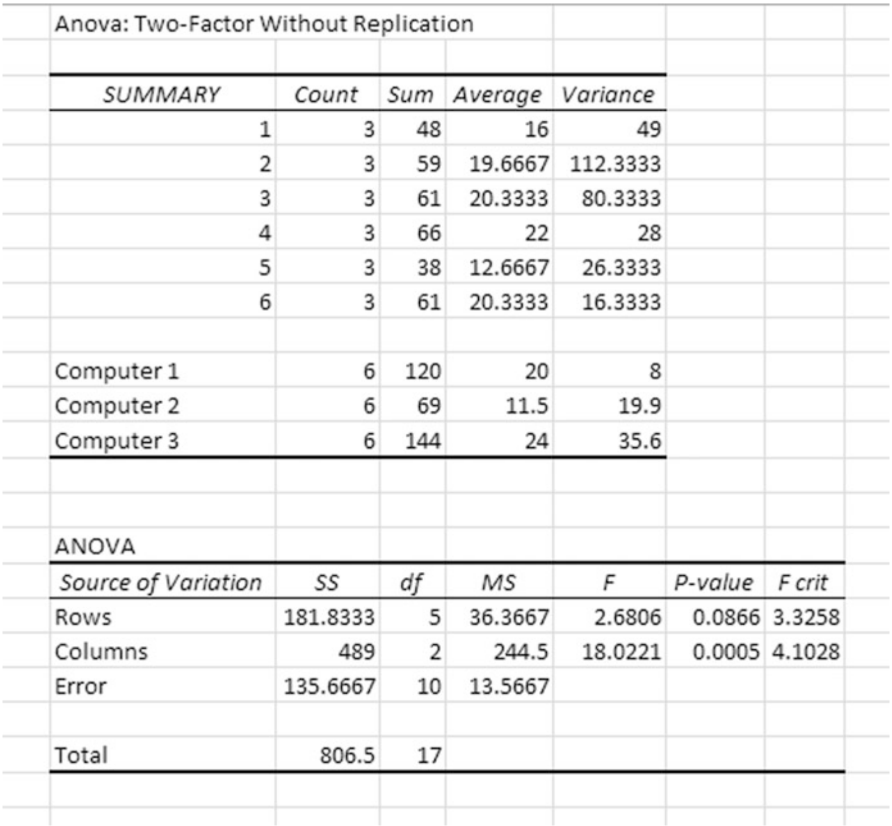
Output options

● Output Range: $F$1

○ New Worksheet Ply:

○ New Workbook

The output as shown in Fig. 11.16 provides summary information about the subjects and the proficiency scores.



Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
1	3	48	16	49
2	3	59	19.6667	112.3333
3	3	61	20.3333	80.3333
4	3	66	22	28
5	3	38	12.6667	26.3333
6	3	61	20.3333	16.3333
Computer 1	6	120	20	8
Computer 2	6	69	11.5	19.9
Computer 3	6	144	24	35.6

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	181.8333	5	36.3667	2.6806	0.0866	3.3258
Columns	489	2	244.5	18.0221	0.0005	4.1028
Error	135.6667	10	13.5667			
Total	806.5	17				

Fig. 11.16 Output

The **Count** should be the same as the number of conditions the subject was tested under; in this case each subject tried three computers. In the bottom of this column the number of times each computer was used (6) is provided.

The **Sum** represents the sum of each person's score on all three computers. At the bottom of this column is the sum of all three different people's scores on this same computer; the total score is by computer not person here.

The **Average** represents the mean scores by office assistant and by computer.

The **Variance** represents the variance in the scores by office assistant and by computer.

In the **ANOVA** table we are interested in two sets of results in the **Columns** row for significance testing. The **Columns** row refers to the different computer outcomes which are what we are interested in understanding.

We can test for significance in two ways:

1. We can compare the **P-value** (.0005) for **Columns** with the alpha value we selected (0.05). Since the **P-value** is less than alpha we can confirm that the difference in the computers scores is real at the 95 % confidence level.
2. We can compare the **F** (18.0221) with the **F crit** (4.1028). Since the **F** (**F**$_{\mathbf{calculated}}$ is shown as **F** in the ANOVA table) is greater than the **F crit**, we can conclude our results are in fact statistically significant at the 95 % confidence level and not just due to chance.

   ◈ If you do run both types of tests and your results are not consistent, go back and check your input is correct and you have run the Excel tools properly. With the same input data, mathematically the results should always give similar results.

Our department should select Computer 1 as it had the high proficiency scores with low variance. If we had selected another 100 samples of six office assistants from the population, we would expect similar results in 95 of those samples.

Example 2: Regression

A sample of office assistants across the San Francisco-Bay Area was created. It was suggested that the data should be analyzed with a regression model to determine if salary was related to years of experience and aptitude test scores. Use a 99 % confidence level. This data is shown in Fig. 11.17.

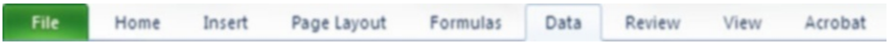
Experience(yrs) X_1	Score (out of 100) X_2	Salary ($000) Y
4	78	24
7	100	43
1	86	23.7
5	82	34.3
8	86	35.8
10	84	38
0	75	22.2
1	80	23.1
6	83	30
6	91	33
9	88	38
2	73	26.6
10	75	36.2
5	81	31.6
6	74	29
8	87	34
4	79	30.1
6	94	33.9
3	70	28.9
3	89	30

Fig. 11.17 Survey data

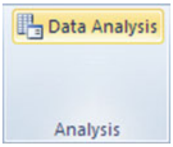
◈ There should only be one label line in Excel otherwise Excel will read the extra label lines as numbers and give you an error.

➢ Input the data from Fig. 11.17

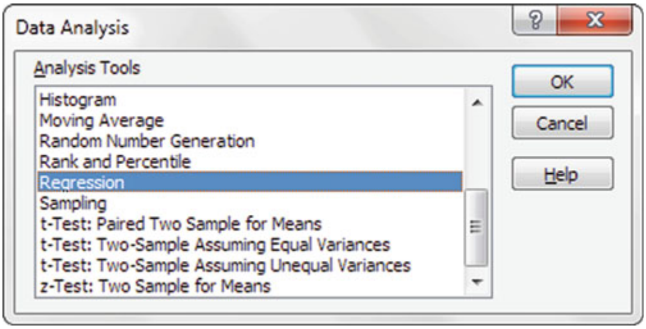
➢ Click the **Data** Tab



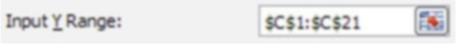
➢ Choose the **Data Analysis** function



- ➢ Choose **Regression from the menu**

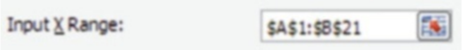

Data Analysis
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

- ➢ **Input y Range**



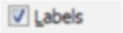
Input Y Range: $C$1:$C$21

- ➢ **Input x Range**



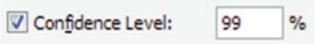
Input X Range: $A$1:$B$21

- ➢ Select **Labels box**



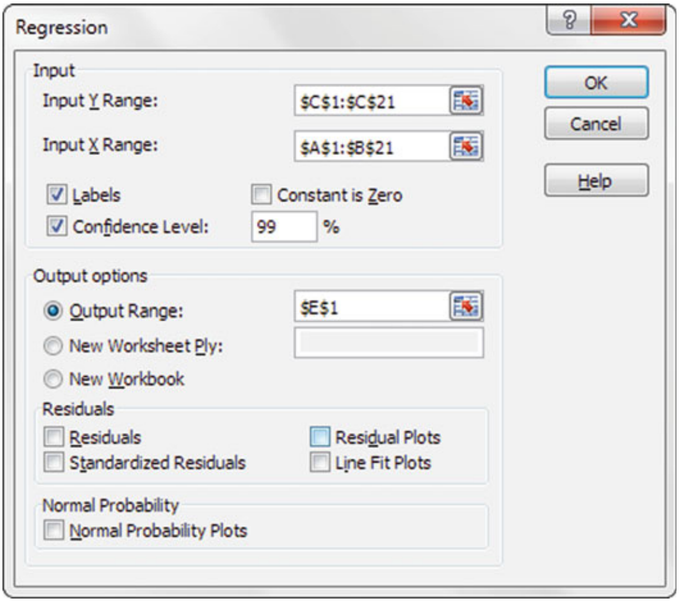
☑ Labels

- ➢ Select **Confidence Level** and enter 99 in **Confidence Level** box



☑ Confidence Level: 99 %

- ➢ Select **Output Range** and click inside the **Output Range** box in order to highlight the cell where you want the output printed
- ➢ Click **OK**



Input Y Range: $C$1:$C$21

Input X Range: $A$1:$B$21

☑ Labels □ Constant is Zero

☑ Confidence Level: 99 %

● Output Range: $E$1

○ New Worksheet Ply:

○ New Workbook

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

□ Normal Probability Plots

The Excel output is provided in Fig. 11.18. Remember the form of the equation will be

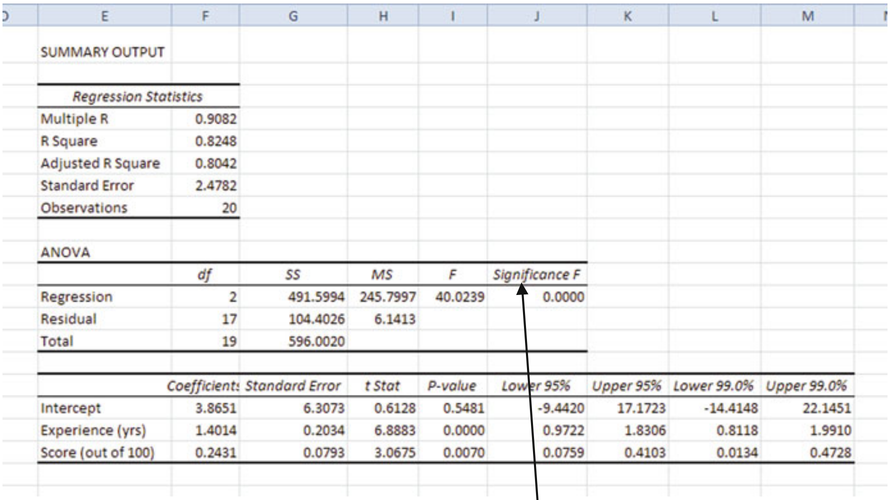
$$\widehat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Where:

y = annual salary ($000)
x_1 = years of experience
x_2 = score on aptitude test

The resulting equation for this model based on the ANOVA results is

$$y = 3.8651 + 1.4014\ x_1 + 0.2431\ x_2$$



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9082
R Square	0.8248
Adjusted R Square	0.8042
Standard Error	2.4782
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	2	491.5994	245.7997	40.0239	0.0000
Residual	17	104.4026	6.1413		
Total	19	596.0020			

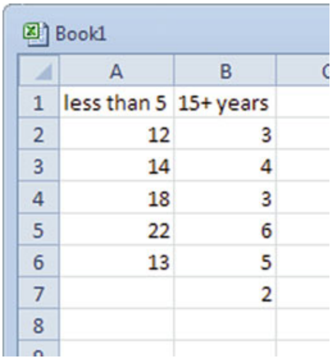
	Coefficient:	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	3.8651	6.3073	0.6128	0.5481	-9.4420	17.1723	-14.4148	22.1451
Experience (yrs)	1.4014	0.2034	6.8883	0.0000	0.9722	1.8306	0.8118	1.9910
Score (out of 100)	0.2431	0.0793	3.0675	0.0070	0.0759	0.4103	0.0134	0.4728

Fig. 11.18 Excel output

To run significance analysis we compare the **Significance F** (p-value) with the alpha value we have selected (0.01). Remember that a confidence level of 99 % is the same as an **α** value of 0.01. The **Significance F** is less than 0.01 so the results are significant at the 99 % confidence level. In other words, if we selected 100 samples from this population, in 99 of those analyses we would get similar results. We are now confident that the sample data has allowed us to build a regression model that we can use for predictive purposes. Remember this does not tell us anything about the individual x values, but tells us about the significance of the overall model.

Example 3: Two Sample for Variances

In this application we can use the one-tailed or two-tailed versions in Excel depending on the model we are trying to test. Let's use the same example but change the confidence level to 95 %.



	A	B
1	less than 5	15+ years
2	12	3
3	14	4
4	18	3
5	22	6
6	13	5
7		2
8		

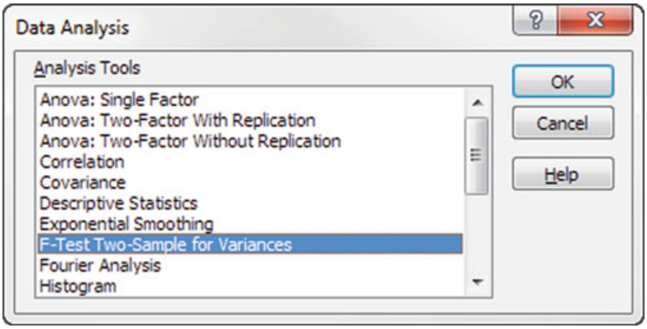
Fig. 11.19 Input

One-Tailed F-Test for Two Sample for Variances

The one-tailed version only tests in one direction that the mean from the first population is either greater than or less than (but not both) the second population mean. The two-tailed version tests against the alternative that the means are not equal. The choice is determined by the problem. For example, suppose that the two groups of employees report on the time they spend handling customer complaints. The first group has less than 5 years of experience, and the second group has at least 15 years of experience. To keep things simple we will assume that based on the original data set we have calculated that the two standard deviations are different and now we want to test their significance; group 1 has sd = 1.47 and group 2 has sd = 4.15. However, there is always the possibility that this difference in the standard deviations could have occurred by chance. We need to test if these results are significant at a 95 % confidence level. A one-sided **F-test** is appropriate in this scenario because of the claim that one sample standard deviation is "smaller than" another sample standard deviation. Input the data from Fig. 11.19.

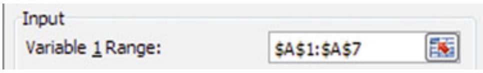
◇ You must put the group with the larger variance in the left hand column, i.e. variable1.

➢ Input the data from Fig. 11.19
➢ Click on the **Data** tab and select **Data Analysis**
➢ Select **F-test Two Sample for Variances** from the drop down menu



➢ Click **OK**

➢ Click inside the **Variable 1 Range box**. Highlight the first column of data including the label.

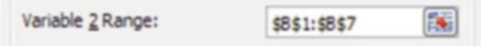


Input

Variable 1 Range: $A$1:$A$7

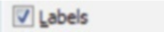
◈ Note the actual range is A1:A6 but Excel will assume a zero value if no data appears. By including A7 we keep the input field symmetric with the second data set input field. You will get the same results whichever approach you use.

➢ Click inside the **Variable 2 Range box**. Highlight the second column of data including the label.



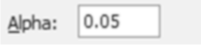
Variable 2 Range: $B$1:$B$7

➢ Check the **Labels** box



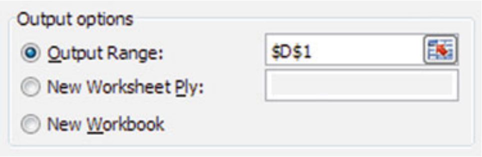
☑ Labels

➢ Select an alpha level of 0.05.



Alpha: 0.05

➢ Click inside the **Output Range box**. Highlight the cell where you want the output.



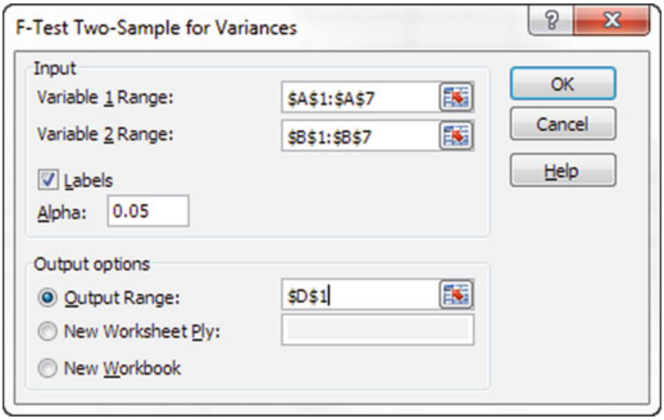
Output options

● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

➢ Click **OK**



F-Test Two-Sample for Variances

Input

Variable 1 Range: $A$1:$A$7

Variable 2 Range: $B$1:$B$7

☑ Labels

Alpha: 0.05

Output options

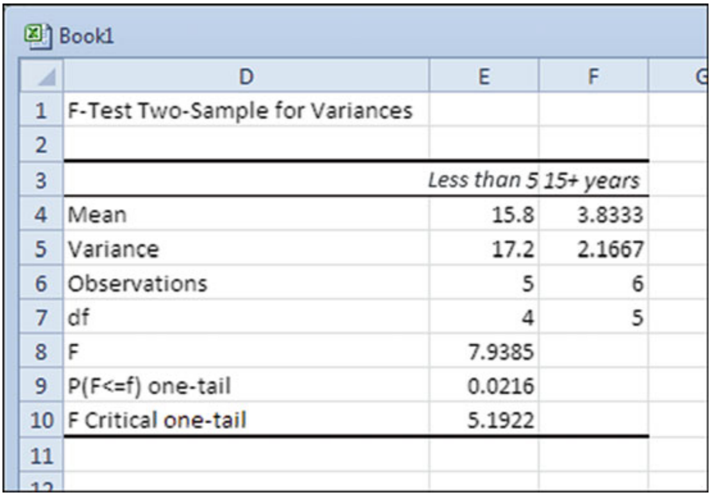
● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

The output data is provided in Fig. 11.20

Make sure you check all of the data have been input by looking at the **Observations**. The **Observations** indicate we have included all of our data (5,6).



Book1

F-Test Two-Sample for Variances

	Less than 5	*15+ years*
Mean	15.8	3.8333
Variance	17.2	2.1667
Observations	5	6
df	4	5
F	7.9385	
P(F<=f) one-tail	0.0216	
F Critical one-tail	5.1922	

Fig. 11.20 Output

The **Mean** and **Variance** are given for each group.

We can use two sets of results for significance testing:

1. We can compare the **P(F <= f) one tail** value (0.0216) with the alpha value we selected (0.05). Since the P value is less than alpha we confirm that the difference in the variance is statistically significant at the 95 % confidence level.
2. We can compare the **F** (7.9385) with the **F Critical one-tail** (5.1922). Since the **F** is greater than the **F Critical one-tail**, we can conclude our results are statistically significant at the 95 % confidence level. If we selected 100 other samples from this population we would get similar results in 95 of those samples.

The variance in satisfaction scores is statistically different; we can assume the less experienced employees really do have a greater standard deviation than those more experienced employees. In other words there is a greater variability amongst the members of the less experienced group in terms of how many complaints they handle. There is less variability in the results of the more experienced group.

Two-Tailed F-Test for Equality of Two Variances

The **two-tailed F-test** determines if one population variance (or standard deviation) is statistically significantly different than another population variance. We have no knowledge nor do we care if one is greater or less than the other, just that they are significantly different.

Two groups of marketing professionals provided the client satisfaction scores out of 100 that they received during the past 12 months. Is the variance in these scores real or simply due to chance at the 95 % confidence level?

◇ Because the Excel **F-test** function was written as a one-sided test, three adjustments are required to use it as a two-sided test

1. *You must put the group with the larger variance in the first column, i.e. variable 1. If you forget to do this **Fcalc (F)** will be less than one, and you need to re-enter the data.*
2. *You divide the alpha value in half when you input into Excel.*
3. *The p-value reported by Excel should be doubled to give the correct p-value for the two-sided hypothesis test.*

◇ **Note:** It is not necessary that both sample sizes be the same.

The two sets of data whose variances are to be compared should be entered in consecutive rows of two columns.

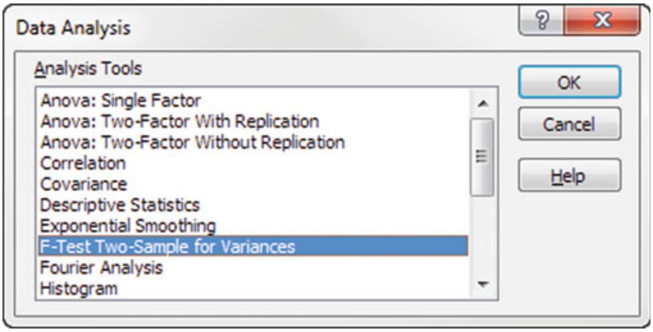
➢ Input the data from Fig. 11.21

Fig. 11.21 Input

grp 1	grp 2
90.47	90.95
92.02	91.30
93.15	91.48
90.98	92.04
91.73	90.70
	91.33

➢ Click on the **Data** tab and select **Data Analysis**

➢ Select **F-Test Two-Sample for Variances** from the drop down menu

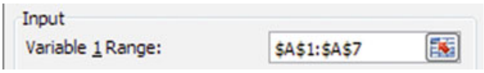

Data Analysis

Analysis Tools

F-Test Two-Sample for Variances

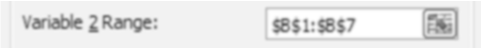
➢ Click **OK**

➢ Click inside the **Variable 1 Range box**. Highlight the first column of data including the label.

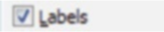

Input

Variable 1 Range: $A$1:$A$7

➢ Click inside the **Variable 2 Range box**. Highlight the second column of data including the label.

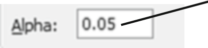

Variable 2 Range: $B$1:$B$7

- ➢ Check the **Labels** box



☑ Labels

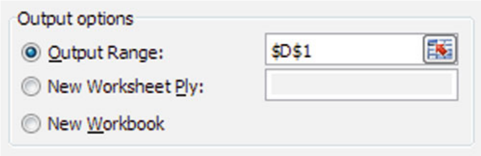
- ➢ Select an alpha level of 0.025.



Alpha: 0.05

Note: Because you want a 2-tailed test you need to divide this value (0.05) by 2 and **input 0.025**

- ➢ Click inside the **Output Range box**. Highlight the cell where you want the output.



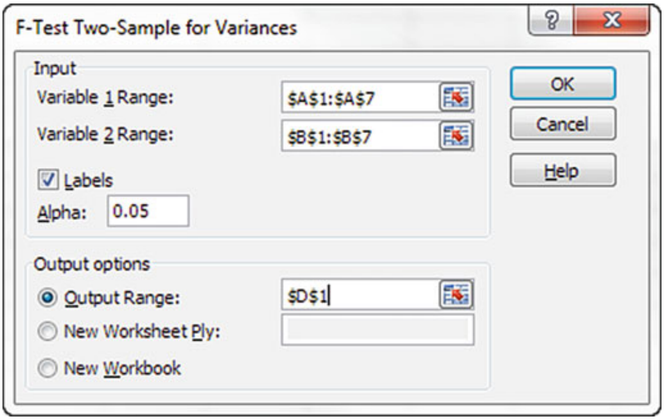
Output options

● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

- ➢ Click **OK**



F-Test Two-Sample for Variances

Input

Variable 1 Range: $A$1:$A$7

Variable 2 Range: $B$1:$B$7

☑ Labels

Alpha: 0.05

Output options

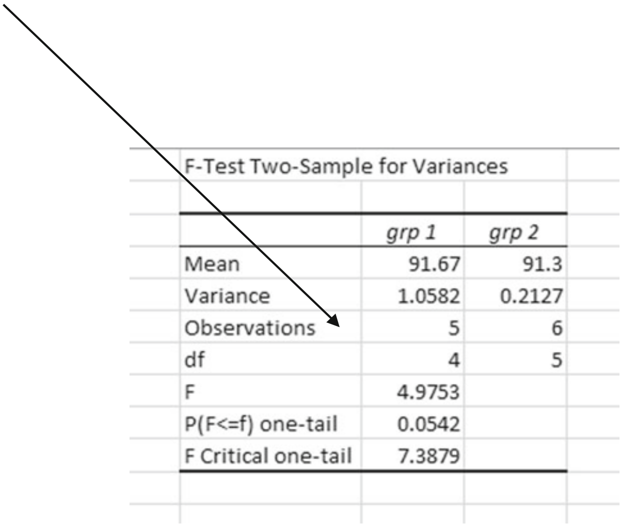
● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

The output data is presented in Fig. 11.22.

Make sure you check all of the data have been input by looking at the **Observations**.



F-Test Two-Sample for Variances		
	grp 1	*grp 2*
Mean	91.67	91.3
Variance	1.0582	0.2127
Observations	5	6
df	4	5
F	4.9753	
P(F<=f) one-tail	0.0542	
F Critical one-tail	7.3879	

Fig. 11.22 Output

The **Mean** and **Variance** are given for each group.

We can use two sets of results for significance testing:

1. Before we can do a comparison, we need to double the **P(F <= f) one tail** to get 2 × 0.0542 = 0.1084. We can now compare this value with the alpha value we selected (0.05). Since the **P value** is greater than alpha we confirm that the difference in the variance is not real at the 95 % confidence level.
2. We can compare the **F** (4.9753) with the **F Critical one tail** (7.3879). Since the **F is** less than the **F Critical one tail**, we can conclude our results are not statistically significant at the 95 % confidence level. If we selected other samples we would not get similar results.

Example 4: Between Group ANOVA

In this application we can use the one-tailed or two-tailed versions in Excel depending on the model we are trying to test.

One-Tail F-Test Between Group ANOVA

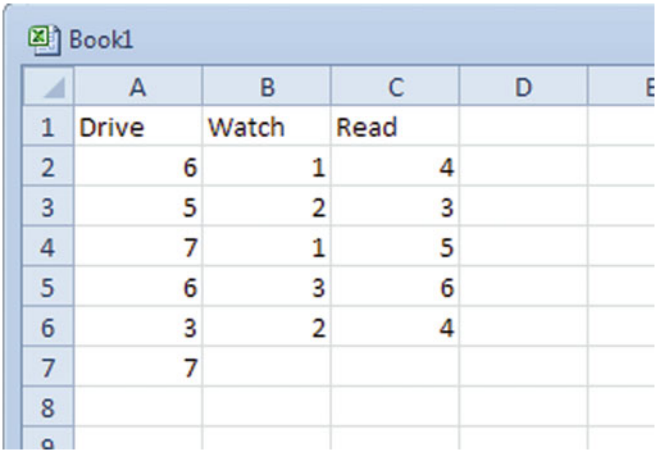
◈ **Note:** No need to have an equal number of cells for this analysis

Sixteen customers were assigned to one of three conditions in which they were given truck performance data. The three conditions were:

Group 1: They were allowed to test drive the actual truck that was described in the performance data sheet.
Group 2: They watched a video on how these trucks are assembled.
Group 3: They were given comparison data on competitor trucks to read.

The customers were then asked to view a commercial on the truck and rate their perceptions of the truck. They used a 7- point Likert scale with 1 as the lowest score. You are interested in knowing if any of these conditions really influenced the perception rating scores with 95 % confidence.

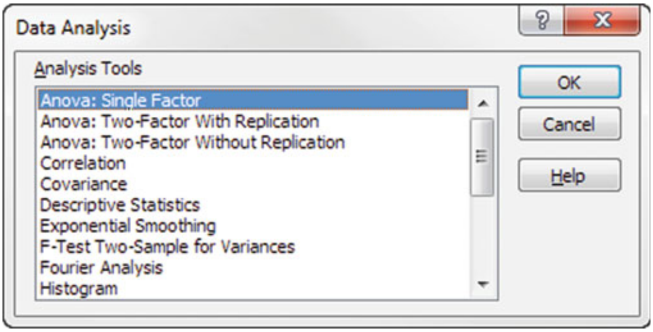
➢ Input the data from Fig. 11.23



	A	B	C	D	E
1	Drive	Watch	Read		
2	6	1	4		
3	5	2	3		
4	7	1	5		
5	6	3	6		
6	3	2	4		
7	7				
8					
9					

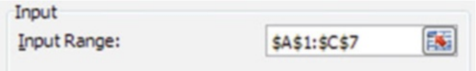
Fig. 11.23 Input

➢ Click on the **Data** tab and click on **Data Analysis** function
➢ From the menu select **ANOVA: Single-Factor**



➢ Click **OK**

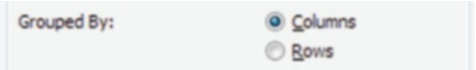
- ➢ Click inside the **Input Range** box and highlight the data input area. Make sure you include the labels to better keep track of the results.



Input

Input Range: $A$1:$C$7

- ➢ In the **Grouped by** box select **Columns**



Grouped By: ● Columns ○ Rows

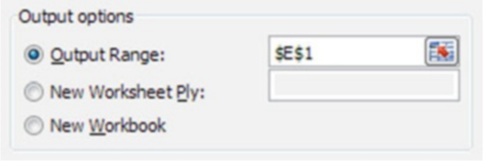
- ➢ Check the **Labels in first row** box

☑ Labels in first row

- ➢ Input an alpha value of **0.05**. This corresponds to a 95 % confidence level.

Alpha: 0.05

- ➢ Select **Output Range** and click inside the **Output Range** box in order to highlight the cell where you want the output printed



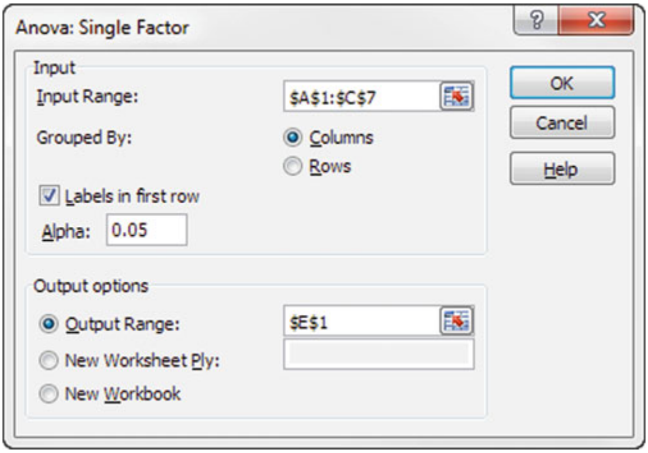
Output options

● Output Range: $E$1

○ New Worksheet Ply:

○ New Workbook

- ➢ Click **OK**



Anova: Single Factor

Input

Input Range: $A$1:$C$7

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Alpha: 0.05

Output options

● Output Range: $E$1

○ New Worksheet Ply:

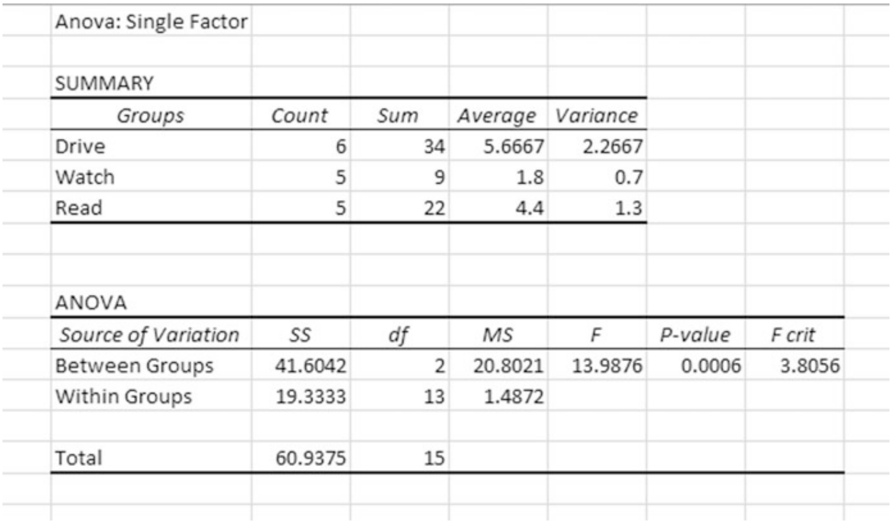
○ New Workbook

The results are provided in Fig. 11.24. Make sure the **Count** adds up to your number of participants (16). The Sum, Average and Variance columns provide the statistics for each condition (category).

The ANOVA table **Between Groups** refers to how differently the three groups ranked the truck.

While the **Within Groups** refers to how differently the customers within each separate group differed from one another.

To determine the significance of these results we can look in two places (Fig. 11.24).



Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Drive	6	34	5.6667	2.2667
Watch	5	9	1.8	0.7
Read	5	22	4.4	1.3

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	41.6042	2	20.8021	13.9876	0.0006	3.8056
Within Groups	19.3333	13	1.4872			
Total	60.9375	15				

Fig. 11.24 Output

1. We can compare the **p-value** (0.0006) with the alpha value we selected (0.05). Since the **p-value** is less than alpha we can confirm that the difference in average scores for each group is real at the 95 % confidence level.
2. We can compare the **F** (13.98761) with the **F crit** (3.8056). Since the **F is** greater than the **F crit**, we can conclude our results are in fact statistically significant at the 95 % confidence level and not just due to chance.

It seems that the group of customers that got to drive the truck had the highest mean perception score about the truck. So if the company really wanted to do a great job marketing perhaps they need to offer more test drives with their vehicles. An important message for the dealerships!

Two-Tail F-Test Between Group ANOVA

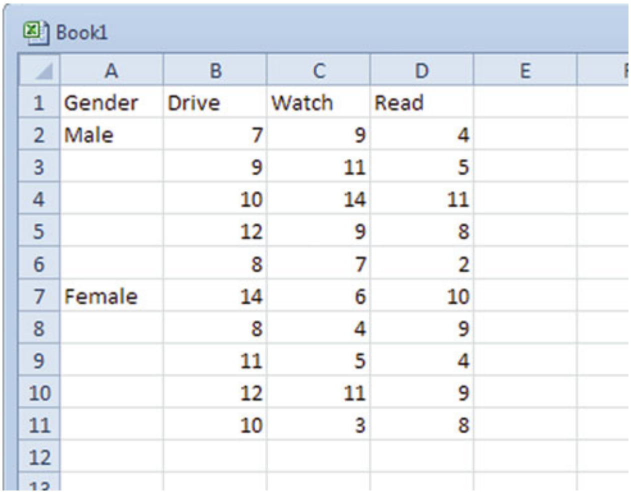
◇ **Note:** You must have an equal number of cells for this analysis.

Thirty customers were assigned to one of three conditions in which they were given a sheet of truck performance data. The three conditions were:

Group 1: They were allowed to test drive the actual truck that was described in the performance data sheet.
Group 2: They watched a video on how these trucks are assembled.
Group 3: They were given comparison data on competitor trucks to read.

The customers were then asked to view a commercial on the truck and rate their perceptions of the truck. They used a 7- point Likert scale with 1 as the lowest score. You are interested in knowing if any of these conditions *as well as their gender* influenced the perception rating scores with 95 % confidence. Now there are two factors so we need to use the **Two-way Between-group ANOVA**

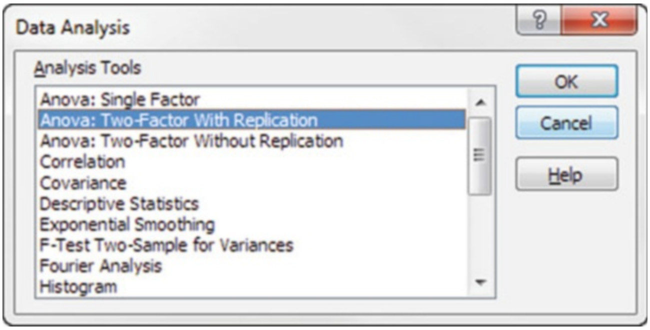
➢ Input the data from Fig. 11.25



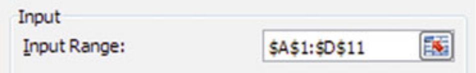
	A	B	C	D	E
1	Gender	Drive	Watch	Read	
2	Male	7	9	4	
3		9	11	5	
4		10	14	11	
5		12	9	8	
6		8	7	2	
7	Female	14	6	10	
8		8	4	9	
9		11	5	4	
10		12	11	9	
11		10	3	8	
12					

Fig. 11.25 Input

➢ Click on the **Data** tab and click on **Data Analysis** function
➢ From the drop down menu select **ANOVA: Two-Factor With Replication**

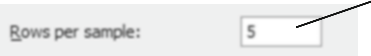


- ➢ Click **OK**
- ➢ Click inside the **Input Range** box and highlight the data input area. Make sure you include the labels to better keep track of the results


Input

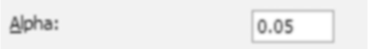
Input Range: $A$1:$D$11

- ➢ In the **Rows per sample** box input **5**. Remember for this Excel function you must have an equal number of rows for each group.

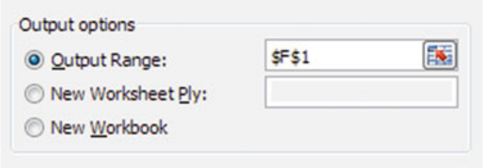

Rows per sample: 5

__Note__: There is no **Labels box** to check because Excel assumes you have included the labels for both factors in your input range. If you have forgotten to include the labels Excel will treat the first row of data as labels. There are 5 rows of input data.

- ➢ Input an alpha value of **0.05** (95 % confidence level)


Alpha: 0.05

- ➢ Select **Output Range** and click inside the **Output Range** box in order to highlight the cell where you want the output printed

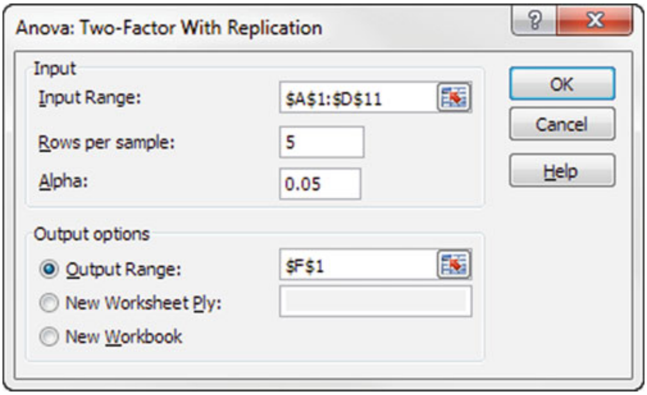

Output options

● Output Range: $F$1

○ New Worksheet Ply:

○ New Workbook

- ➢ Click **OK**


Anova: Two-Factor With Replication

Input

Input Range: $A$1:$D$11

Rows per sample: 5

Alpha: 0.05

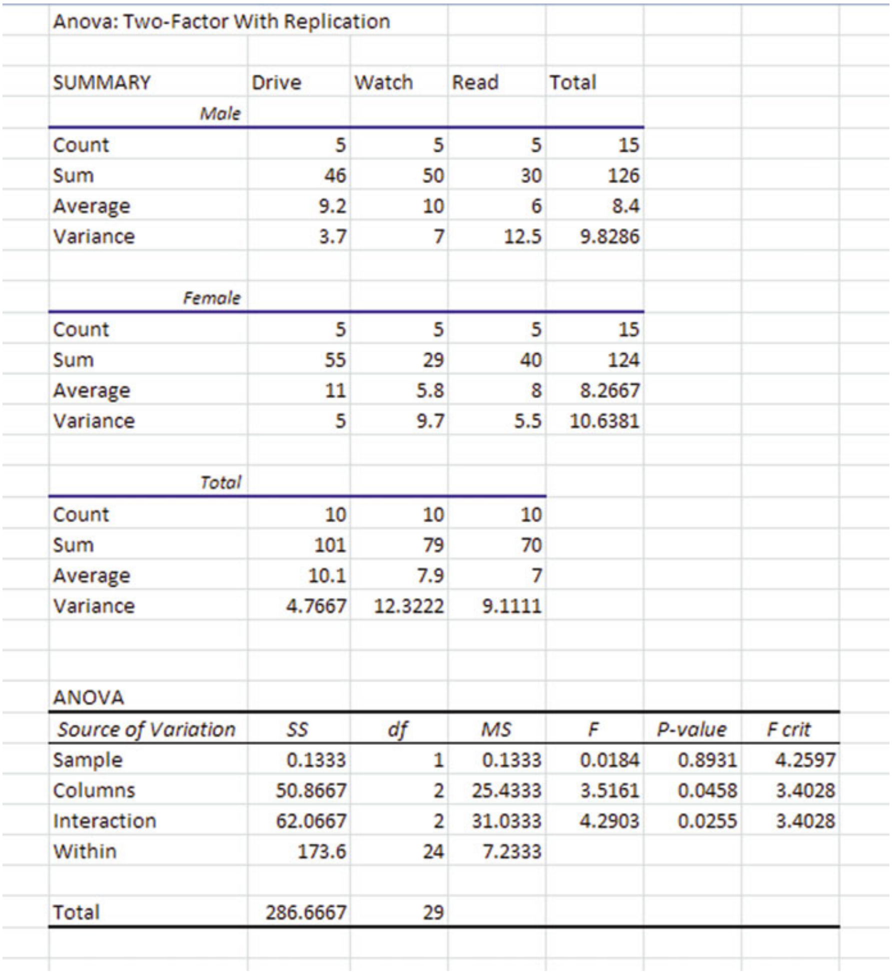
Output options

● Output Range: $F$1

○ New Worksheet Ply:

○ New Workbook

The output data is presented in Fig. 11.26.



Anova: Two-Factor With Replication

SUMMARY	Drive	Watch	Read	Total
Male				
Count	5	5	5	15
Sum	46	50	30	126
Average	9.2	10	6	8.4
Variance	3.7	7	12.5	9.8286
Female				
Count	5	5	5	15
Sum	55	29	40	124
Average	11	5.8	8	8.2667
Variance	5	9.7	5.5	10.6381
Total				
Count	10	10	10	
Sum	101	79	70	
Average	10.1	7.9	7	
Variance	4.7667	12.3222	9.1111	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	0.1333	1	0.1333	0.0184	0.8931	4.2597
Columns	50.8667	2	25.4333	3.5161	0.0458	3.4028
Interaction	62.0667	2	31.0333	4.2903	0.0255	3.4028
Within	173.6	24	7.2333			
Total	286.6667	29				

Fig. 11.26 Output

Make sure the **Count** adds up to your number of participants (15).

The **Sum, Average** and **Variance** columns provide the statistics for each condition (category) within each gender, as well as for an overall total.

To determine the significance of these results we can look in two places within the ANOVA output of Fig. 11.25.

1. If any of the **P-values** are less than the alpha value we selected (0.05), the results are statistically significant at the 95 % confidence level
2. If any of the three obtained **F-test** values are greater than the **critical F-test,** the results are statistically significant at the 95 % confidence level

The **Sample** refers to the first column where we defined the gender categories. The **Columns** refers to the three conditions we had in this study (drive, watch, read). The **Interaction** refers to the interaction effect between the gender and the condition.

In this case the **Columns P-values** for the **condition** (drive, watch, read) of 0.0458 and the **interaction** (gender x condition) of 0.0255 are significant since they are both less than our alpha value of 0.05.

The F crit values provide the same significance information. **F crit** is less than the **F-test** values for the conditions (**Columns: 0.0458**) and for the (**Interaction: 0.0255**) so both of these are statistically significant at 95 % confidence.

The sample achieved significance. When we divided the groups into gender groups the differences amongst the conditions are statistically important.

In summary this means that the conditions of driving, watching and reading have a real effect on the customers' perception ratings of the truck. However gender also has an effect. The females gave higher scores when they test drove the trucks whereas the males gave higher scores when they got to watch how the trucks were being built. These results are statistically significant.

t-Test

The t-test can be used effectively in a variety of situations:

- **Example 1 (Regression Problems):** Unlike the **F-test** which tests the statistical significance of the overall regression model, the **t-test** looks for statistical significance for each independent variable in the model. This analysis is completed using the Regression tool in Excel. Results are located in the associated ANOVA table.
- **Example 2 (Equality of Two Means):** The **t-test** can also be used to test if two population means are equal. It does this by comparing the ratio of the two means; if the means are equal, the ratio of variances will be 1.
- **Example 3 (Before-After Models):** The **t-test** may be used in before-after models, where the same respondents are measured before and after an intervention. It can be used when you match pairs of respondents on a variable related to the dependent variable. This type of testing is referred to as: paired-sample t-test, correlated t-test, or dependent-sample t-test.

Excel provides t-test analysis using individual tests located in the **Data Analysis** tab: **t-test: Paired Two Sample for Means, t-test: Two-Sample Assuming Equal Variance, t-test: Two-Sample Assuming Unequal Variances**. However all of these tests can also be run using the **T.TEST** function in the formula tab.

Rejection Rule When do we know that our results are real and not just due to chance? In other words when can we reject the null hypothesis:

$$\text{Reject } H_0 \text{ if } t_{\text{calculated}} > t_{\text{critical}}$$

The **critical value of t ($\mathbf{t_{critical}}$)** can be thought of as the largest value to occur by chance for the given degrees of freedom. If the $\mathbf{t_{calculated}}$ is larger than $\mathbf{t_{critical}}$, then the null hypothesis of equivalent group means is rejected. In other words your results are simply due to chance and are statistically significant.

The **p-value** is the probability that a value of **t** greater than or equal to the $\mathbf{t_{calculated}}$ could have occurred by chance if there were no difference in the variances. The **p-value** is the fractional area of the right tail of the **t-distribution**. $\mathbf{t_{calculated}}$ is indicated as **F** in the Excel output. When the **p-value** is less than your chosen confidence level, you have statistically significant results so keep going!

Basic Descriptions of t-Test Applications

This first section will describe the general application of each problem type by way of an example and the second section will walk through the step by step Excel applications for all three problem types.

Example 1: Regression Problems

This tests the statistical significance of the each independent variable in the input model for this sample output, in the regression model. The overall model is defined by:

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_0$$

In Fig. 11.26 we have included only one independent variable in the input model, for simplicity in the explanation of significance. The test results are shown in the third output table. The results of interest are indicating whether the independent variable (**X Variable 1**) is contributing in a statistically significant way in this regression model. Excel does not provide the $\mathbf{t_{crtical}}$ value so we can only use the **P- value**. The **P–value** (0.618085) is the probability associated with the $\mathbf{t_{calculated}}$ (shown as **t Stat**) of 0.539637. In other words there is a 0.618 probability of getting a t-value greater than 0.54. This is where the confidence level or alpha value comes in. Select the desired level for risk or alpha value. Let's select a 95 % confidence level so we have an alpha of 0.05. Now compare the **P–value** with the selected alpha value; since the **P–value** of 0.618 is greater than 0.05 that is not so good. We do not have statistically significant results at the 95 % confidence level for this independent variable. Our regression model will not yield useful results the way it is currently set up. So back to the drawing board and rethink the model.

Example 2: t-Test for Equality of Means

This test can be a two-tailed test or a one-tailed test. The two-tailed version tests against the alternative that the means are not equal. The one-tailed version only tests in one direction that the mean from the first population is either greater than or less than (but not both) the second population mean. The choice is determined by the problem. For example, if we are testing a new process, we may only be interested in knowing if the new mean is less than the mean of the old process. The **t-test** can be used to answer the following questions:

1. Do two samples come from populations with equal means? (two-tail)
2. Does a new process, treatment, or test reduce the mean of the current process? (one-tail)

A **t-test** returns the two-tailed probability that the means in sample 1 and sample 2 are not significantly different. We can use this function to determine whether two samples have different means. For example, given satisfaction scores from accountants and from engineers, you can test whether these professions have different average satisfaction scores.

Example 3: t-TEST Paired Samples

This **t-test** is used in before-after models, where the same respondents are measured before and after an intervention. It can also be used when you match pairs of respondents on a variable related to the dependent variable. Because this **t-test** is always a comparison of two means, it is a two sample test.

The two sets of data whose means are to be compared should be entered in consecutive rows of two columns. It is not necessary that both sample sizes be the same.

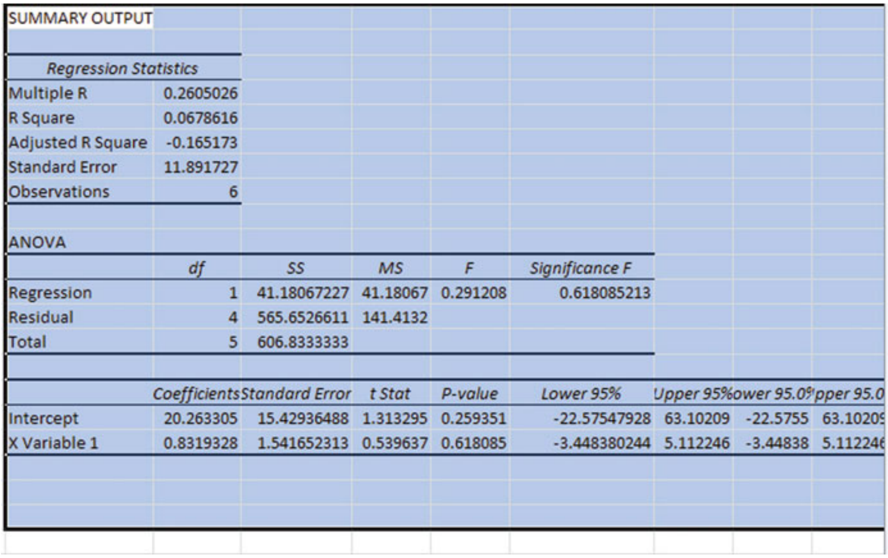
Excel

Excel examples are provided for each of the problems types:

- Example 1: Regression Problems
- Example 2: Equality of Two Means
- Example 3: Before-After Models

Example 1: Regression Problems

A sample of office assistants across the San Francisco-Bay Area was created. It was suggested that the data should be analyzed with a regression model to determine if salary was related to years of experience and aptitude test scores. We want a 99 % confidence level. This data is shown in Fig. 11.27.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.2605026
R Square	0.0678616
Adjusted R Square	-0.165173
Standard Error	11.891727
Observations	6

ANOVA

	df	SS	MS	F	Significance F
Regression	1	41.18067227	41.18067	0.291208	0.618085213
Residual	4	565.6526611	141.4132		
Total	5	606.8333333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0
Intercept	20.263305	15.42936488	1.313295	0.259351	-22.57547928	63.10209	-22.5755	63.10209
X Variable 1	0.8319328	1.541652313	0.539637	0.618085	-3.448380244	5.112246	-3.44838	5.112246

Fig. 11.27 Output

As a reminder the form of the regression model will be

$$y = b_o + b_1 x_1 + b_2 x_2$$

where

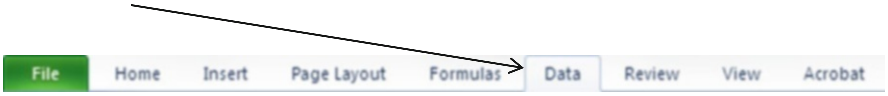
y = annual salary ($000)
x_1 = years of experience
x_2 = score on aptitude test

➢ Input the data from Fig. 11.28

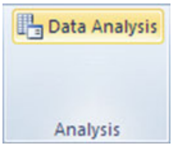
Experience (yrs) X_1	Score (out of 100) X_2	Salary ($000) Y
4	78	24
7	100	43
1	86	23.7
5	82	34.3
8	86	35.8
10	84	38
0	75	22.2
1	80	23.1
6	83	30
6	91	33
9	88	38
2	73	26.6
10	75	36.2
5	81	31.6
6	74	29
8	87	34
4	79	30.1
6	94	33.9
3	70	28.9
3	89	30

Fig. 11.28 Input

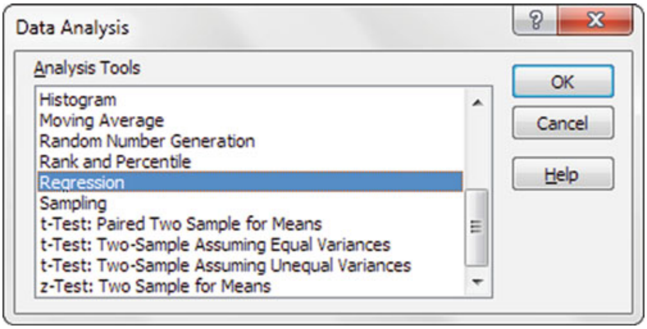
➢ Select **Data** from the menu bar



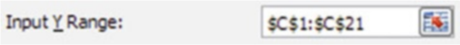
➢ Select the **Data Analysis** function



➢ Select **Regression**

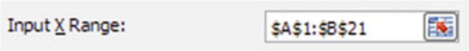

Data Analysis
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

- ➢ Click inside the **Input y Range** box to select the correct data



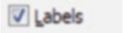
Input Y Range: $C$1:$C$21

- ➢ Click inside the **Input x Range** box to select the correct data



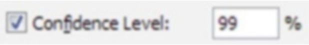
Input X Range: $A$1:$B$21

- ➢ Check the **Labels** box



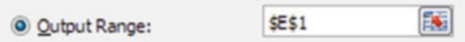
☑ Labels

- ➢ Select **Confidence Level.** Enter 99 in **Confidence Level** box



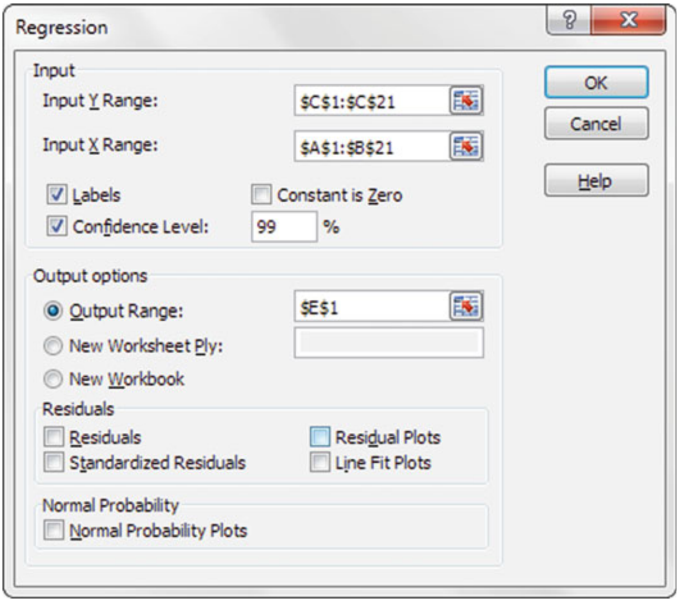
☑ Confidence Level: 99 %

- ➢ Select **Output Range** and click inside the **Output Range** box in order to highlight the cell where you want the output printed



● Output Range: $E$1

- ➢ Click **OK**



Regression

Input

Input Y Range: $C$1:$C$21

Input X Range: $A$1:$B$21

☑ Labels □ Constant is Zero

☑ Confidence Level: 99 %

Output options

● Output Range: $E$1

○ New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

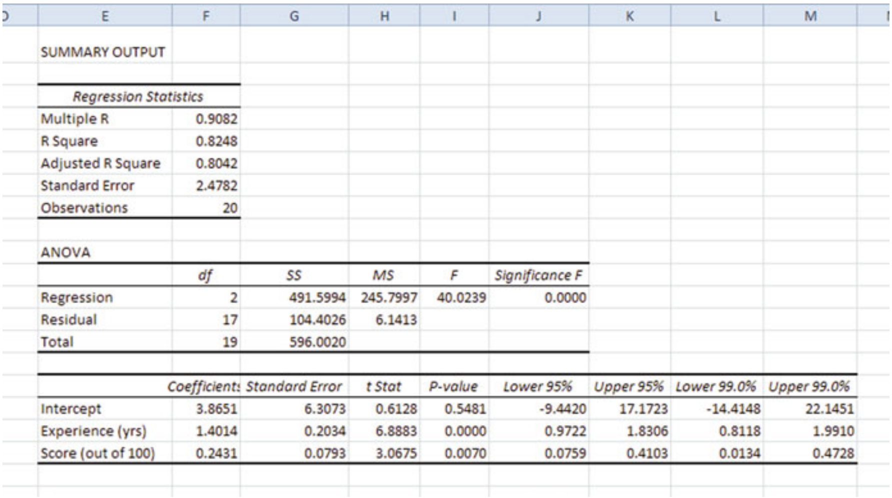
□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

The Excel output is provided in Fig. 11.29
The resulting equation is given in the ANOVA table in Fig. 11.29

$$y = 3.8651 + 1.4014x_1 + 0.2431x_2$$



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9082
R Square	0.8248
Adjusted R Square	0.8042
Standard Error	2.4782
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	2	491.5994	245.7997	40.0239	0.0000
Residual	17	104.4026	6.1413		
Total	19	596.0020			

	Coefficient:	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	3.8651	6.3073	0.6128	0.5481	-9.4420	17.1723	-14.4148	22.1451
Experience (yrs)	1.4014	0.2034	6.8883	0.0000	0.9722	1.8306	0.8118	1.9910
Score (out of 100)	0.2431	0.0793	3.0675	0.0070	0.0759	0.4103	0.0134	0.4728

Fig. 11.29 Output

To perform the t-test significance analysis, we compare the **p-values** with the selected alpha value of 0.01 (99 % confidence level). The results in the third table in the ANOVA section of Fig. 11.28 **P-value** column indicate that both Experience (x_1) and Score (x_2) are statistically significant. Both **p-values** are less than 0.01. In other words, if we selected 100 samples from this population, in 99 of those analyses we would get similar results. This suggests that we can use these two independent variables with 99 % confidence when predicting our independent variable which, in this case, is sales. Good model!

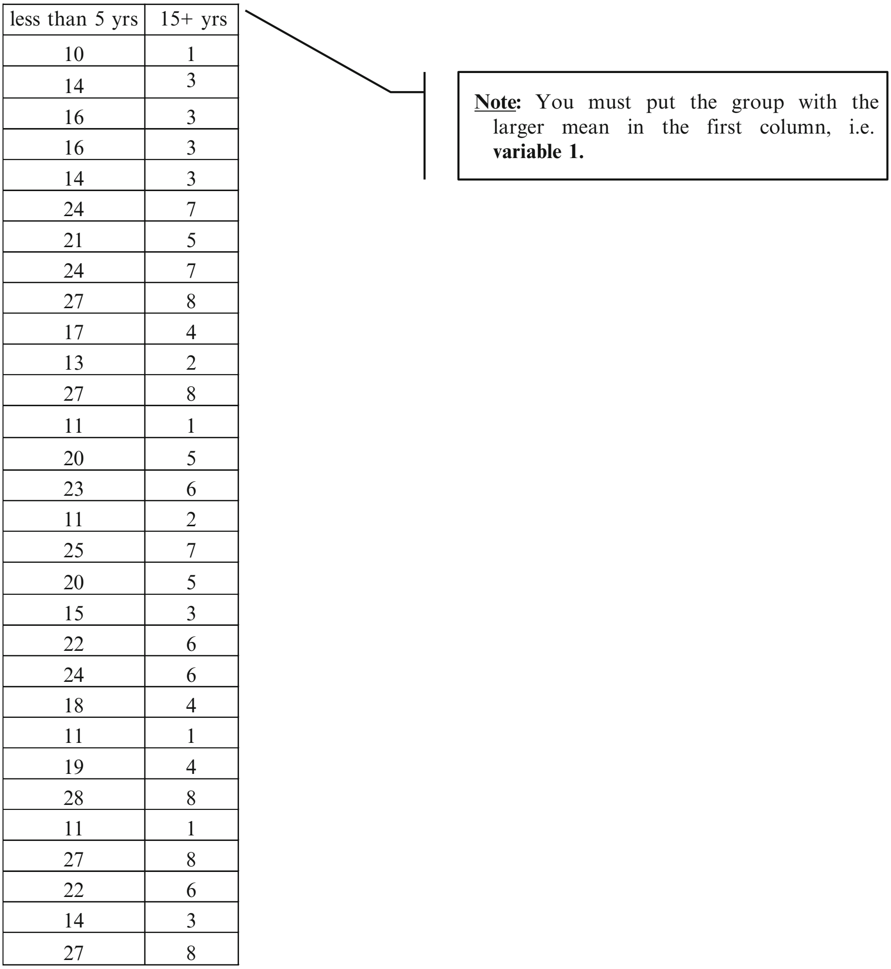
Example 2: t-Test for Equality of Means

This test can be run as a two-tailed test or a one-tailed test depending on model you are trying to solve.

- The One-tailed ***t***-test for Equality of Means

A one-sided (or one-tailed) test is appropriate if researchers want to know if the population mean of one of the data sets is larger than that of another set of data; or, equivalently, if the population mean of one set of data is smaller than that of another set of data. For example, suppose that the two groups of employees report on the time (hours) they spend handling customer complaints. The first group has less than 5 years of experience, and the second group has at least 15 years of experience. We have already calculated that the two means are different; group 1 spends 19.03 h handling complaints on average and group 2 spends 4.62 h. However, there is always the possibility that this difference in the means could have occurred by chance. We need to test if these results are significant. A one-sided **t-test** is appropriate in the scenario that one sample mean is "smaller than" the other sample mean. In this case, the more experienced employees on average spend less time handling customer complaints. Assume a 95 % confidence level.

➢ Input the data from Fig. 11.30



less than 5 yrs	15+ yrs
10	1
14	3
16	3
16	3
14	3
24	7
21	5
24	7
27	8
17	4
13	2
27	8
11	1
20	5
23	6
11	2
25	7
20	5
15	3
22	6
24	6
18	4
11	1
19	4
28	8
11	1
27	8
22	6
14	3
27	8

Note: You must put the group with the larger mean in the first column, i.e. **variable 1.**

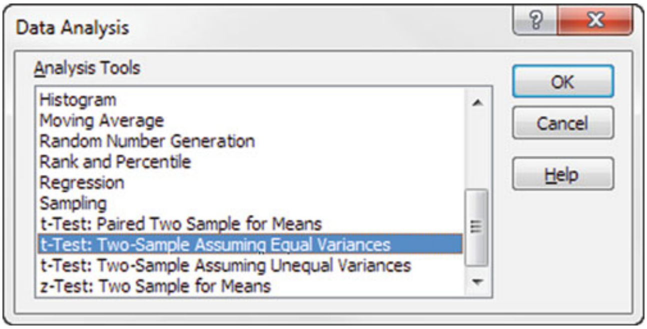
Fig. 11.30 Input

Before you select the correct Excel analysis tool you need to determine if you know the population variances. If you do, are they equal? Here are the recommended t-test tools based on this information:

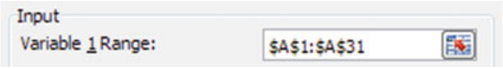
Problem type	Excel tool
Equal variances	**t-Test: Two-Sample Assuming Equal Variances**
Unequal variances	**t-Test: Two-Sample Assuming Unequal Variances**
Unknown variances with sample size >30	**t-Test: Two-Sample Assuming Equal Variances**
Unknown variances with sample size <30	**t-Test: Two-Sample Assuming Unequal Variances**
Unknown variances but assumed to be unequal based on other information	**t-Test: Two-Sample Assuming Unequal Variances**

In this case let's assume we know that the population variances are equal. But you would follow the same instructions if you had selected the case for unequal variances.

- ➢ Click on the **Data** tab and select **Data Analysis** function
- ➢ Select **t-Test: Two-Sample Assuming Equal Variances** from the menu

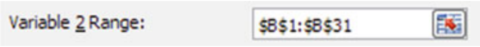

Data Analysis
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

- ➢ Click **OK**
- ➢ Click inside the **Variable 1 Range box**. Highlight the first column of data including the label.

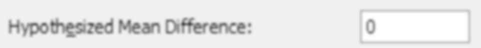

Input

Variable 1 Range: `$A$1:$A$31`

- ➢ Click inside the **Variable 2 Range box**. Highlight the second column of data including the label.

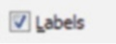

Variable 2 Range: `$B$1:$B$31`

➢ Enter **0** in the **Hypothesized Mean Difference** box, as we are testing if these two means are equal.



Hypothesized Mean Difference: 0

➢ Check the **Labels** box

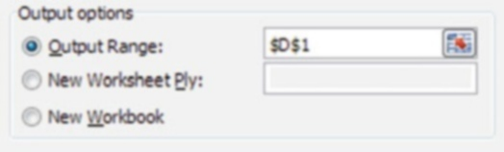


☑ Labels

➢ Select an alpha level of **0.05**

Alpha: 0.05

➢ Select **Output Range** and click inside the **Output Range** box in order to highlight the cell where you want the output printed



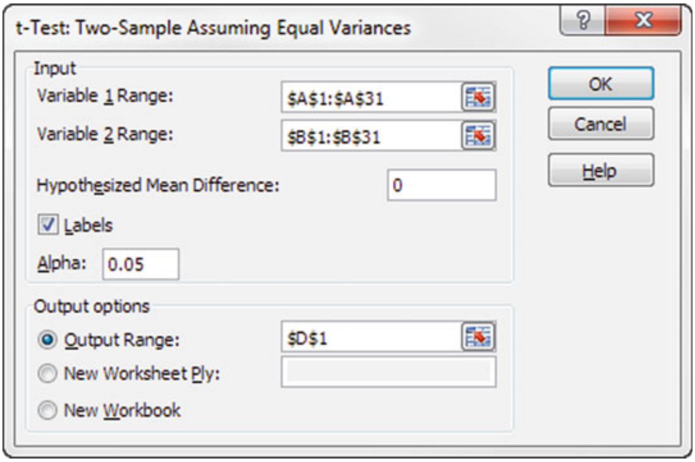
Output options

● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

➢ Click **OK**



t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range: $A$1:$A$31

Variable 2 Range: $B$1:$B$31

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

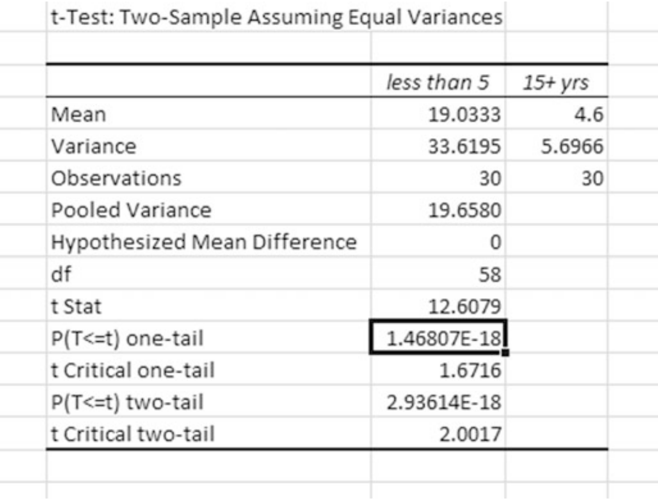
Output options

● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

The results will appear as follow in the next output table (Fig. 11.31).



t-Test: Two-Sample Assuming Equal Variances

	less than 5	*15+ yrs*
Mean	19.0333	4.6
Variance	33.6195	5.6966
Observations	30	30
Pooled Variance	19.6580	
Hypothesized Mean Difference	0	
df	58	
t Stat	12.6079	
P(T<=t) one-tail	1.46807E-18	
t Critical one-tail	1.6716	
P(T<=t) two-tail	2.93614E-18	
t Critical two-tail	2.0017	

Fig. 11.31 Output

Make sure you check all of the data have been input by looking at the **Observations**. The **Observations** indicate we have included all of our data (30).

The **Mean** and **Variance** are given for each group.

Pooled Variance is the weighted (using degrees of freedom) average of the two sample variances. This value is probably not of much use to you.

Hypothesized mean Difference we set to zero to test if the two means are in fact the same.

df is just degrees of freedom for the test and is calculated by ($n_1 + n_2 - 2$).

tStat is the t-value calculated based on the input data (12.6079)

The next four outputs are used for our significance testing. This problem is only testing using one-tail.

One-Tailed Test

1. We can compare the **P(T <= t) one tail** value (1.46807E-18) with the alpha value we selected (0.05). Since this p value is less than alpha we confirm that the difference in the means is statistically significant at the 95 % confidence level.
2. We can compare the **tStat** (12.6079) with the **t Critical one-tail** (1.6716). Since the **tStat** is greater than the **tcrit**, we can conclude our results are statistically significant at the 95 % confidence level. If we selected 100 other samples from this population we would get similar results in 95 of those samples.

The variance in satisfaction scores is statistically different; we can assume the less experienced employees really do have a greater mean than those more experienced employees. In other words the less experienced group spends more time on average handling complaints than their more experienced counterparts. This may be worth looking into in terms of understanding best practices and managing time efficiently.

Two-Tailed Test

The **two-tailed t-test** determines if one population mean is statistically significantly different than another population mean. We have no knowledge nor do we care if one is greater or less than the other, just that they are significantly different. The previous problem would have required a two-tailed analysis if it had been worded "*Is the difference in average time spent handling complaints statistically significant between the two groups of employees with different experience levels?*"

To analyze these results we would follow the same logic as before but word the conclusion somewhat differently.

1. We can compare the **P(T <= t) two tail** value (2.93614E-18) with the alpha value we selected (0.05). Since the P value is less than alpha we confirm that the difference in the means is statistically significant at the 95 % confidence level.
2. We can compare the **tStat** (12.6079) with the **t Critical two-tail** (2.0017). Since the **tStat** is greater than the **tcrit**, we can conclude our results are statistically significant at the 95 % confidence level. If we selected 100 other samples from this population we would get similar results in 95 of those samples.

In this case we can only conclude the means are statistically significantly different at the 95 % confidence level. We cannot infer one mean is significantly larger than the other. Training may in fact increase or decrease the scores.

Example 3: Before-After Models

A group of marketing professionals provided the client satisfaction scores out of 100 that they received during the first quarter of this year. The range of possible scores is 0–100 with 100 as the highest possible score. The same marketing professionals then attended a company training program on effective communications and submitted their client satisfaction scores in the following quarter of the year.

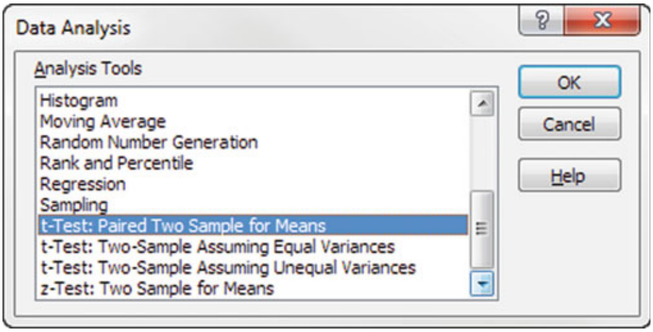
- Is the difference in these scores real or simply due to chance at the 95 % confidence level? This would be a **two-tailed t-test** as we don't care if one is greater than the other, just that they are different.
- Are the average employee satisfaction scores higher after training? In other words was training effective? This is a **one-tailed t-test** since it determines if one population mean is greater or less than the other in a statistically significantly way.

➢ Input the following data from Fig. 11.32

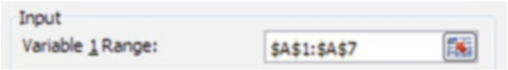
Pre-training scores	Post training scores
90.47	90.95
92.02	91.30
93.15	91.48
90.98	92.04
91.73	90.70
94.00	91.33

Fig. 11.32 Input

- ➢ Click on the **Data** tab and select **Data Analysis** function
- ➢ Select **t-test: Paired Two-Sample for means**

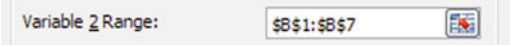

Data Analysis
?
X
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

- ➢ Click **OK**
- ➢ Click inside the **Variable 1 Range** box to select the correct data

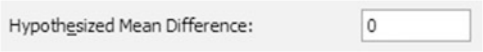

Input

Variable 1 Range: $A$1:$A$7

- ➢ Click inside the **Variable 2 Range** box to select the correct data


Variable 2 Range: $B$1:$B$7

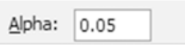
- ➢ Enter **0** in the **Hypothesized Mean Difference** box, as we are testing if these two means are equal


Hypothesized Mean Difference: 0

- ➢ If you included labels in your input range check the **Labels** box.

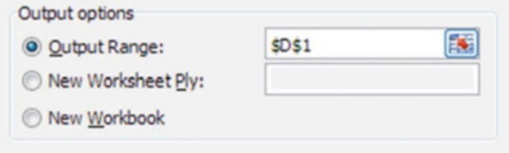
☑ Labels

- ➢ Enter the level of significance or "**Alpha**" value. The default that should appear is 0.05 which, of course, is a 95 % confidence level. But you can choose whatever alpha value you want.



Alpha: 0.05

- ➢ Select **Output Range** and click inside the **Output Range** box in order to highlight the cell where you want the output printed



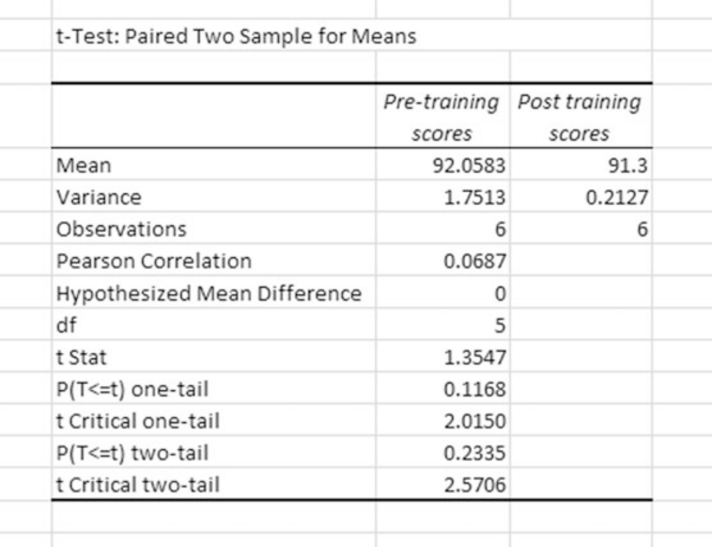
Output options

● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

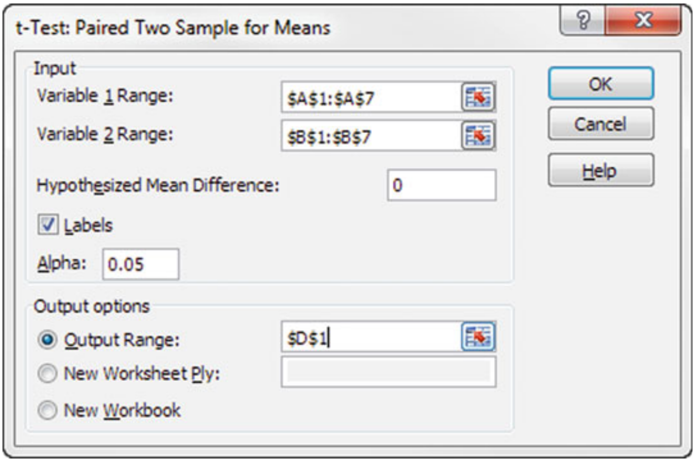
- ➢ Click **OK** to see the output of the function (Fig. 11.33)



t-Test: Paired Two Sample for Means

	Pre-training scores	*Post training scores*
Mean	92.0583	91.3
Variance	1.7513	0.2127
Observations	6	6
Pearson Correlation	0.0687	
Hypothesized Mean Difference	0	
df	5	
t Stat	1.3547	
P(T<=t) one-tail	0.1168	
t Critical one-tail	2.0150	
P(T<=t) two-tail	0.2335	
t Critical two-tail	2.5706	

Fig. 11.33 Output



t-Test: Paired Two Sample for Means

Input

Variable 1 Range: $A$1:$A$7

Variable 2 Range: $B$1:$B$7

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

Output options

● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

Output data is provided in Fig. 11.33.

Make sure you check all of the data have been input by looking at the **Observations**. The **Observations** indicate we have included all of our data (6).

The **Mean** and **Variance** are given for each group.

Pearson Correlation is the r value **(correlation coefficient)**. This number tells us that there is a positive relationship between the two sets of results, but it is very weak at only 0.07.

Hypothesized mean Difference we set to zero to test if the two means are in fact the same.

df is just degrees of freedom for the test and is calculated (n − 1)

tStat is the t-value calculated based on the input data

The next sections will provide instructions on completing the significance analysis.

One-Tailed Test

We can compare the **P(T <= t) one tail** value (0.1168) with the alpha value we selected (0.05). Since the **P value** is greater than alpha we confirm that the difference in the variance is not statistically significant at the 95 % confidence level.

We can compare the **tStat** (1.3547) with the **t Critical one-tail** (2.0150). Since the **tStat** is less than the **t Critical one-tail**, we can conclude our results are not statistically significant at the 95 % confidence level. If we selected other samples we would not get similar results.

The mean client satisfaction scores are not statistically different; we can assume the client satisfaction scores are not higher after training. If we are striving for higher scores, than we should actually test for that. It seems that training may not have been effective.

Two-Tailed Test

To analyze these results with a two tailed test, we would follow the same logic as before but word the conclusion somewhat differently.

We can compare the **P(T <= t) two tail** value (0.2335) with the alpha value we selected (0.05). Since the P value is greater than alpha we confirm that the difference in the means is not statistically significant at the 95 % confidence level.

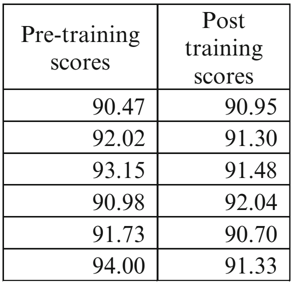
We can compare the **tStat** (1.35467) with the **t Critical two-tail** (2.5706). Since the **tStat** is less than the **t Critical two-tail**, we can conclude our results are not statistically significant at the 95 % confidence level. If we selected other samples we would not get similar results.

In conclusion the communications training did not have a statistically significant effect on the client satisfaction scores of the employees. There must be another more important factor that is accounting for the client satisfaction results. Or perhaps the type of training that was given was not appropriate for this context.

T.TEST

The **T.TEST** tool can also provide the p-values for each of the previous three examples using the Excel tools in **Data Analysis**. The difference with the **T.TEST** results is that you only get information about the p-value, no critical value information and only one result at a time. In the previous **Data Analysis** tools you can get multiple results within each analysis.

➢ Input data from Fig. 11.34



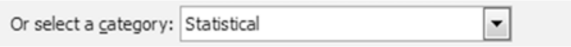
Pre-training scores	Post training scores
90.47	90.95
92.02	91.30
93.15	91.48
90.98	92.04
91.73	90.70
94.00	91.33

Fig. 11.34 Input

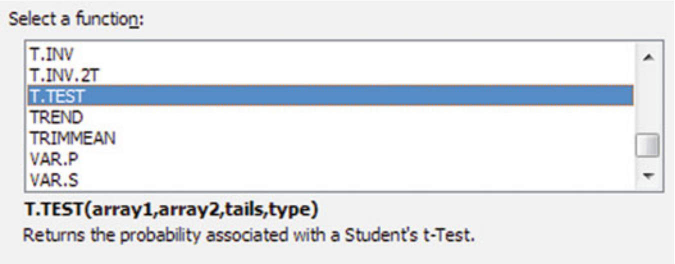
To access the **T.TEST** tool:

➢ Select a cell, and click the Function button

➢ In the select a category section, select **Statistical**


Or select a category: Statistical

➢ Select **T.TEST**, for Function.


Select a function:

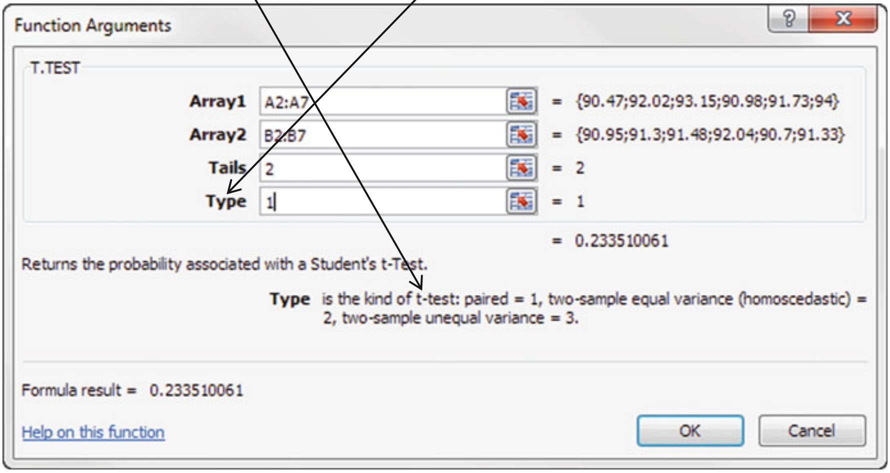
- T.INV
- T.INV.2T
- T.TEST
- TREND
- TRIMMEAN
- VAR.P
- VAR.S

T.TEST(array1,array2,tails,type)

Returns the probability associated with a Student's t-Test.

◇ **Caution:** Do not include labels in your array data with this tool. Since it is operating as a formula you just want the actual data values to be included.

The test type is input in the last box labeled **Type**. Each type is listed at the bottom of the input page.


Function Arguments

T.TEST

Array1	A2:A7	= {90.47;92.02;93.15;90.98;91.73;94}
Array2	B2:B7	= {90.95;91.3;91.48;92.04;90.7;91.33}
Tails	2	= 2
Type	1	= 1

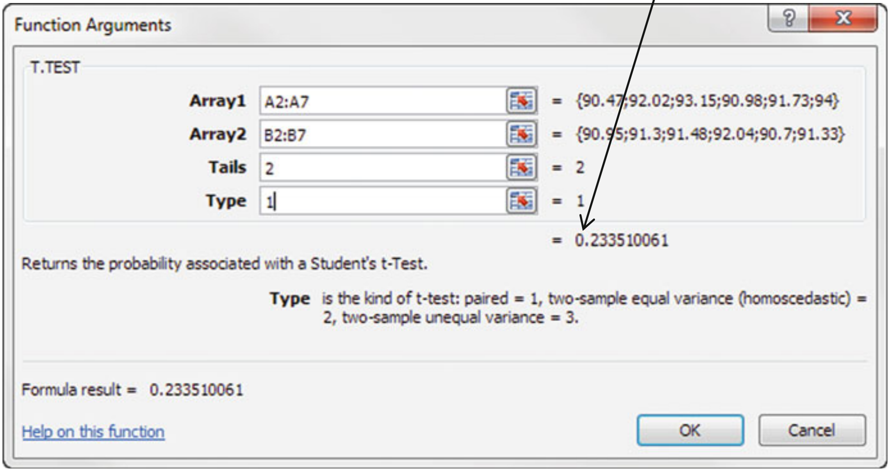
= 0.233510061

Returns the probability associated with a Student's t-Test.

Type is the kind of t-test: paired = 1, two-sample equal variance (homoscedastic) = 2, two-sample unequal variance = 3.

Formula result = 0.233510061

- **Paired two sample- Two Tailed:** provides the p value of 0.23351 as shown on the input screen



Function Arguments

T.TEST

Array1 A2:A7 = {90.47;92.02;93.15;90.98;91.73;94}

Array2 B2:B7 = {90.95;91.3;91.48;92.04;90.7;91.33}

Tails 2 = 2

Type 1 = 1

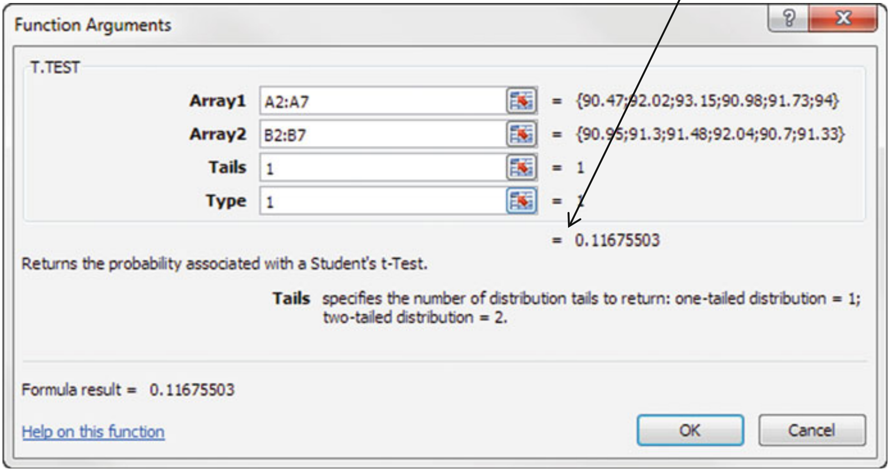
= 0.233510061

Returns the probability associated with a Student's t-Test.

Type is the kind of t-test: paired = 1, two-sample equal variance (homoscedastic) = 2, two-sample unequal variance = 3.

Formula result = 0.233510061

- **Paired two sample- One Tailed:** provides the p value of 0.11675503 as shown on the input screen



Function Arguments

T.TEST

Array1 A2:A7 = {90.47;92.02;93.15;90.98;91.73;94}

Array2 B2:B7 = {90.95;91.3;91.48;92.04;90.7;91.33}

Tails 1 = 1

Type 1 = 1

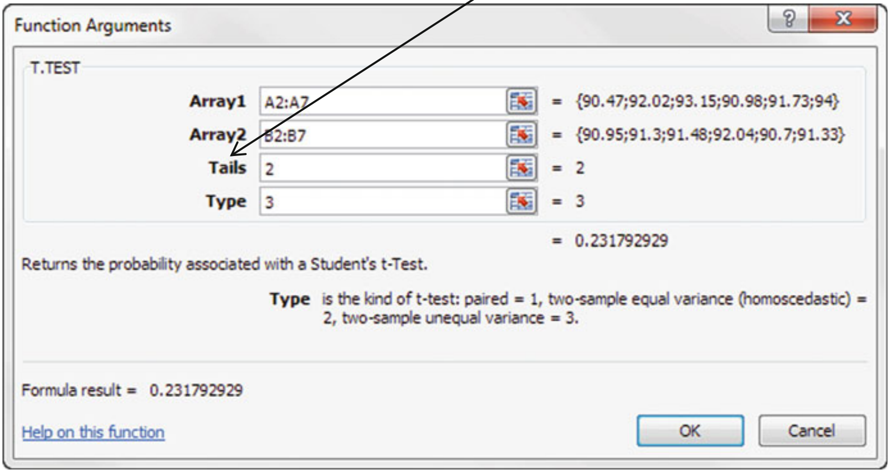
= 0.11675503

Returns the probability associated with a Student's t-Test.

Tails specifies the number of distribution tails to return: one-tailed distribution = 1; two-tailed distribution = 2.

Formula result = 0.11675503

- **Two- sample equal variance (homoscedastic)** provides the **p value** of 0.214504933 as shown on the input screen. Note this can also be run as a one-tailed test by changing the **Tails** to the value of “2”.



Function Arguments

T.TEST

Array1 A2:A7 = {90.47;92.02;93.15;90.98;91.73;94}

Array2 B2:B7 = {90.95;91.3;91.48;92.04;90.7;91.33}

Tails 2 = 2

Type 3 = 3

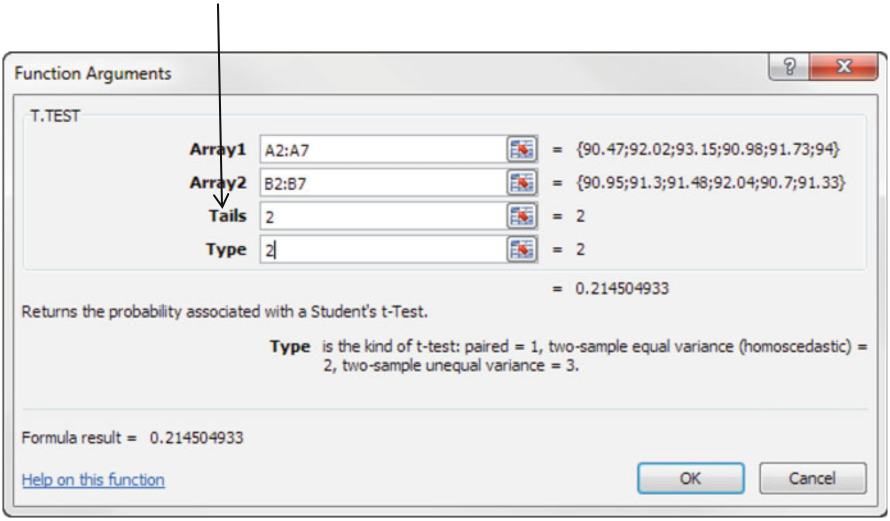
= 0.231792929

Returns the probability associated with a Student's t-Test.

Type is the kind of t-test: paired = 1, two-sample equal variance (homoscedastic) = 2, two-sample unequal variance = 3.

Formula result = 0.231792929

- **Two- sample unequal variance** provides the **p value** of 0.23179 as shown on the input screen. Note this can be run as a one-tailed test by changing the **Tails** to the value of “2”.



Function Arguments

T.TEST

Array1 A2:A7 = {90.47;92.02;93.15;90.98;91.73;94}

Array2 B2:B7 = {90.95;91.3;91.48;92.04;90.7;91.33}

Tails 2 = 2

Type 2 = 2

= 0.214504933

Returns the probability associated with a Student's t-Test.

Type is the kind of t-test: paired = 1, two-sample equal variance (homoscedastic) = 2, two-sample unequal variance = 3.

Formula result = 0.214504933

Common Excel Pitfalls

- Make sure you work with absolute values when working with critical values
- Two-way between-group **ANOVA** must include labels in input arrays
- Don't include labels in the array in **T.TEST** since this tool operates as a formula
- **One-tailed t-test for Equality of Means** requires including the group with the larger mean in the first column

Final Thoughts and Activities

Practice Problems and Case Studies

1. A professional golf ball manufacturing company wanted to compare the distance traveled by golf balls produced using each of four different designs. Ten balls were manufactured with each design and were brought to the local golf course for the club professional to test. The order in which the balls were hit with the same club from the first tee was randomized so that the pro did not know which type of ball was being hit. All 40 balls were hit in a short period of time during which the environmental conditions were essentially the same. The results (distance traveled in yards) for the four designs are stored in the file **Ball**.

   (a) At the 0.05 level of significance, is there evidence of a difference in the mean distances traveled by the golf balls with different designs?
   (b) What assumptions are necessary in (a)?

2. You are planning on stocking your grocery counter with chewing gum but want to know if customers really want this product. The other branches of this store (population) sell on average 10 packs of gum per customer every 2 months (population average). You have limited time and limited money so you survey a very small sample of your customers, n = 15. You ask them how often they have purchased gum in the last 60 days. You want a 95 % confidence level ($\alpha = 0.05$) that all of your customers (population) will also purchase on average 10 packs every 2 months. Test to make sure the average result you calculate for your customers is in fact statistically significantly different from the population average of 10 packs per customer every 60 days. The data is listed below:

   Number of packages of gum purchased in the last 60 days

Packs of gum purchased in 60 days
10
2
4
6

(continued)

(continued)

Packs of gum purchased in 60 days
0
15
22
3
8
8
7
10
7
7

Discussion Boards

1. All criminals should take lie detector tests, regardless of false positives.
2. The death penalty discourages others from committing murder in the United States.
3. What are the ethical issues involved in hypothesis testing?

Group Activity

1. Choose a company with data available online. Discuss the company's reliability claims. What type of data would this company collect as part of its market research? Discuss the issue of valid research results.
2. There is more than one way to measure reliability but often the actual statistical source is not provided. Complete a web search to see if you can produce any reliability results, using a statistic other than a Cronbach's alpha.
3. Explore the following hypotheses using the web:

   H_0: Employees are not more productive working in cubicles.
   H_α: Employees are more productive working in cubicles.

Parting Thought

A statistician is a mathematician broken down by age and sex . . .

Problem Solutions

1. A professional golf ball manufacturing company wanted to compare the distance traveled by golf balls produced using each of four different designs. Ten balls were manufactured with each design and were brought to the local golf course for the club professional to test. The order in which the balls were hit with the same club from the first tee was randomized so that the pro did not know which type of ball was being hit. All 40 balls were hit in a short period of time during which the environmental conditions were essentially the same.

   Answer:

Anova: Single factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Design1	10	2069.64	206.964	5.246427
Design2	10	2188.66	218.866	30.66718
Design3	10	2269.38	226.938	23.18213
Design4	10	2289.72	228.972	16.10697

ANOVA

Source of variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between groups	2990.99	3	997.00	53.03	0.00	2.87
Within groups	676.82	36	18.80			
Total	3667.814	39				

   (a) At the 0.05 level of significance, is there evidence of a difference in the mean distances traveled by the golf balls with different designs?

   Because $F = 53.03 > F_{\text{crit}} = 2.8$, there is a significant difference in the mean distances traveled by the golf balls with different designs

   (b) What assumptions are necessary in (a)?

   The assumptions are that the golf ball samples are randomly and independently selected and the variances are constant.

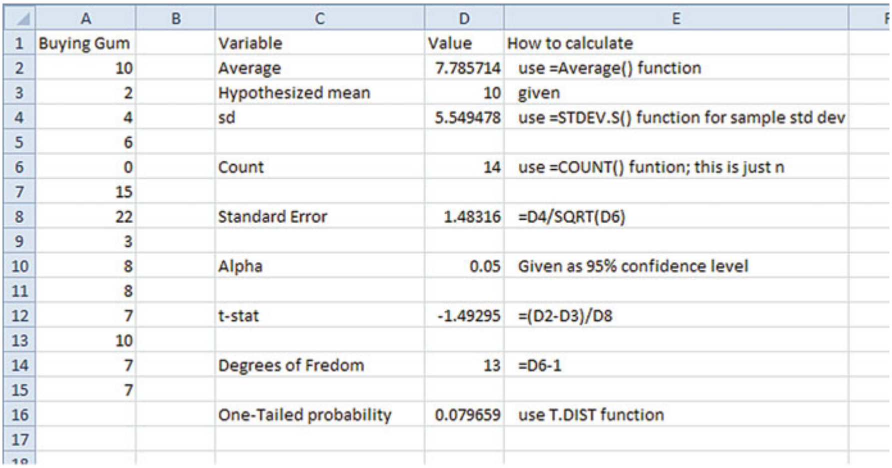
2. You are planning on stocking your grocery counter with chewing gum but want to know if customers really want this product. The other branches of this store (population) sell on average ten packs of gum per customer every 2 months (population average). You have limited time and limited money so you survey a very small sample of your customers, n = 15. You ask them how often they

have purchased gum in the last 60 days. You want a 95 % confidence level ($\alpha = 0.05$) that all of your customers (population) will also purchase on average 10 packs every 2 months. Test to make sure the average result you calculate for your customers is in fact statistically significantly different from the population average of 10 packs per customer every 60 days. The data is listed below:
Number of packages of gum purchased in the last 60 days

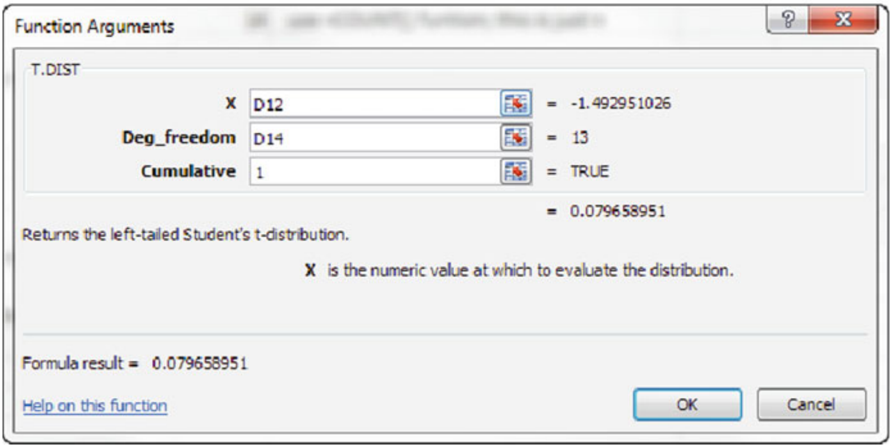
Packs of gum purchased in 60 days
10
2
4
6
0
15
22
3
8
8
7
10
7
7

Answer:

In this problem we use a **one-tailed t-test** to see if the two averages really are different (alternative hypothesis), or if in fact the sample should be assumed to have the same average as the population (null hypothesis).



	A	B	C	D	E	F
1	Buying Gum		Variable	Value	How to calculate	
2	10		Average	7.785714	use =Average() function	
3	2		Hypothesized mean	10	given	
4	4		sd	5.549478	use =STDEV.S() function for sample std dev	
5	6					
6	0		Count	14	use =COUNT() funtion; this is just n	
7	15					
8	22		Standard Error	1.48316	=D4/SQRT(D6)	
9	3					
10	8		Alpha	0.05	Given as 95% confidence level	
11	8					
12	7		t-stat	-1.49295	=(D2-D3)/D8	
13	10					
14	7		Degrees of Fredom	13	=D6-1	
15	7					
16			One-Tailed probability	0.079659	use T.DIST function	
17						



Function Arguments

T.DIST

X: D12 = -1.492951026

Deg_freedom: D14 = 13

Cumulative: 1 = TRUE

= 0.079658951

Returns the left-tailed Student's t-distribution.

X is the numeric value at which to evaluate the distribution.

Formula result = 0.079658951

Where:

X is the t-value you just calculated
Deg_Freedom is the degrees of freedom (n − 1)
Cumulative is where we let Excel know we want the left hand tail only. You type in the word **TRUE** or the number **1**.

In this case we have a left hand tail of 0.08 (one tailed probability) which exceeds our alpha value of 0.05. This means our results are not significantly different. The average gum purchases per customer that our sample indicated of approximately 8 packs in 2 months is not significantly different from the population average of 10 packs. Therefore we can assume that in fact the population average of 10 packs can be used to describe these customers as well. If our results had been statistically significant we could safely assume our sample average was real and feel safe to use it to describe our population of customers. Our store owner should go ahead and stock gum in his store as he can expect to sell about as much as the other stores.

Another way of proving significance is to use the t_c calculation where we compare the calculated t with the critical t (based on degrees of freedom). Remember when the $t_{\text{calculated}} < t_{\text{critical}}$ the results are not statistically significant.

Chapter 12
Significance Tests Part 2

Key Concepts

Chi-square test and z-test.

Discussion

The significance tests covered in this chapter include the X^2 test and the z test. The conditions for selecting either of these significance tests are included in this chapter. See Fig. 12.1.

Remember, significance only says that the difference is actually true. Never allow statistics to overrule common sense.

The choice of the best technique depends on the type of questions to be answered, the number of variables and the measurement scale, sample size and how much is known about the population and whether you need a one tailed or two tailed test. Remember none of these tests are necessary if you have access to the entire population. These tests simply advise you whether you should trust your sample results . . . are these statistically significant results?

Significance Tests

The next two sections will define the X^2 test and the z-test with examples. The Excel tools will be discussed for each of these significance tests.

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_12,

Test	When to Use	One tailed	Two tailed
F-test (Chapter 11)	Repeated measure using ANOVA. Same respondents under different conditions	x	
	Overall regression model significance	NA	NA
	Equality of 2 group means	x	x
	Between Group ANOVA. Test several variables within your group of data	x	x
t-test (Chapter 11)	Equality of means when population std dev unknown. Can use with small samples	x	x
	Matched Pair test. Before and after analysis with same respondents	x	x
	Regression. Tests each regression co-efficient	NA	NA
Chi-squared (Chapter 12)	Goodness of Fit for model	NA	NA
	Independence of variables	NA	NA
z-test (for normally distributed data) (Chapter 12)	Testing the mean (or proportion) of a population versus a standard	x	x
	Comparing the means (proportions) of two populations, with large sample size (n ≥ 30)	x	x
	Matched Pair test. Before and after analysis with same respondents Samples should large (n ≥ 30)	x	x

Fig. 12.1 Selecting the correct significance test

X^2 Test

The X^2 test can be used effectively under the several conditions:

- **Example 1 Goodness-of-Fit:** for estimating how closely an observed distribution of one variable matches its expected distribution
- **Example 2 Independence of Two Variables:** for estimating whether two random variables are independent

The **Chi Square statistic** compares the tallies or counts of categorical responses between two (or more) independent groups. There are a number of features of the social world we characterize through categorical variables: religion, political preferences etc. It can be used to compare the observed frequency distribution with the expected distribution.

If we are concerned about whether the data distribution is normal the quick-and-dirty Excel normality test is simply to throw the data into an Excel histogram and eyeball how close the shape of the graph matches with a normal distribution. If there is a still a question, the next (and easiest) normality test is the Chi-Square Goodness-Of-Fit test.

The Chi-Square Goodness-Of-Fit test is less well known than some other normality tests such as the Kolmogorov-Smirnov test, the Anderson-Darling test, or the Shapiro-Wilk test. The Chi-Square Goodness-Of-Fit test is a lot less complicated, and a whole lot easier to implement in Excel than any of the more well-known normality tests.

Rejection Rule. When the computed X^2 statistic exceeds the critical value for a given probability level, then we can reject the null hypothesis. If you select an $\alpha = 0.05$ (CI = 95 %), then you are accepting a 5 % chance that your significance is wrong. If the rejection rule provided statistically significant results, there would still be a 5 % chance the results are only due to chance.

Example 1: Goodness-of-Fit Test

Consider the case of a gambler playing craps (dice game) in a local casino. The casino and the gamblers are always interested in fairness and cheating in games of chance. Since such games usually involve money, there is an incentive for both sides to try to rig the games in their favor.

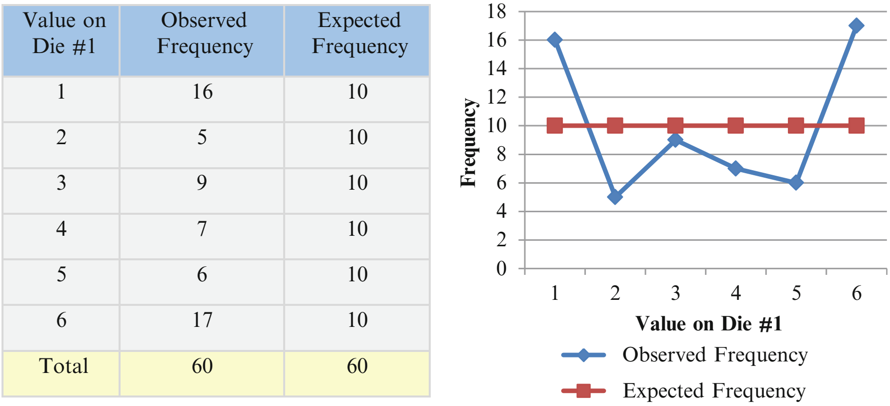
The goodness-of-fit test can be used to examine cheating in gambling. In the game of craps, two dice are thrown. Most dice used in wagering have six sides, with each side having a value of one, two, three, four, five, or six. If the dice being used are fair, the chance of any particular number coming up is the same: 1 in 6. However, if the dice have been tampered with, often referred to as “loaded”, then certain numbers will have a greater likelihood of appearing, while others will have a lower likelihood. In other words we compare the observed distribution of data with the expected distribution.

Consider the example in which the casino’s resident statistician collects data from one of the craps tables (Fig. 12.2). The two dice were thrown 60 times and he recorded the outcomes for one of those dice. The expected value for each possible outcome is 10 in 60 or more simply 1 in 6. These values are given in the last column.

In the Observed Frequency column there are more 1’s and 6’s than expected, and fewer 2’s, 3’s, 4’s and 5’s than expected. The chi-square function can be used to estimate the likelihood that the values observed on die #1 occurred by pure chance, and therefore prove that there is no relationship between the observed frequency data.

The first step is to compare the observed and expected values. This information is input to Excel to calculate the significance level (p-value) using the chi-square statistic (CHISQ.TEST).

In this example the expected data follow a uniform distribution but often the chi square test is used when the expected data follow a normal distribution or other distributions.



Value on Die #1	Observed Frequency	Expected Frequency
1	16	10
2	5	10
3	9	10
4	7	10
5	6	10
6	17	10
Total	60	60

Fig. 12.2 Frequency table for die #1

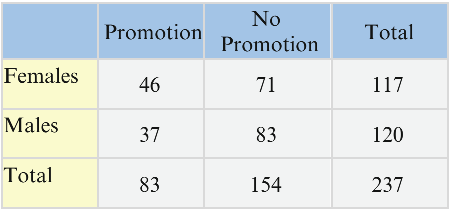
You may also need the actual chi-square statistic for reporting purposes. So the second step calculates X^2 by using the CHIINV function in Excel (the inverse of the chi-square) using the p-value and the degrees of freedom.

The p-value is different from the chi-square statistic itself; it is the probability of observing a chi-square at least as large as the one you determine for the data set, if the null hypothesis is true. The null hypothesis is that the two variables (observed frequency and expected frequency in this case) are unrelated. The alternate hypothesis is that the two variables are related in some way. Generally, if the p-value is lower than some predetermined level, typically .05 (α level), then we reject the null hypothesis. By rejecting the null hypothesis means there is some relationship between the expected and observed variables and you can continue with your analysis. In this particular case that would mean the observed outcomes are not due to chance . . . probably loaded dice!

Example 2: Independence of Two Variables

The other primary use of the chi-square test is to examine whether two variables are independent or not. In this case there are data sets associated with each of the variables, unlike the goodness-of-fit application where you have only observed data for one variable. In this significance test independence means that the two factors (variables) are not related. Typically in social science research, we're interested in finding factors that are related – education and unemployment, stock price and CEO salary, age and consumer behavior. In this case, the chi-square can be used to assess whether two variables are independent or not.

More generally, we say that the y variable is "not correlated with" the x variable if more of one is not associated with more of another. If two categorical variables are correlated their values tend to move together, either in the same direction (+r) or in the opposite (−r). In most cases we hope for a relationship between the two factors.



	Promotion	No Promotion	Total
Females	46	71	117
Males	37	83	120
Total	83	154	237

Fig. 12.3 Input data

In this example the two variables are gender and promotions. We have recorded how many promotions were handed out and how many people did not get promotions. Within those two data sets we kept track of gender. We want to know whether male or female nurses at a local hospital are more likely to gain promotions within the first 2 years of employment with a confidence level of 95 % ($\alpha = .05$). In other words are the promotions independent of gender? Is there a statistically significant relationship between gender and promotions (H_a) or are these differences simply due to chance (H_o)? The data are provided in the Fig. 12.3.

In Fig. 12.3 there are 237 nurses at this hospital, 117 of whom are female while 120 are male. At first review of the data it appears that female nurses have received more promotions (46) than male nurses (37). The null hypothesis is that the two variables are independent, that the likelihood of getting promoted is the same for males and females.

Rejection Rule. The goodness of fit problems, when the calculated X^2 statistic ($X^2{}_{calc}$) exceeds the critical X^2 value ($X^2{}_{\alpha}$) for your selected probability level, then we can reject the null hypothesis of equal distributions . If you select an $\alpha = 0.05$ (CI = 95 %) then you are accepting a 5 % chance that your significance finding is wrong. If the rejection rule provided significant results, there will still be a 5 % chance the results are not significant. If the results were indicated as non-significant, there would still be a 5 % chance the results are significant.

Excel

Excel examples will be provided for each of the two X^2 problem types:

- Example 1: Goodness-of-Fit
- Example 2: Independence of Two Variables

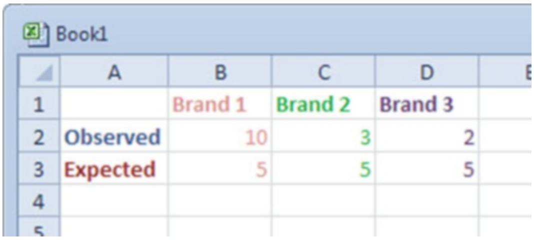
Example 1: Goodness of Fit

You are planning on stocking your grocery counter with chewing gum so you need to know the preferred brand. You have limited time and limited money so you survey a very small sample of your customers, n = 15. You ask them which is their preferred brand of gum {Brand 1, Brand 2, Brand 3}. You want a 90 % confidence level (α = 0.10). Test to make sure the results are in fact statistically significant. The data is listed below:

Brand 1: 10
Brand 2: 3
Brand 3: 2

This is a problem involving goodness of fit with the X^2 test.

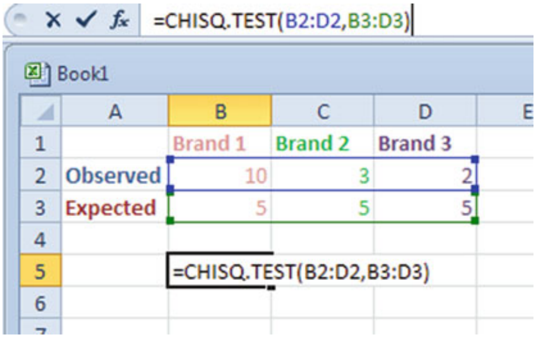
➢ Enter the data and the labels as indicated in Fig. 12.3



	A	B	C	D	E
1		Brand 1	Brand 2	Brand 3	
2	Observed	10	3	2	
3	Expected	5	5	5	
4					

To calculate the expected values, we know that the brands all have an equal chance of being selected and there are 15 possible outcomes, so each gum has a 1 in 5 chance of being selected.

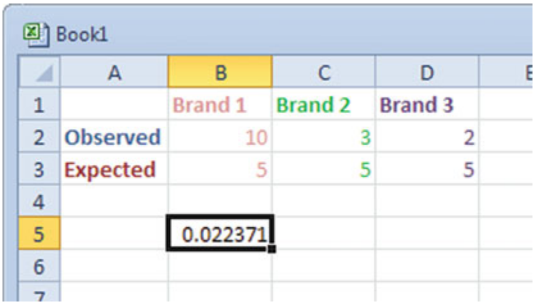
➢ Click on cell B5
➢ Click on white input ribbon and type = CHISQ.TEST(B2:D2, B3:D3)



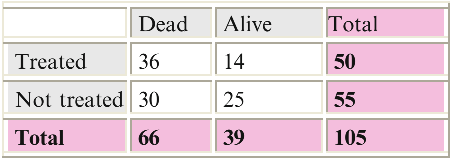
=CHISQ.TEST(B2:D2,B3:D3)

	A	B	C	D	E
1		Brand 1	Brand 2	Brand 3	
2	Observed	10	3	2	
3	Expected	5	5	5	
4					
5		=CHISQ.TEST(B2:D2,B3:D3)			
6					

➢ Click on the checkmark



	A	B	C	D	E
1		Brand 1	Brand 2	Brand 3	
2	Observed	10	3	2	
3	Expected	5	5	5	
4					
5		0.022371			
6					



	Dead	Alive	Total
Treated	36	14	**50**
Not treated	30	25	**55**
Total	**66**	**39**	**105**

Fig. 12.4 Number of plants that survived a treatment

This gives the result **0.0224.** This p-value indicates that there is a significant difference between the observed frequencies and the expected frequencies, which is unlikely to be simply due to chance error since the calculated p-value is less than the alpha value of 0.05. These results are statistically significant; in other words, you should stock more Brand 1 because your customers really are more likely to purchase this type of gum.

You could also run the significance test using the X^2 statistic test where you compare the X^2_{calc} with the X^2_{critical}. If $X^2_{\text{calc}} > X^2_{\text{critical}}$ you can reject H_0 and acknowledge that the observed distribution is real and not just due to chance.

Example 2: Testing Independence

Your company conducted an insecticide trial on a group of plants. The data represents two data sets in which one group of plants received insecticide and the other group did not. Within these two data sets you also recorded whether the plant lived or died. You hypothesized that the plants being treated with the insecticide would survive better than those that did not receive the insecticide (H_a). The null hypothesis (H_o), that you hope is not true, would be that there is no statistically significant relationship between plant survival and insecticide treatment.

H_o: The survival of the plants is independent of insecticide treatment
H_a: The survival of the plants is associated with insecticide treatment

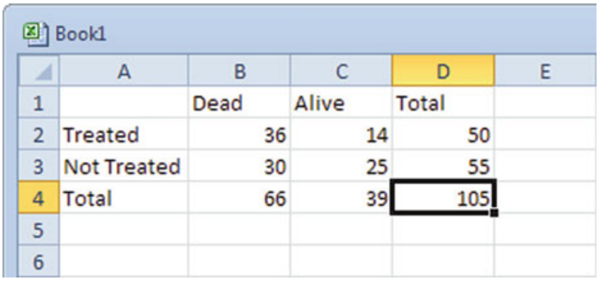
⟐ Note $H_a = H_\alpha$ in terms of terminology

You conduct the study and collect the data as shown in Fig. 12.4.

Before you build your marketing claims you need to be sure that there is a real relationship between the insecticide treatment and the number of surviving plants at the 95 % confidence level. Your company wants to avoid being sued for false advertising.

- ➢ Input **Dead** into B1
- ➢ Input **Alive** into C1
- ➢ Input **Total** into D1
- ➢ Input **Treated** into A2
- ➢ Input **Not Treated** into A3
- ➢ Input **Total** into A4

➢ Fill in cells B2 through D4 with the data as given in Fig. 12.4

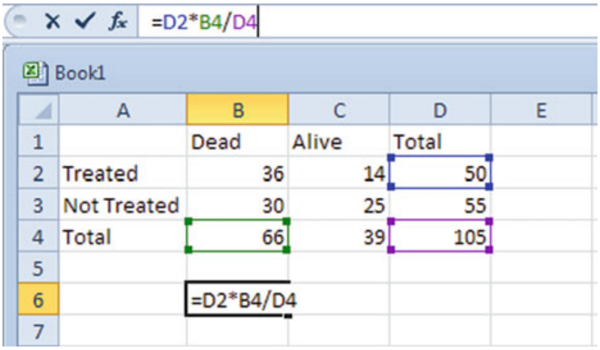


Book1

	A	B	C	D	E
1		Dead	Alive	Total	
2	Treated	36	14	50	
3	Not Treated	30	25	55	
4	Total	66	39	105	
5					
6					

Calculate expected values E(x) = (row total × column total)/N

➢ In cell B6 type = **D2 * B4/D4** into the input ribbon



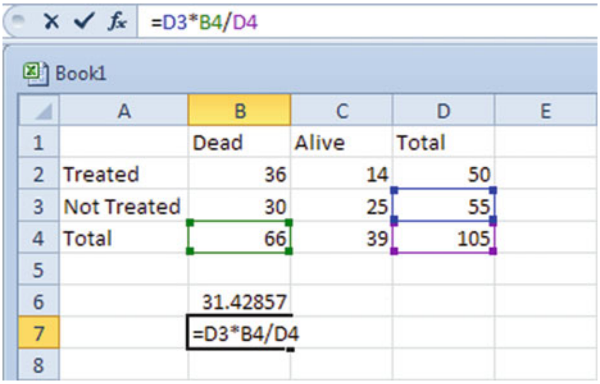
=D2*B4/D4

Book1

	A	B	C	D	E
1		Dead	Alive	Total	
2	Treated	36	14	50	
3	Not Treated	30	25	55	
4	Total	66	39	105	
5					
6		=D2*B4/D4			
7					

➢ Click green checkmark

➢ In cell B7 type = **D3 * B4/D4** into the input ribbon



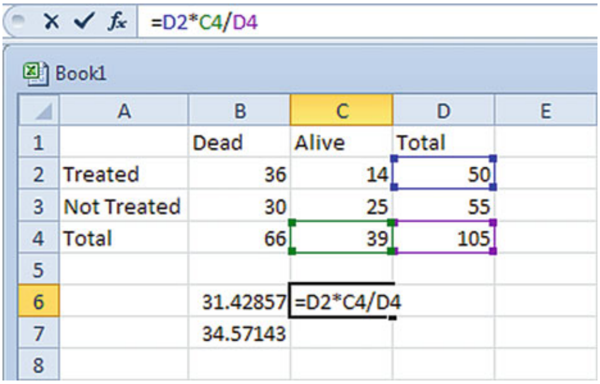
=D3*B4/D4

Book1

	A	B	C	D	E
1		Dead	Alive	Total	
2	Treated	36	14	50	
3	Not Treated	30	25	55	
4	Total	66	39	105	
5					
6		31.42857			
7		=D3*B4/D4			
8					

➢ Click the green checkmark

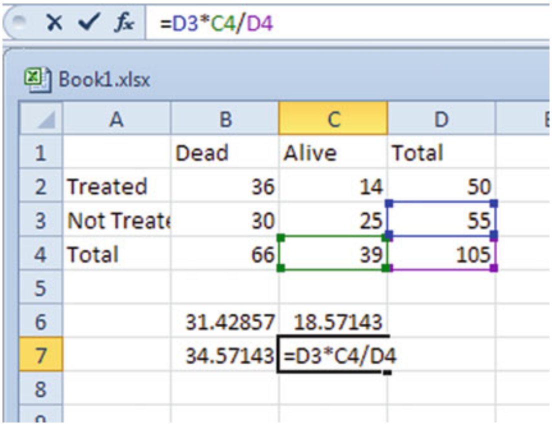
➢ In cell C6 type = **D2 * C4/D4** into the input ribbon



=D2*C4/D4

	A	B	C	D	E
1		Dead	Alive	Total	
2	Treated	36	14	50	
3	Not Treated	30	25	55	
4	Total	66	39	105	
5					
6		31.42857	=D2*C4/D4		
7		34.57143			
8					

- ➢ Click the green checkmark
- ➢ In cell C7 type = **D3 * C4/D4** into the input ribbon



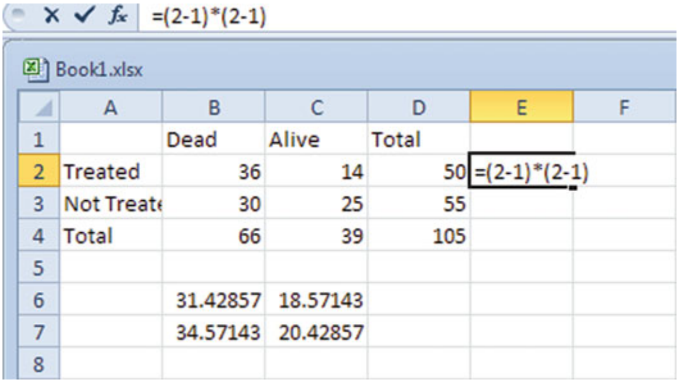
=D3*C4/D4

	A	B	C	D
1		Dead	Alive	Total
2	Treated	36	14	50
3	Not Treate	30	25	55
4	Total	66	39	105
5				
6		31.42857	18.57143	
7		34.57143	=D3*C4/D4	
8				

- ➢ Click the green checkmark

Calculate **degrees of freedom**: (number of columns minus one) * (number of rows minus one). No real easy way to do this except to actually count the rows and columns in your pivot table and use Excel as a calculator.

- ➢ In cell E2 type = **(2 − 1) * (2 − 1)** into the input ribbon

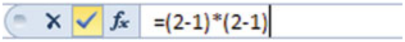


=(2-1)*(2-1)

Book1.xlsx

	A	B	C	D	E	F
1		Dead	Alive	Total		
2	Treated	36	14	50	=(2-1)*(2-1)	
3	Not Treat	30	25	55		
4	Total	66	39	105		
5						
6		31.42857	18.57143			
7		34.57143	20.42857			
8						

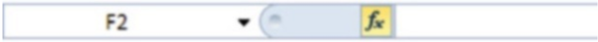
- ➢ Click the green checkmark



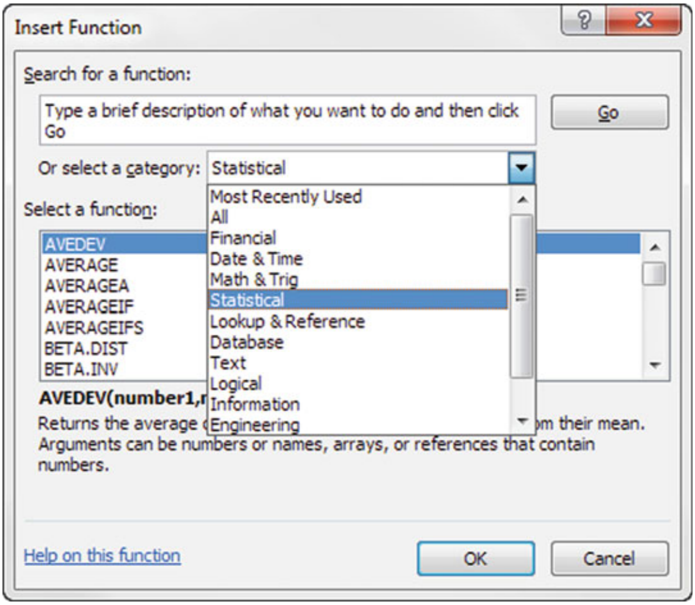
=(2-1)*(2-1)

Calculate the p-value

- ➢ Click on cell F2
- ➢ Click on the ***fx*** button



- ➢ A dialog box will appear. Select **Statistical** from drop-down menu next to “or select a category”



Insert Function

Search for a function:

Type a brief description of what you want to do and then click Go

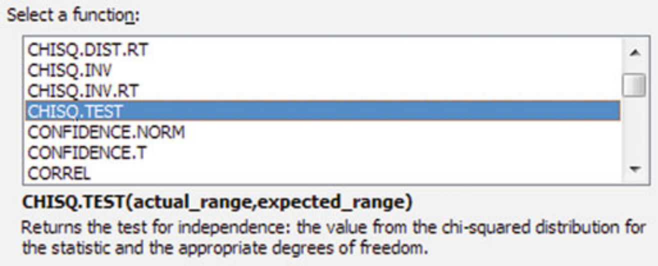
Or select a category: Statistical

Select a function:

AVEDEV(number1,

Returns the average ... m their mean. Arguments can be numbers or names, arrays, or references that contain numbers.

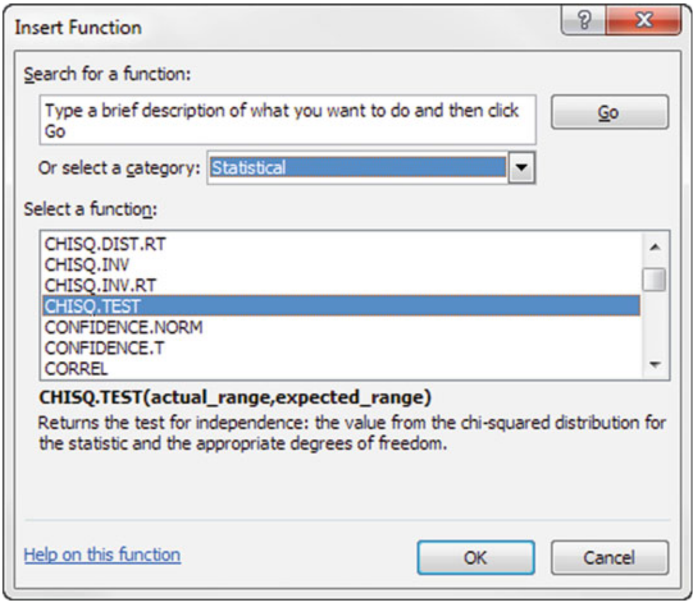
➢ Select **CHISQ.TEST** under “Select a function:”



CHISQ.TEST(actual_range,expected_range)

Returns the test for independence: the value from the chi-squared distribution for the statistic and the appropriate degrees of freedom.

➢ Click **OK**

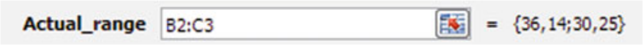


Or select a category: Statistical

CHISQ.TEST(actual_range,expected_range)

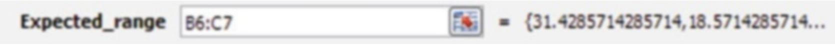
Returns the test for independence: the value from the chi-squared distribution for the statistic and the appropriate degrees of freedom.

➢ Select the **Actual Range** (cells B2 through C3)



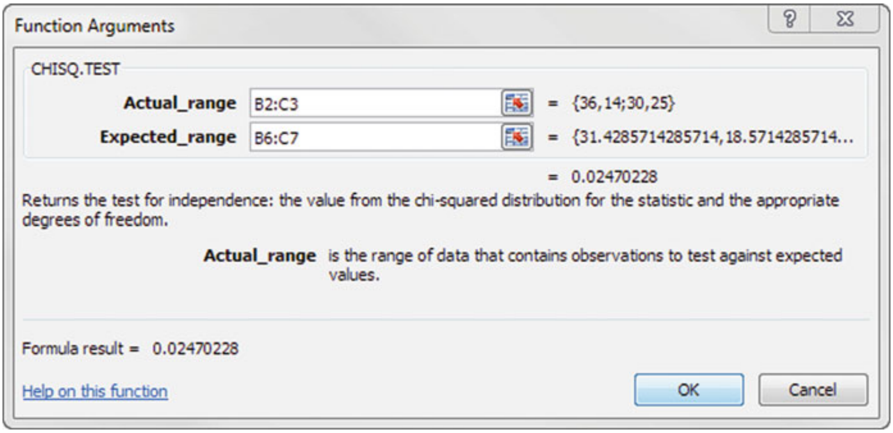
Actual_range B2:C3 = {36,14;30,25}

➢ Place cursor on **Expected Range** (cells B5 through C7)



Expected_range B6:C7 = {31.4285714285714,18.5714285714...

➢ Click **OK**



Function Arguments

CHISQ.TEST

Actual_range: B2:C3 = {36,14;30,25}

Expected_range: B6:C7 = {31.4285714285714,18.5714285714...

= 0.02470228

Returns the test for independence: the value from the chi-squared distribution for the statistic and the appropriate degrees of freedom.

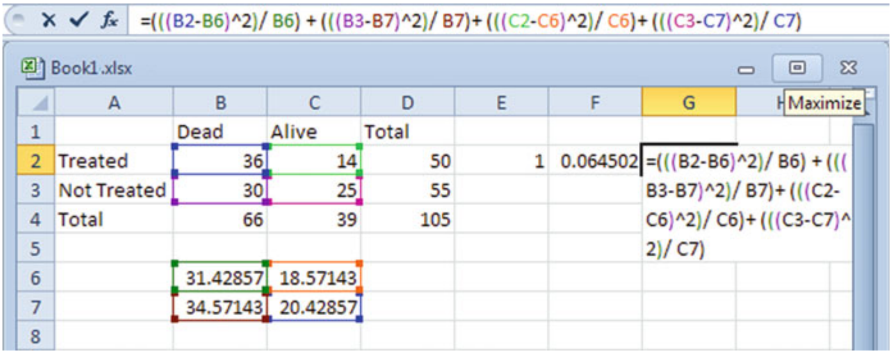
Actual_range is the range of data that contains observations to test against expected values.

Formula result = 0.02470228

Calculate the actual test statistic X^2 if you need it for reporting purposes.

◇ Note: In most cases knowing the p-value is sufficient for basic analysis

➢ In cell G2, type "=(((B2 − B6)^2)/B6) + (((B3 − B7)^2)/B7) + (((C2 − C6)^2)/C6) + (((C3 − C7)^2)/C7)"



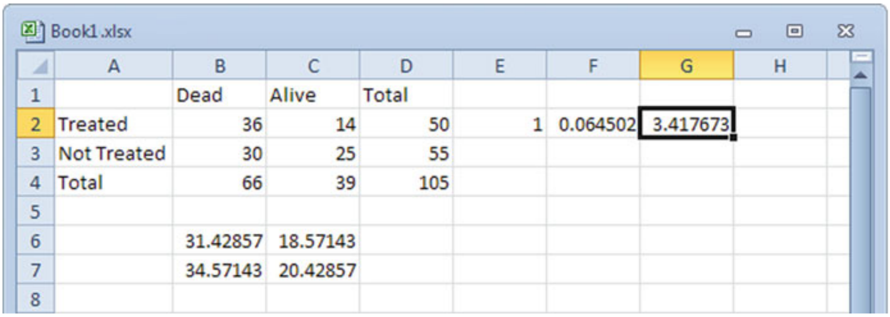
=(((B2-B6)^2)/ B6) + (((B3-B7)^2)/ B7)+ (((C2-C6)^2)/ C6)+ (((C3-C7)^2)/ C7)

Book1.xlsx

	A	B	C	D	E	F	G
1		Dead	Alive	Total			
2	Treated	36	14	50	1	0.064502	=(((B2-B6)^2)/ B6) + (((B3-B7)^2)/ B7)+ (((C2-C6)^2)/ C6)+ (((C3-C7)^2)/ C7)
3	Not Treated	30	25	55			
4	Total	66	39	105			
5							
6		31.42857	18.57143				
7		34.57143	20.42857				
8							

➢ Click the green checkmark

The final worksheet will look like Fig. 12.5. The reported outcome statistics using just the p-value data would read X^2 (1) = 3.42 p = 0.065. These are not statistically significant results for our required level of confidence since the calculated p value (0.065) is more than the p value of our selected confidence level (0.05). Likewise, if we had calculated the $X^2_{critical}$ we would have found that the $X^2_{calc} < X^2_{critical}$. We can state that H_o is true; the survival of the plants is not dependent on the insecticide treatment. We cannot claim our insecticide has a statistically significant relationship with the health of plants.



	A	B	C	D	E	F	G	H
1		Dead	Alive	Total				
2	Treated	36	14	50	1	0.064502	3.417673	
3	Not Treated	30	25	55				
4	Total	66	39	105				
5								
6		31.42857	18.57143					
7		34.57143	20.42857					
8								

Fig. 12.5 Output

z-Test

The **z-test** is used when you **know the population mean and standard deviation**. Thus, it is obviously more accurate than the t-test where you must approximate these input statistics.

◇ The z-test works best with data sets that have at least 30 data points.

The z-test can be used in the following situations:

- **Example 1 (Mean or proportion vs. a standard):** The recent CPA exam results which are normally distributed in California averaged 78 % but the US standard is 82 %. We could use a z-test to determine if these results are real or simply due to chance in a 90 % confidence interval. The data follow a normal distribution.
- **Example 2 (Testing the means or proportions of two populations/Matched pair test):** To use this test to compare the means of two populations you need large ($n \geq 30$) samples. The average employee ethics score for the Plano office was 66 %. The average score for the Detroit office was 71 %. The z-test can help determine if these results are statistically significant or just due to chance with 90 % confidence. The data follow a normal distribution.

The second part of this test includes the "Matched pair" which is used to compare the means before and after something is done to the samples. The z-test is used when the samples are large ($n \geq 30$). The variable is the difference between the before and after measurements. The employees of Aaron Building Supplies completed a workshop on effective communication. The company is wondering if the workshop is effective so they measure communication patterns before and after the workshop.

This test establishes the critical z values in a normal distribution for a given confidence level. It determines if your sample statistic falls within the confidence level. In the **z-test** the critical value is determined solely by your chosen significance level, whereas in a **t-test** the critical values are determined by the significance level and the sample size.

Rejection Rule.

$$\text{Reject } H_0 \text{ if } \text{ABS}[z_{\text{calculated}}] > \text{ABS } [z_{\text{critical}}]$$

The α indicates the confidence level you choose for the analysis. In other words the $z_{\alpha/2}$ is the critical or boundary value defining the rejection areas in the tail(s). If you select an $\alpha = 0.05$ (CI = 95 %) then you are accepting a 5 % chance that your significance finding is wrong. On the flipside if the results were indicated as non-significant, there would still be a 5 % chance the results are significant. Remember the only way you can be 100 % sure with your results is to use the entire population of data available, otherwise there is always a probability associated with your results.

Excel

Excel examples will be provided for each of the z-test problem types.

Example 1: Mean or proportion vs. a standard
Example 2: Testing the means or proportions of two populations/Matched pair test

The Z.TEST Excel tool will also be described with an example as an alternate means of comparing mean or proportion data with a standard.

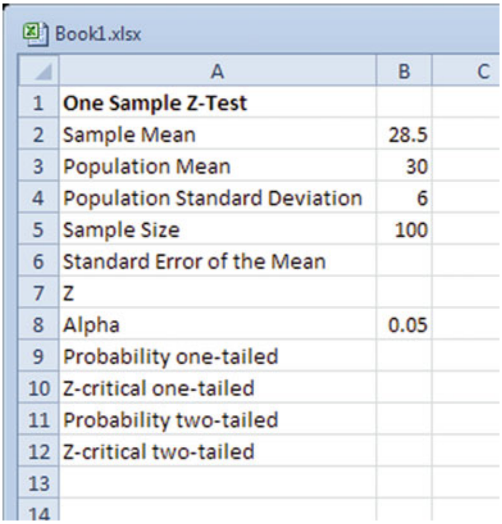
Example 1: Z-Test One Sample Mean Versus a Standard

◇ **Note: This can also be used in the same way for Proportion vs. a Standard**

Of all of the individuals (population) who develop sunburn, suppose the mean recovery time for individuals is 30 days with standard deviation equal to 6 days. A pharmaceutical company manufacturing a certain type of aloe lotion wishes to determine whether the lotion shortens, extends, or has no effect on the recovery time. The company chooses a random sample of 100 individuals who have used the lotion, and determines that the mean recovery time for these individuals was 28.5 days. They want the analysis to incorporate a 95 % confidence level (alpha = 0.05)

(a) Is the lotion effective in reducing the recovery time? (One-tailed z-test)
(b) Does the lotion have any effect at all? (Two-tailed z-test)

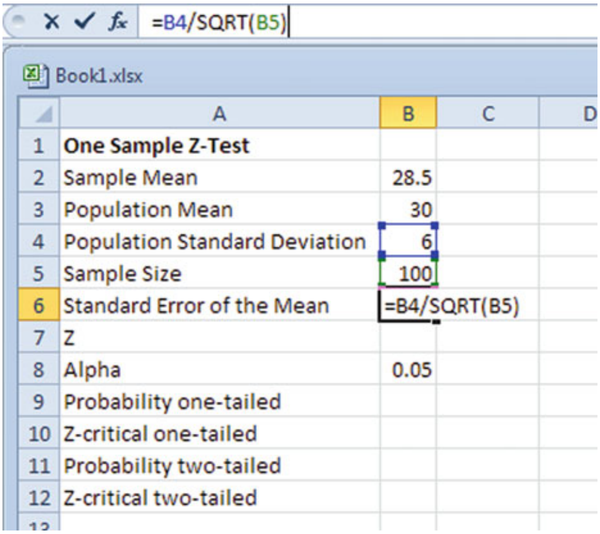
➢ Prepare an output table by typing in labels and data from Fig. 12.6



	A	B	C
1	**One Sample Z-Test**		
2	Sample Mean	28.5	
3	Population Mean	30	
4	Population Standard Deviation	6	
5	Sample Size	100	
6	Standard Error of the Mean		
7	Z		
8	Alpha	0.05	
9	Probability one-tailed		
10	Z-critical one-tailed		
11	Probability two-tailed		
12	Z-critical two-tailed		
13			
14			

Fig. 12.6 Input

➢ To calculate **Standard Error of the Mean** input the formula in cell B6 (across from its label) = **B4/SQRT (B5)**

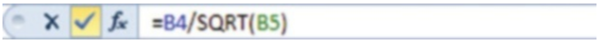


=B4/SQRT(B5)

	A	B	C	D
1	**One Sample Z-Test**			
2	Sample Mean	28.5		
3	Population Mean	30		
4	Population Standard Deviation	6		
5	Sample Size	100		
6	Standard Error of the Mean	=B4/SQRT(B5)		
7	Z			
8	Alpha	0.05		
9	Probability one-tailed			
10	Z-critical one-tailed			
11	Probability two-tailed			
12	Z-critical two-tailed			
13				

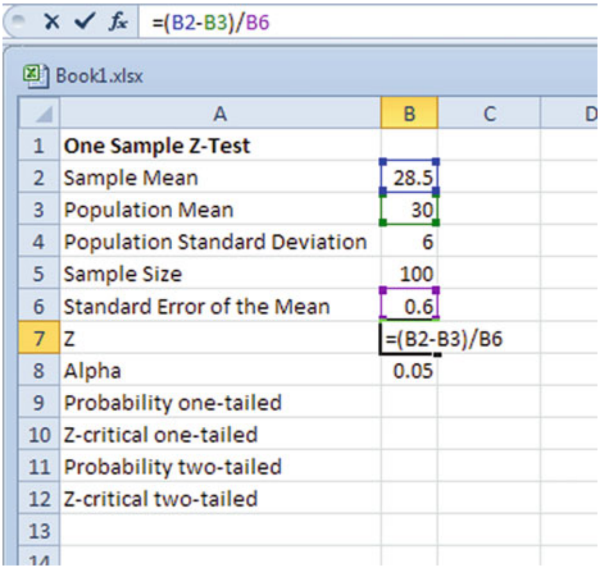
◇ Remember the standard error of the mean provides an approximation of the average amount by which the sample mean deviates from the population mean.

➢ Click the **checkmark**



=B4/SQRT(B5)

➢ Click on cell B7. To calculate the z-statistic click on the white formula ribbon and insert = **(B2 − B3)/(B6)**



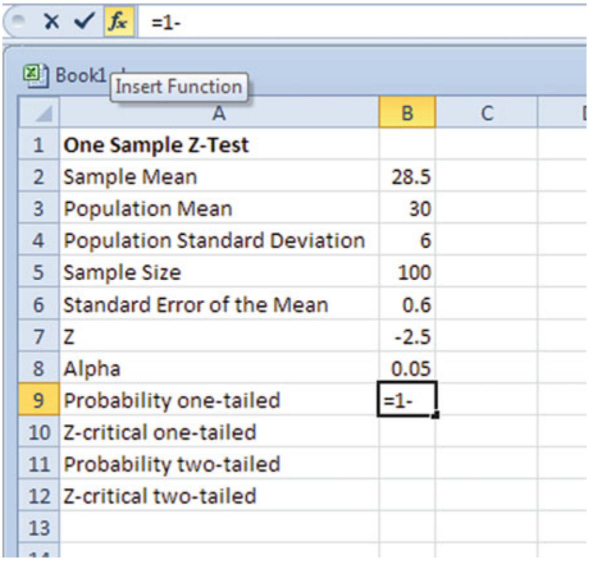
=(B2-B3)/B6

Book1.xlsx

	A	B	C	D
1	**One Sample Z-Test**			
2	Sample Mean	28.5		
3	Population Mean	30		
4	Population Standard Deviation	6		
5	Sample Size	100		
6	Standard Error of the Mean	0.6		
7	Z	=(B2-B3)/B6		
8	Alpha	0.05		
9	Probability one-tailed			
10	Z-critical one-tailed			
11	Probability two-tailed			
12	Z-critical two-tailed			
13				

➢ Click the checkmark

➢ Click on cell B9. To calculate the probability for a one-tailed test click on the white formula ribbon and insert = **1-.** Now click on the *fx* symbol to the left of the input ribbon

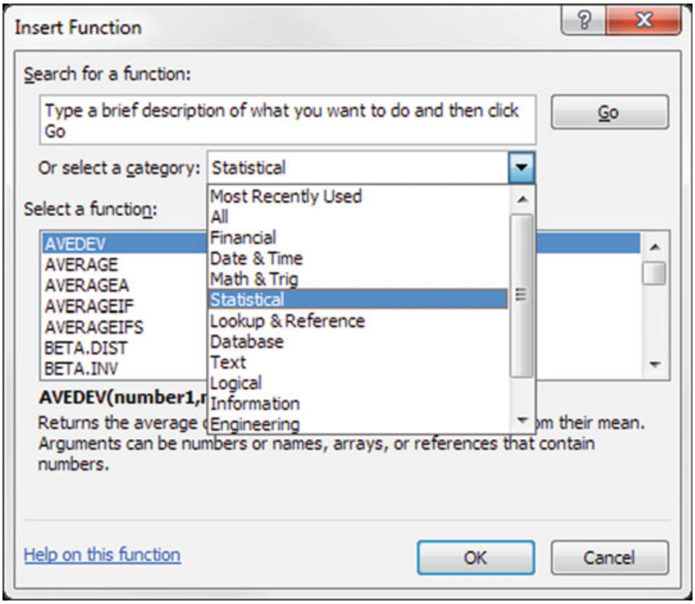


=1-

Book1

	A	B	C
1	**One Sample Z-Test**		
2	Sample Mean	28.5	
3	Population Mean	30	
4	Population Standard Deviation	6	
5	Sample Size	100	
6	Standard Error of the Mean	0.6	
7	Z	-2.5	
8	Alpha	0.05	
9	Probability one-tailed	=1-	
10	Z-critical one-tailed		
11	Probability two-tailed		
12	Z-critical two-tailed		
13			

- ➢ In the pop-up window, select the category **Statistical**



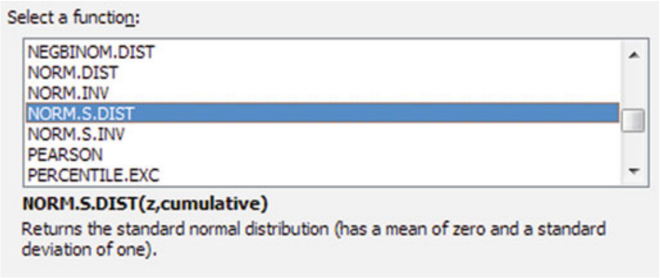
Type a brief description of what you want to do and then click Go

Or select a category: Statistical

AVEDEV(number1,r

Returns the average (m their mean. Arguments can be numbers or names, arrays, or references that contain numbers.

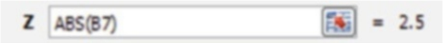
- ➢ Select **NORM.S.DIST** (Remember from Chap. 4 **NORM.S.DIST** tells you the distribution to the left of this z statistic for a mean of zero and a SD of one)



NORM.S.DIST(z,cumulative)

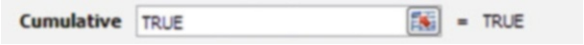
Returns the standard normal distribution (has a mean of zero and a standard deviation of one).

- ➢ Click **OK**
- ➢ Type **ABS(B7)** for **Z** which provides the absolute value which makes the task of interpreting the statistical significance of z easier



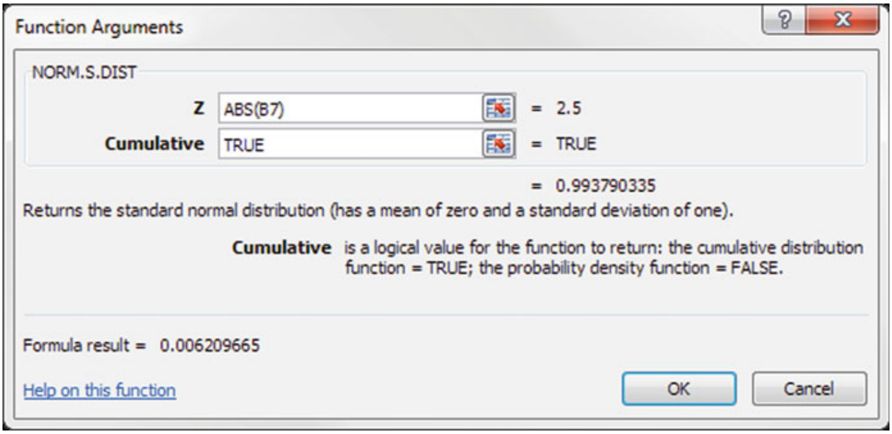
Z ABS(B7) = 2.5

- ➢ Type **TRUE** in the cumulative box to ensure the cumulative (everything to the left) distribution is calculated



Cumulative TRUE = TRUE

➢ Click **OK**



NORM.S.DIST

Z ABS(B7) = 2.5

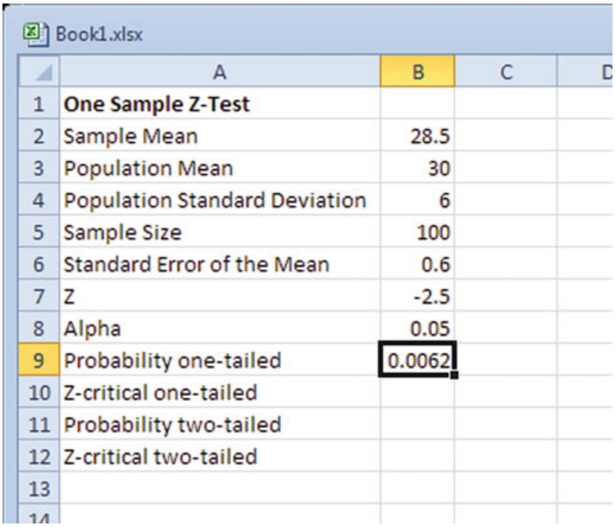
Cumulative TRUE = TRUE

= 0.993790335

Returns the standard normal distribution (has a mean of zero and a standard deviation of one).

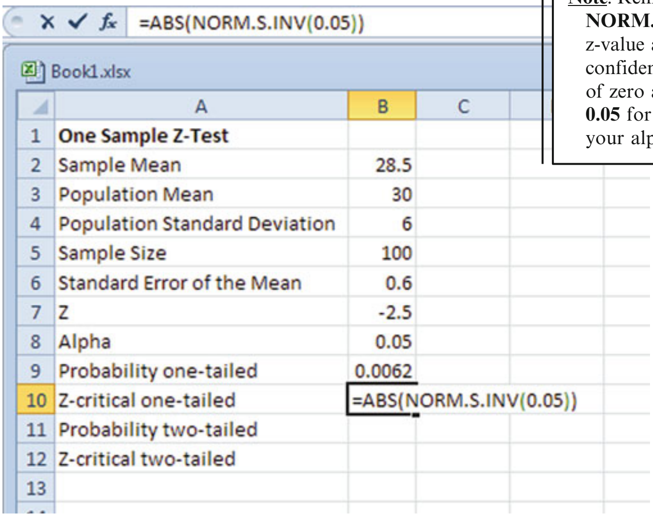
Cumulative is a logical value for the function to return: the cumulative distribution function = TRUE; the probability density function = FALSE.

Formula result = 0.006209665



	A	B	C	D
1	**One Sample Z-Test**			
2	Sample Mean	28.5		
3	Population Mean	30		
4	Population Standard Deviation	6		
5	Sample Size	100		
6	Standard Error of the Mean	0.6		
7	Z	-2.5		
8	Alpha	0.05		
9	Probability one-tailed	0.0062		
10	Z-critical one-tailed			
11	Probability two-tailed			
12	Z-critical two-tailed			
13				
14				

➢ Click on cell B10. To calculate the z-Critical for a one-tailed test click on the white formula ribbon and insert = **ABS(NORM.S.INV(0.05))**



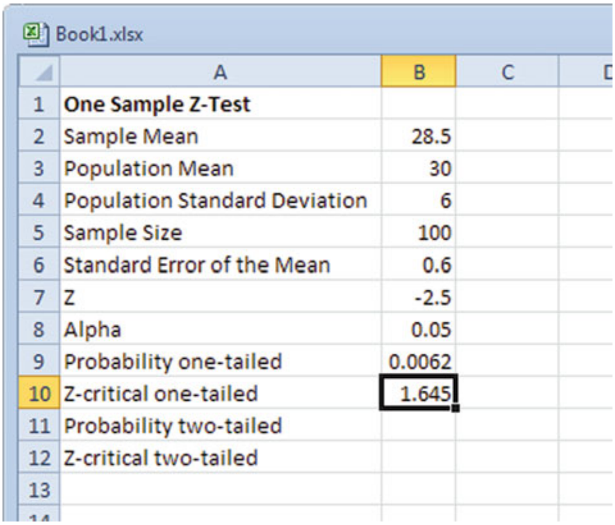
=ABS(NORM.S.INV(0.05))

	A	B	C
1	**One Sample Z-Test**		
2	Sample Mean	28.5	
3	Population Mean	30	
4	Population Standard Deviation	6	
5	Sample Size	100	
6	Standard Error of the Mean	0.6	
7	Z	-2.5	
8	Alpha	0.05	
9	Probability one-tailed	0.0062	
10	Z-critical one-tailed	=ABS(NORM.S.INV(0.05))	
11	Probability two-tailed		
12	Z-critical two-tailed		
13			

Note: Remember from Chapter 4 **NORM.S.INV** tells you the z-value associated with your confidence interval usinga mean of zero and a SD of one. Use **0.05** for the **Probability** which is your alpha value for a CI of 95%.

➢ Click the **Checkmark**

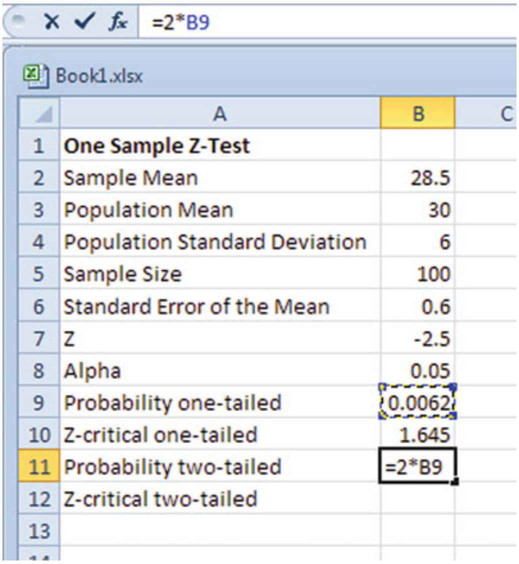
- Note: You will get an error if you do not include closing brackets at the end of the formula. So remember to add the brackets.



	A	B	C
1	**One Sample Z-Test**		
2	Sample Mean	28.5	
3	Population Mean	30	
4	Population Standard Deviation	6	
5	Sample Size	100	
6	Standard Error of the Mean	0.6	
7	Z	-2.5	
8	Alpha	0.05	
9	Probability one-tailed	0.0062	
10	Z-critical one-tailed	1.645	
11	Probability two-tailed		
12	Z-critical two-tailed		
13			

- Note: you will always get this same answer for z-critical in all problems whenever you use a confidence level of 95 % in a one tailed test. You could also look this up in the standard z- table.

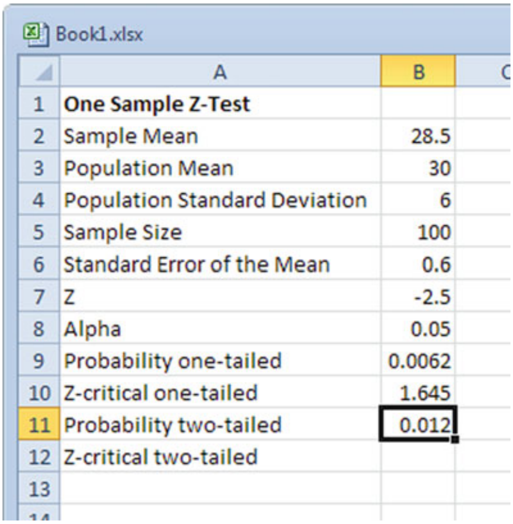
➢ Click on cell **B11.** Because the two-tailed probability is just twice the one-tailed probability type in the formula ribbon = **2 * B9**



=2*B9

	A	B
1	**One Sample Z-Test**	
2	Sample Mean	28.5
3	Population Mean	30
4	Population Standard Deviation	6
5	Sample Size	100
6	Standard Error of the Mean	0.6
7	Z	-2.5
8	Alpha	0.05
9	Probability one-tailed	0.0062
10	Z-critical one-tailed	1.645
11	Probability two-tailed	=2*B9
12	Z-critical two-tailed	
13		

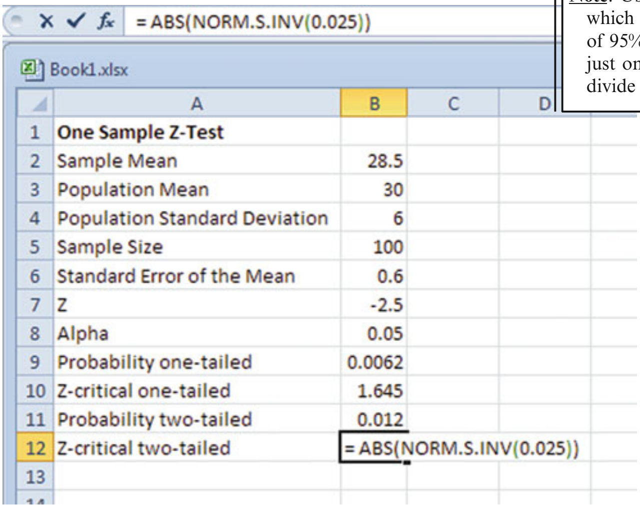
➢ Click on the **Checkmark**



	A	B
1	**One Sample Z-Test**	
2	Sample Mean	28.5
3	Population Mean	30
4	Population Standard Deviation	6
5	Sample Size	100
6	Standard Error of the Mean	0.6
7	Z	-2.5
8	Alpha	0.05
9	Probability one-tailed	0.0062
10	Z-critical one-tailed	1.645
11	Probability two-tailed	0.012
12	Z-critical two-tailed	
13		

➢ Click on cell B12. To calculate the z-Critical for a two-tailed test click on the white formula ribbon and insert = **ABS(NORM.S.INV(0.025))**

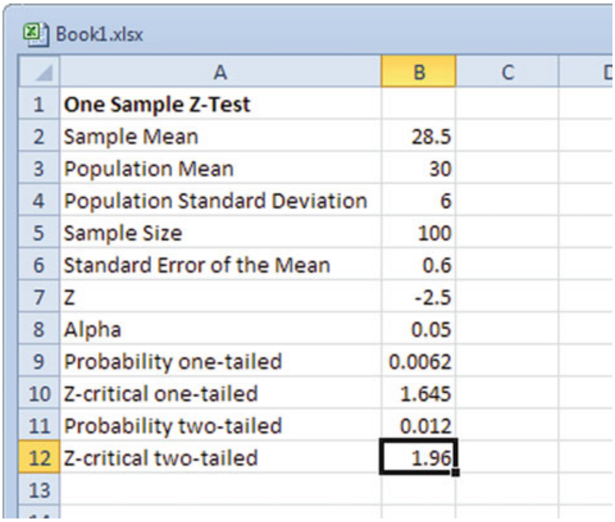
Note: Use **0.025** for **Probability**, which is your alpha value for a CI of 95% divided into two tails not just one. As such we need to divide the alpha value in half.



= ABS(NORM.S.INV(0.025))

	A	B	C	D
1	**One Sample Z-Test**			
2	Sample Mean	28.5		
3	Population Mean	30		
4	Population Standard Deviation	6		
5	Sample Size	100		
6	Standard Error of the Mean	0.6		
7	Z	-2.5		
8	Alpha	0.05		
9	Probability one-tailed	0.0062		
10	Z-critical one-tailed	1.645		
11	Probability two-tailed	0.012		
12	Z-critical two-tailed	= ABS(NORM.S.INV(0.025))		
13				

➢ Click on the **Checkmark**



	A	B	C	D
1	**One Sample Z-Test**			
2	Sample Mean	28.5		
3	Population Mean	30		
4	Population Standard Deviation	6		
5	Sample Size	100		
6	Standard Error of the Mean	0.6		
7	Z	-2.5		
8	Alpha	0.05		
9	Probability one-tailed	0.0062		
10	Z-critical one-tailed	1.645		
11	Probability two-tailed	0.012		
12	Z-critical two-tailed	1.96		
13				

One-Tailed Results

The calculated **z-statistic** (**Z**) for this problem is −2.5. The probability of obtaining this $\mathbf{z_{calculated}}$ in the one-tailed test is 0.006 (Probability one-tailed). We check to see if the probability of getting this **z-statistic** is less than our alpha value.

Since 0.006 is less than 0.05 (Alpha) we can conclude that our results of obtaining a **z-statistic** of −2.5 are statistically significant.

Based on our chosen alpha value of 0.05, we obtain a critical $\mathbf{z_{critical}}$ that we need to compare to the absolute value of the $\mathbf{z_{calculated}}$ of 2.5. Since the $\mathbf{z_{critical}}$ of 1.64 (Z-critical one-tail) is less that the obtained **z value** (absolute value) of 2.5 (Z), the mean value of the sample is real and not just due to chance. Therefore the shorter recovery time of 28.5 days versus 30 days indicates the lotion is effective.

Two-Tailed Results

For the two-tailed test the calculated **z statistic (Z)** for this problem is still −2.5. The probability of obtaining this **z value** in the two-tailed test is 0.012 (Probability two-tailed). We check to see if the probability of getting this **z-statistic** is less than our alpha value. Since 0.012 is less than 0.05 (Alpha) we can conclude that our results of obtaining a **z-statistic** of −2.5 are statistically significant.

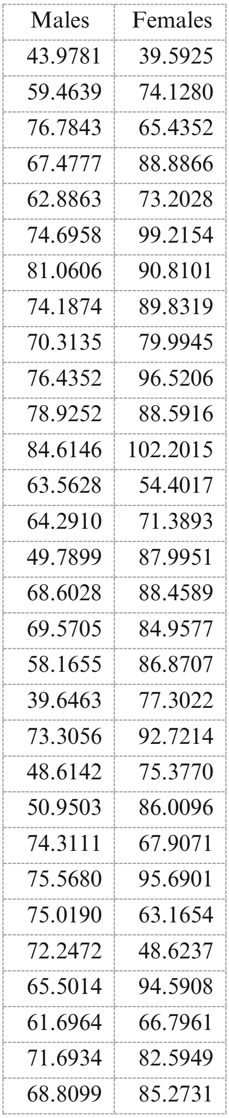
Based on our chosen alpha value of 0.05, we obtain a $\mathbf{z_{critical}}$ that we need to compare to the obtained absolute value of the **z statistic** of 2.5 (Z). Since the $\mathbf{z_{critical}}$ of 1.96 (Z-critical two-tailed) is less that the obtained **z value** (absolute value) of 2.5, the mean value of the sample is real and not just due to chance. Therefore the mean recovery time associated with the use of this lotion is significantly different for the mean of the population. With the two tailed test we don't know if the sample mean is greater than or less than the population mean, just that the difference is real and not simply due to chance.

Example 2: Testing the Means of Two Populations

◇ **Note:** This can also be used in the same way for Two Sample Proportions. This can also be used in a Matched Pair application where you need to test whether before and after results are in fact significant.

In this example we have data from the two samples, and want to test that the differences between the two are real and not just due to chance. The two samples are collected from the same sales organization; the first column included only male respondents and the second column represents female respondents. These employees recorded how many complaints they received about the company from customers over a 12 month window. Is there a real difference between the average complaints received by male employees vs. female employees?

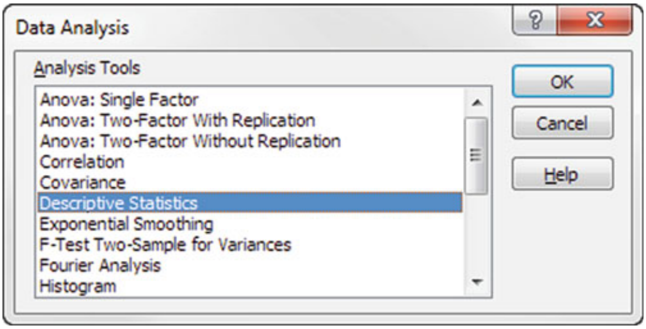
First enter in the data for your two samples, one in Column A and the other in Column B (Fig. 12.7). Note to use this test the number of entries in each column does not have to be the same, but the rule of thumb is to have at least 30 entries in each column for the test to work properly.



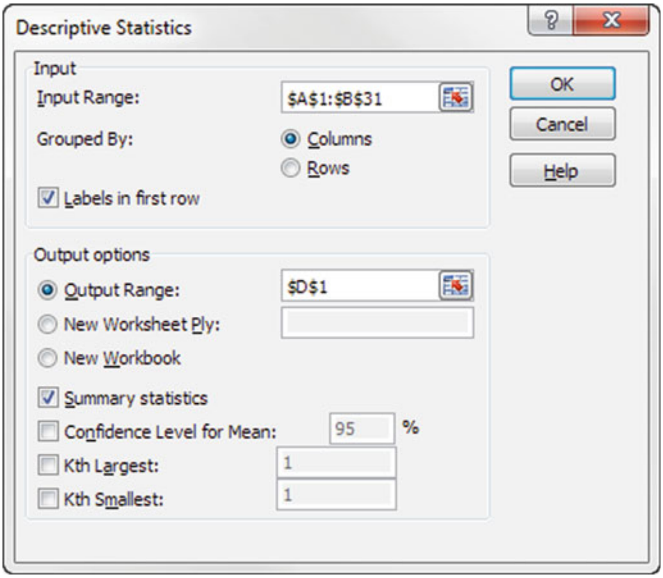
Males	Females
43.9781	39.5925
59.4639	74.1280
76.7843	65.4352
67.4777	88.8866
62.8863	73.2028
74.6958	99.2154
81.0606	90.8101
74.1874	89.8319
70.3135	79.9945
76.4352	96.5206
78.9252	88.5916
84.6146	102.2015
63.5628	54.4017
64.2910	71.3893
49.7899	87.9951
68.6028	88.4589
69.5705	84.9577
58.1655	86.8707
39.6463	77.3022
73.3056	92.7214
48.6142	75.3770
50.9503	86.0096
74.3111	67.9071
75.5680	95.6901
75.0190	63.1654
72.2472	48.6237
65.5014	94.5908
61.6964	66.7961
71.6934	82.5949
68.8099	85.2731

Fig. 12.7 Input data

➢ Open the **Data** tab and select the **Data Analysis** tool, then select **Descriptive Statistics**



➢ Highlight the two columns of data as your input field. Check the **Labels in first row box**. Select **Output Range**, click inside the box, and highlight a cell for you output location. Check the **Summary statistics** box.


Descriptive Statistics

Input

Input Range: $A$1:$B$31

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

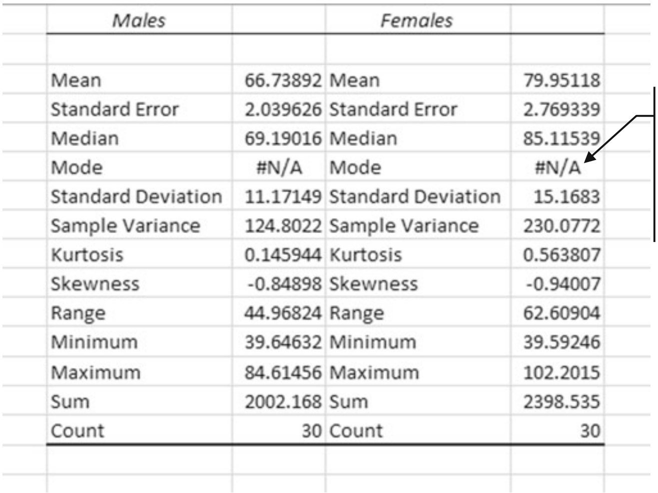
☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

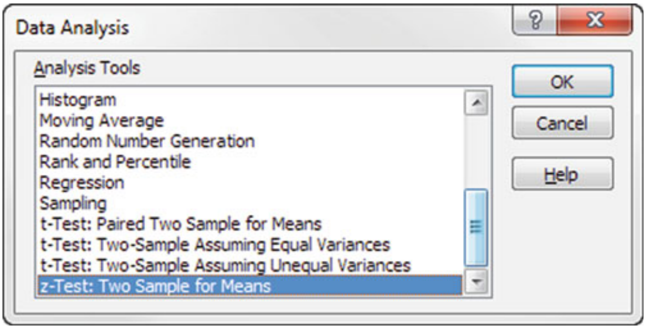
➢ Click OK



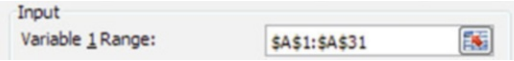
Males		Females	
Mean	66.73892	Mean	79.95118
Standard Error	2.039626	Standard Error	2.769339
Median	69.19016	Median	85.11539
Mode	#N/A	Mode	#N/A
Standard Deviation	11.17149	Standard Deviation	15.1683
Sample Variance	124.8022	Sample Variance	230.0772
Kurtosis	0.145944	Kurtosis	0.563807
Skewness	-0.84898	Skewness	-0.94007
Range	44.96824	Range	62.60904
Minimum	39.64632	Minimum	39.59246
Maximum	84.61456	Maximum	102.2015
Sum	2002.168	Sum	2398.535
Count	30	Count	30

Note: The "**#N/A**" indicates that no number in the data set is replicated; thus there is no mode.

➢ Select the **Data Analysis** tool and select **z-Test: Two Sample for Means**

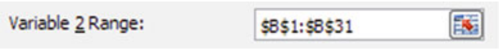

Data Analysis
?
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

➢ In the **Variable 1 Range** box, indicate the range of the first sample in Column A **($A$1:$A$31)**



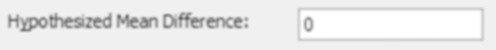
Variable 1 Range: $A$1:$A$31

➢ In **Variable 2 Range,** indicate the range of the second sample in Column B **($B$1:$B$31)**



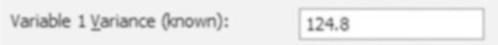
Variable 2 Range: $B$1:$B$31

➢ In the **Hypothesized Mean Difference** box, enter the figure **0**. Since the null hypothesis, which we are attempting to refute, says that both samples come from the same population, we are hypothesizing that the difference between the means for the two populations is 0.



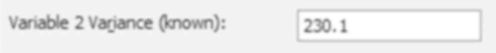
Hypothesized Mean Difference: 0

- In the **Variable 1 Variance (known)** box enter the value for the variance for the sample in the A Column from the Descriptive Statistics output. For this sample enter **124.8.** Remember that the variance is just the square of the standard deviation.



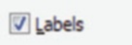
Variable 1 Variance (known): 124.8

- In the **Variable 2 Variance (known)** box enter the corresponding figure for the variance of the second sample. For this sample enter **230.1**



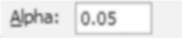
Variable 2 Variance (known): 230.1

- Check **Labels**



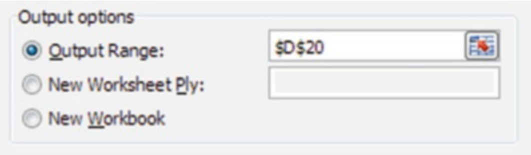
☑ Labels

- In the **Alpha:** box the number 0.05 should already appear. If it is empty or shows a number different from 0.05, then enter the number 0.05. The Alpha figure indicates the confidence level for this test. A figure of 0.05 states that you want to be 95 % certain of the result or, in other words, that you want the probability of being wrong to be .05 or lower.



Alpha: 0.05

- Select **Output Range:** then in the **Output Range:** box, click on the cell where you want the output to appear. The Output Range table will take up 3 columns and 12 horizontal rows.



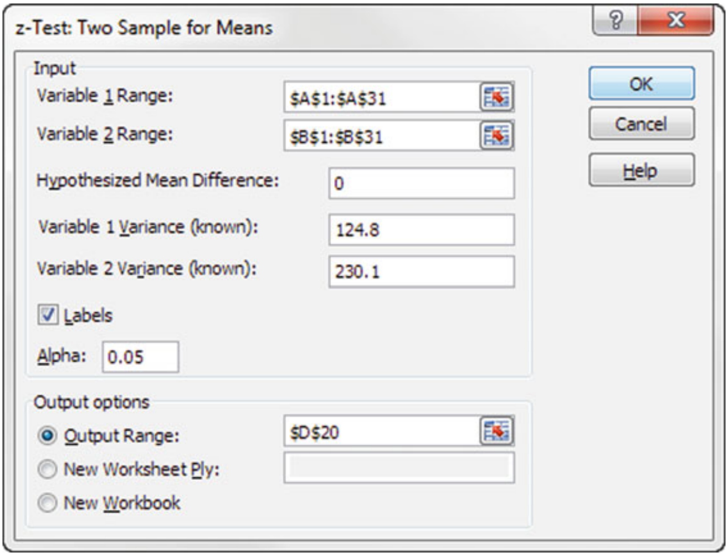
Output options

● Output Range: $D$20

○ New Worksheet Ply:

○ New Workbook

➢ Click **OK**



z-Test: Two Sample for Means

Input

Variable 1 Range: $A$1:$A$31

Variable 2 Range: $B$1:$B$31

Hypothesized Mean Difference: 0

Variable 1 Variance (known): 124.8

Variable 2 Variance (known): 230.1

☑ Labels

Alpha: 0.05

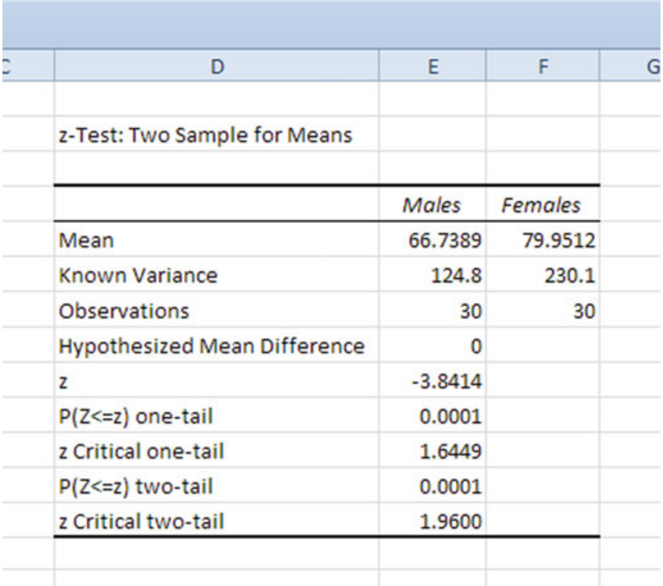
Output options

● Output Range: $D$20

○ New Worksheet Ply:

○ New Workbook

The output table in Fig. 12.8 has the heading: **z-Test: Two Samples for Means**. You will need to widen the left hand column of the table in order to read the names of the items. In the table there are results for the following items: **Mean, Known Variance, Observations, Hypothesized Mean Difference, z, P(Z <= z) one-tail, Z critical One-tail, P(Z <= z) two-tail, Z critical two-tail.**



z-Test: Two Sample for Means

	Males	*Females*
Mean	66.7389	79.9512
Known Variance	124.8	230.1
Observations	30	30
Hypothesized Mean Difference	0	
z	-3.8414	
P(Z<=z) one-tail	0.0001	
z Critical one-tail	1.6449	
P(Z<=z) two-tail	0.0001	
z Critical two-tail	1.9600	

Fig. 12.8 Output

The **Mean** values give the arithmetical average for each sample. It should be the same as the value for the **Mean** in your Descriptive Statistics chart you generated earlier.

The **Known Variance** similarly gives the variance for each sample and corresponds to the variance results generated earlier (these are the values you entered into the **z-Test** dialogue box).

The **Observations** is the number of items in each sample.

The **Hypothesized Mean Difference** should be 0 (the value you entered in the z-Test dialogue box earlier).

The **z value** indicates in standard deviation units how far from the mean the value for the difference between your two samples is located. Remember that the normal curve for all the differences between all the possible pairs of samples from the population has a mean of 0. Your two samples did not have the same mean; thus they fall away from the mean in the normal distribution. The **z value** tells you how far away the difference falls.

Following the **z value** there are four lines of results, two concerning one-tail and two concerning two-tail testing. You will use one or the other of these pairs of figures, not both. The one you use will depend upon on how you defined your alternative hypothesis, or in other words, how you define failure.

If you have set up your alternative hypothesis such that it argues that one of the populations will have a higher value than the other on your variable of interest, then you are interested only in one end of the distribution curve. This means you should use the one-tail output. In this case an example might be whether the women's higher mean number of customer complaints (80) is real when compared to the mean number of customer complaints that the men receive (67).

However, if your alternative hypothesis simply asserts that there will be a significant difference between the populations, without saying which will be higher or lower, then you need the two-tail output. For example, "There is a difference in the number of customer complaints about the company that male employees receive versus the number female employees receive."

In Fig. 12.8 the P(Z <= z) values indicate the probability that the two populations are the same, so you want these values to be very small. In Fig. 12.8 the one-tail p-test of 0.0001 suggests that the difference in the mean number of customer complaints received by women is in fact higher than those received by the men. The results are real and not just due to chance. We can be more specific by choosing our level of accepted risk with this outcome by choosing a confidence level. If we choose a 95 % confidence level, then the α value is 0.05. The one-tail p-test of 0.0001 is smaller than the value of 0.05. This means that we can be 95 % confident that the mean number of customer complaints received by women is in fact higher than those received by the men. This finding is statistically significant.

The two-tail p-test of 0.0001 is also very small indicating that the difference in the means of the two samples is real and not due to chance. If we choose a 95 % confidence level we accept that 95 out of 100 times we will get means that are different, but 5 out the 100 times we will not get these results. We can say that the means for customer complaints about the company are different for men and women and this difference is mathematically significant at the 95 % confidence level.

Another method to determine if the difference in means is real is to compare the z value with the **z-Critical** value. If the **z value** is less than **the z-Critical** value, you affirm the null hypothesis; if the z value is greater than the **z-Critical** value, you reject it, which says your difference in the means is real. In this example the **z value** of 3.84 is greater than the **z-Critical one-tail** (1.6449) and **z-Critical two-tail** (1.96). In both of these models the difference in the means is real and not due to chance. Remember that by changing the confidence level in a **z-test,** the **z-Critical** values will also change.

Z.TEST Tool for Comparing a Mean or Proportion with a Standard

Excel also provides a **Z.TEST** tool in the formula function area. This is another method for comparing a mean or proportion with a standard. The **Z.TEST** function returns the one-tailed p-value of a z-test. It can be modified to also return two-tailed results. The input format to use this **Z.TEST** function in Excel is:

$$\mathbf{Z.TEST}(\text{array of data, the population mean, the population std. deviation})$$

However we could get the same results if we used our NORMDIST formula from Chap. 4:

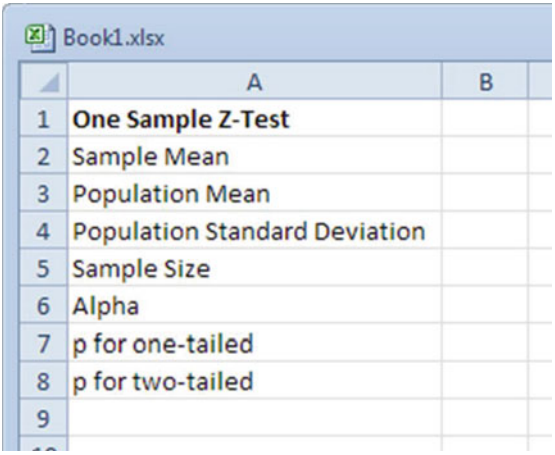
$$\begin{aligned}&= 1 - \text{NORMSDIST}((\text{sample mean} - \text{population mean})/\\&\quad(\text{population SD}/\text{SQRT}(\text{sample size})))\end{aligned}$$

Example Problem

In this section we will use the **Z.TEST** as the Excel tool. This example is based on collecting engineering proficiency quotient scores which happen to follow a normal distribution. The population standard deviation is 15 points, and the mean is 60 points. A sample of 30 engineers is collected and you need to know if based on this sample, that the sample mean is significant or a chance occurrence, at a 95 % confidence level

➢ Input the data in Fig. 12.9

Scores
43.98
59.46
76.78
67.48
62.89
74.70
81.06
74.19
70.31
76.44
78.93
84.61
63.56
64.29
49.79
68.60
69.57
58.17
39.65
73.31
48.61
50.95
74.31
75.57
75.02
72.25
65.50
61.70
71.69
68.81



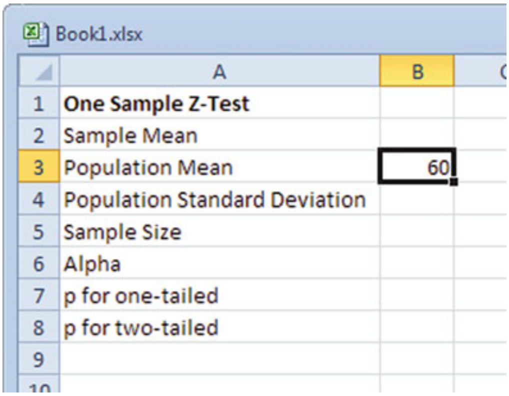
Book1.xlsx

	A	B
1	**One Sample Z-Test**	
2	Sample Mean	
3	Population Mean	
4	Population Standard Deviation	
5	Sample Size	
6	Alpha	
7	p for one-tailed	
8	p for two-tailed	
9		

Fig. 12.9 Input data and template

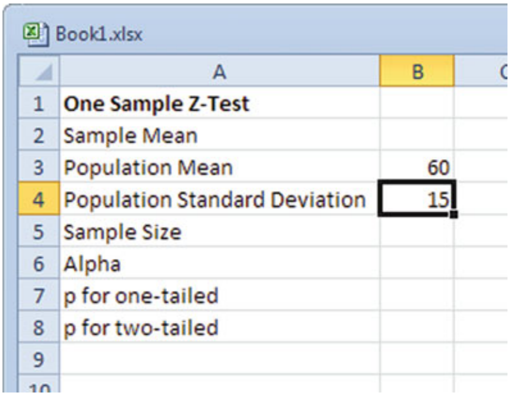
➢ Select cell D1 in your blank Excel worksheet, and then paste the entries so that the array data fills cells D1:D31 in your worksheet.

➢ Input the population mean of **60** in B3



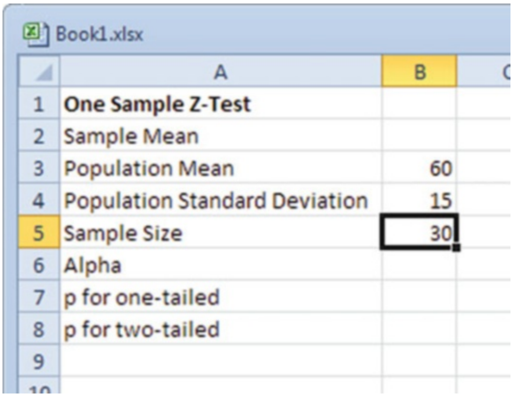
	A	B
1	**One Sample Z-Test**	
2	Sample Mean	
3	Population Mean	60
4	Population Standard Deviation	
5	Sample Size	
6	Alpha	
7	p for one-tailed	
8	p for two-tailed	
9		
10		

➢ In cell B4 input the population std. deviation of **15**



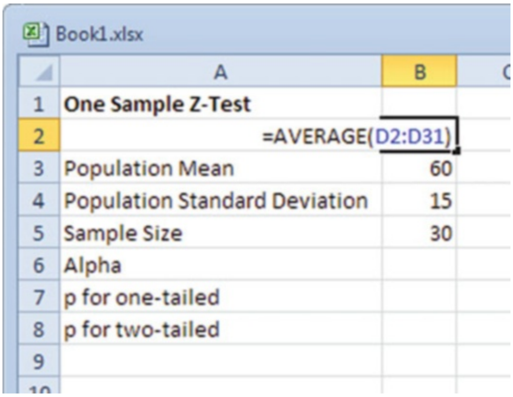
	A	B
1	**One Sample Z-Test**	
2	Sample Mean	
3	Population Mean	60
4	Population Standard Deviation	15
5	Sample Size	
6	Alpha	
7	p for one-tailed	
8	p for two-tailed	
9		
10		

➢ In cell B5 input the sample size of **30**



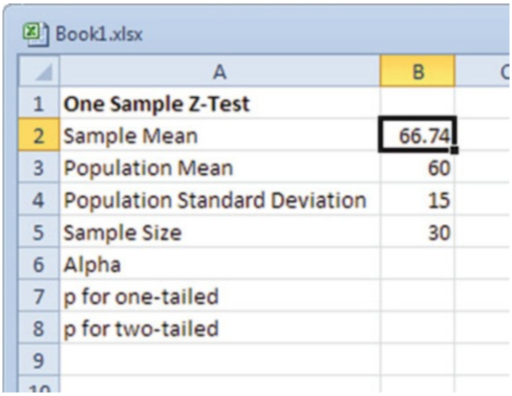
	A	B
1	**One Sample Z-Test**	
2	Sample Mean	
3	Population Mean	60
4	Population Standard Deviation	15
5	Sample Size	30
6	Alpha	
7	p for one-tailed	
8	p for two-tailed	
9		
10		

➢ Click on cell B2. Calculate the mean for the input data by clicking on the formula ribbon and typing = **AVERAGE(D2:D31)**



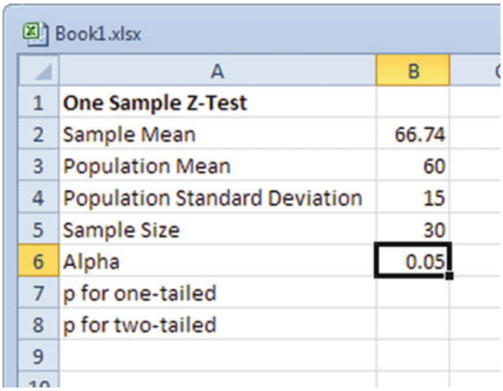
	A	B
1	**One Sample Z-Test**	
2	=AVERAGE(D2:D31)	
3	Population Mean	60
4	Population Standard Deviation	15
5	Sample Size	30
6	Alpha	
7	p for one-tailed	
8	p for two-tailed	
9		

The average engineering proficiency score for this sample is **66.74** (Cell B2).



	A	B
1	**One Sample Z-Test**	
2	Sample Mean	66.74
3	Population Mean	60
4	Population Standard Deviation	15
5	Sample Size	30
6	Alpha	
7	p for one-tailed	
8	p for two-tailed	
9		

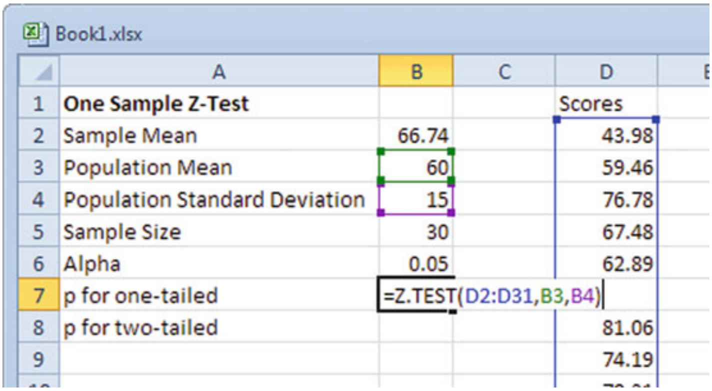
➢ Click on cell B6 and input the chosen alpha value **0.05**



	A	B
1	**One Sample Z-Test**	
2	Sample Mean	66.74
3	Population Mean	60
4	Population Standard Deviation	15
5	Sample Size	30
6	Alpha	0.05
7	p for one-tailed	
8	p for two-tailed	
9		

➢ Click on cell B7. Calculate the **z-statistic** by clicking on the formula ribbon and typing = **Z.TEST(D2:D31,B3,B4)**

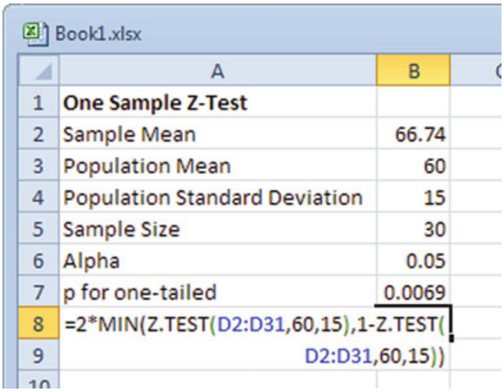
◇ <u>**Note**</u>: When the population SD is not known, the optional third argument in Z.TEST can be left blank. In this case, Excel calculates the SD from the sample data.



Book1.xlsx

	A	B	C	D
1	**One Sample Z-Test**			Scores
2	Sample Mean	66.74		43.98
3	Population Mean	60		59.46
4	Population Standard Deviation	15		76.78
5	Sample Size	30		67.48
6	Alpha	0.05		62.89
7	p for one-tailed	=Z.TEST(D2:D31,B3,B4)		
8	p for two-tailed			81.06
9				74.19

The p-value returned by Z.TEST for the one-sided (or one-tailed) test is in cell B7 (0.0069). Z.TEST confirms that if the true mean of the underlying normal distribution from which all 30 engineering scores were drawn is in fact 60 points, and the true standard deviation is 15, a sample mean that is higher than 66.74 would occur with probability 0.0069.



Book1.xlsx

	A	B
1	**One Sample Z-Test**	
2	Sample Mean	66.74
3	Population Mean	60
4	Population Standard Deviation	15
5	Sample Size	30
6	Alpha	0.05
7	p for one-tailed	0.0069
8	=2*MIN(Z.TEST(D2:D31,60,15),1-Z.TEST(	
9	D2:D31,60,15))	

Since we chose a significance level of 0.05 for this one-tailed test and the p-value of 0.0069 is less than this significance level, we have established significant results. These results are not due to chance.

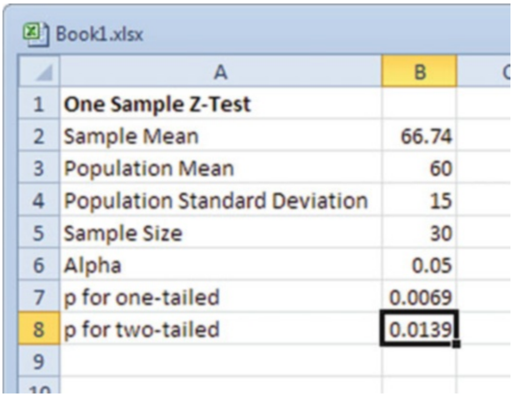
Z.TEST is designed to give a one-tailed result, but it can be adapted to give a two-tailed result. For this example the question might be whether we expect the sample mean of the proficiency scores to be either larger or smaller than 66.74 points.

The formula for the two-tailed z-test is:

$$= \mathbf{2} * \mathbf{MIN}\left(\begin{array}{l}\mathbf{Z.TEST}(\mathbf{array\ ofsample\ data}, \mathbf{population\ mean}, \mathbf{population\ SD}), \\ \mathbf{1} - \mathbf{Z.TEST}(\mathbf{array}, \mathbf{population\ mean}, \mathbf{population\ SD})\end{array}\right)$$

➢ Click on cell B8. Input into the formula ribbon

$$= \mathbf{2} * \mathbf{MIN}(\mathbf{Z.TEST}(\mathbf{D2} : \mathbf{D31}, \mathbf{60}, \mathbf{15}), \mathbf{1} - \mathbf{Z.TEST}(\mathbf{D2} : \mathbf{D31}, \mathbf{60}, \mathbf{15}))$$



Book1.xlsx

	A	B
1	**One Sample Z-Test**	
2	Sample Mean	66.74
3	Population Mean	60
4	Population Standard Deviation	15
5	Sample Size	30
6	Alpha	0.05
7	p for one-tailed	0.0069
8	p for two-tailed	0.0139
9		

The p-value returned by **Z.TEST** for the two-sided test in cell B8 (0.0139) confirms that if the true mean of the underlying normal distribution from which all 30 engineering scores were drawn is in fact 60 points, and the true standard deviation is 15, a sample mean that is higher or lower than 66.74 would occur with probability 0.0139.

Since we choose a significance level of 0.05 for this two-tailed test, and the p-value of 0.0139 is less than this significance level, we have established significant results. These results are not due to chance. We are 95 % confident that the average engineering proficiency quotient score is 60.

Common Excel Pitfalls

- z-test works best with $n > 30$
- Make sure you have an equal amount of "(" as you do ")". If you do not have enough closed parentheses, Excel will give you an error.
- When doing the z-test make sure that you are using the correct formulas and templates for your problem. Using the wrong formula or template will produce incorrect answers.
- By default answers and calculations will display as many decimals as it can to fill a cell. To reduce the number of digits displayed after the decimal, format the cell to display no more than four decimal places.

Final Thoughts and Activities

Practice Problems and Case Studies

1. A teacher wanted to know how well the gifted students in her class perform relative to her other classes. She administers a standardized test with a mean of 50 points and standard deviation of 10 points. Her class of 31 students has an average score of 55. Test if this new sample mean has statistical significance at the 95 % confidence level.
2. A rental car company claims the mean time to rent a car on their website is 60 s with a standard deviation of 30 s. A random sample of 36 customers attempted to rent a car on the website. The mean time to rent was 75 s. Is this enough evidence to contradict the company's claim?
3. Suppose that in a particular company, the overall mean and standard deviation of scores on a safety awareness test are 100 points, and 12 points, respectively. Our interest is in the scores of 55 employees in a new office that just opened in Plano, Texas. They achieved a mean score of 96. Is their mean score significantly lower than the overall company mean: that is, would the scores of the employees in this office be comparable to another simple random sample of 55 employees from the company, or are their scores surprisingly low? In other words, if we took another sample would we get similar results?
4. Many companies are finding that customers are using various types of online content before purchasing products. The following results are the percentages of adults and youths who use various sources of online content. Suppose the survey is based on 100 adults and 100 youths.

	Use online content	
Type of online content	Adult	Youth
Customer product ratings/reviews	74	84
For sale listings with seller ratings	72	80
For sale listings without seller ratings	61	68
Online classified ads	60	69
Message-board posts	60	74
Web blogs	58	70
Dating side profiles/personals	52	62
Peer-generated and peer-reference information	52	71
Peer-posted event listings	49	74

For "Customer product ratings/reviews" and "For sale listings with seller ratings" determine whether there is a difference between adults and youths in the proportion who use the type of online content at the 0.05 level of confidence. Use the template provided in **Chi-SquareTest** file.

5. A company is considering an organizational change by adopting the use of self-managed work teams. To assess the attitudes of employees of the company toward this change, a sample of 400 employees is selected and asked whether they favor the institution of self-managed work teams in the organization. Three responses are permitted: favor, neutral, or oppose. The results of the survey, cross-classified by type of job and attitude toward self-managed work teams are summarized as follow:

	Self-managed work teams			
Type of job	Favor	Neutral	Oppose	Total
Hourly worker	110	49	70	229
Supervisor	20	15	29	64
Middle management	37	17	25	79
Upper management	26	10	8	44
Total	193	91	132	416

At the 0.05 level of significance, is there evidence of a relationship between attitude toward self-managed work and type of job? Use the template provided in **Chi-SquareTest** file.

6. We want to know whether male or female nurses at a local hospital are more likely to gain promotions within the first 2 years of employment within a confidence level of 95 % ($\alpha = .05$). In other words are the promotions independent of gender? Is there a statistically significant relationship between gender and promotions (H_a) or are these differences simply due to chance (H_o)? The data are provided bellow.

	Promotion	No promotion	Total
Females	46	71	117
Males	37	83	120
Total	83	154	237

Discussion Boards

1. If you want to inspire confidence, give plenty of statistics. It does not matter that they should be accurate, or even intelligible, as long as there is enough of them.
2. Economic data is not strong enough to bear the weight of elaborate mathematics and statistics.
3. The formal and mathematical theory of statistics was largely invented in the 1880s by eugenicists. Eugenics lost favor because of misunderstood statistics.
4. No one really knows how many people are malnourished.

Group Activity

1. What's the weak global economy to do when one of its strongest players, China, is growing at its slowest pace in 3 years? What organizations are responsible for reporting global economic statistics? How do they collect this data and how do they define economic growth?
2. The Bureau of Labor Statistics issues the "U.S. Employment Report" monthly. However, it focuses on the gain or decline from the previous month. This does not take into account seasonal variations, such as teachers who go on unpaid leave in the summer. For this reason, a better way to look at employment statistics is year-over-year. Compare and contrast the two methods.
3. Explore the World Hunger Organization's most recent report (www.worldhunger.org). Discuss the statistical pros and cons of this report.

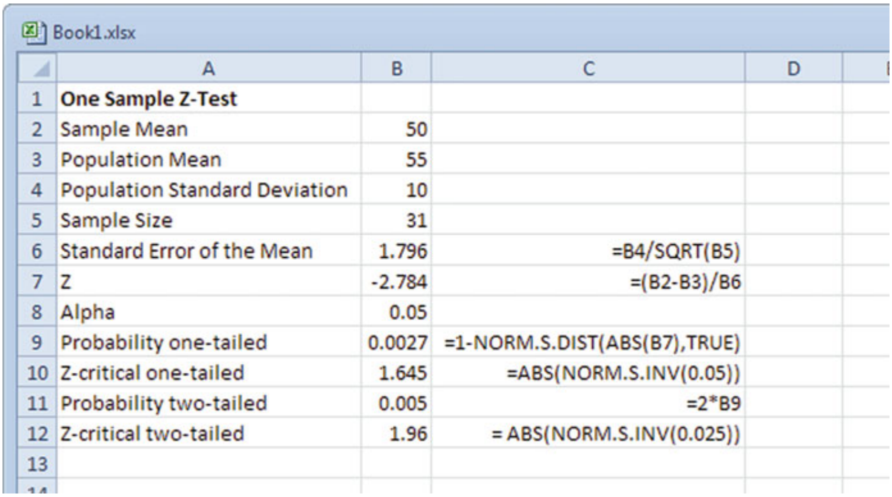
Parting Thought

An unsophisticated forecaster uses statistics as a drunken man uses lamp posts . . . for support rather than for illumination.

Problem Solutions

1. A teacher wanted to know how well the gifted students in her class perform relative to her other classes. She administers a standardized test with a mean of 55 points and standard deviation of 10 points. Her class of 31 students has an

average score of 50. Test if this new sample mean has statistical significance at the 95 % confidence level.



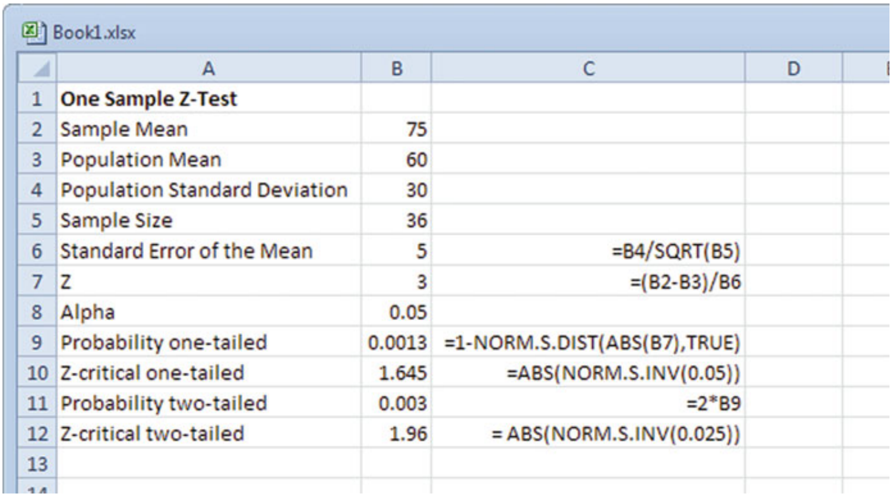
	A	B	C	D
1	**One Sample Z-Test**			
2	Sample Mean	50		
3	Population Mean	55		
4	Population Standard Deviation	10		
5	Sample Size	31		
6	Standard Error of the Mean	1.796	=B4/SQRT(B5)	
7	Z	-2.784	=(B2-B3)/B6	
8	Alpha	0.05		
9	Probability one-tailed	0.0027	=1-NORM.S.DIST(ABS(B7),TRUE)	
10	Z-critical one-tailed	1.645	=ABS(NORM.S.INV(0.05))	
11	Probability two-tailed	0.005	=2*B9	
12	Z-critical two-tailed	1.96	= ABS(NORM.S.INV(0.025))	
13				

Testing for significance we use the two-tailed results since we are only interested if the means are different, not if one is larger or smaller than the other.

- The p-value of 0.005(B11) is smaller than the alpha value of 0.05 (B8). Therefore these results are statistically significant.
- The z-critical value of 1.96 (B12) is less than the z-calculated ABS (2.78) (B7) which also reinforces that these results are statistically significant. Remember when doing this test you need to work with absolute values.

The teacher can use this statistically significant data to show that the gifted students in her class perform differently from her other classes. The difference in the standardized test means is real at the 95 % confidence level. If you wanted to know if they performed better or worse you would need to redefine the problem that way and use the one-tailed test.

2. A rental car company claims the mean time to rent a car on their website is 60 s with a standard deviation of 30 s. A random sample 36 customers attempted to rent a car on the website. The mean time to rent was 75 s. Is this enough evidence to contradict the company's claim with 95 % confidence?



	A	B	C	D
1	**One Sample Z-Test**			
2	Sample Mean	75		
3	Population Mean	60		
4	Population Standard Deviation	30		
5	Sample Size	36		
6	Standard Error of the Mean	5	=B4/SQRT(B5)	
7	Z	3	=(B2-B3)/B6	
8	Alpha	0.05		
9	Probability one-tailed	0.0013	=1-NORM.S.DIST(ABS(B7),TRUE)	
10	Z-critical one-tailed	1.645	=ABS(NORM.S.INV(0.05))	
11	Probability two-tailed	0.003	=2*B9	
12	Z-critical two-tailed	1.96	= ABS(NORM.S.INV(0.025))	
13				

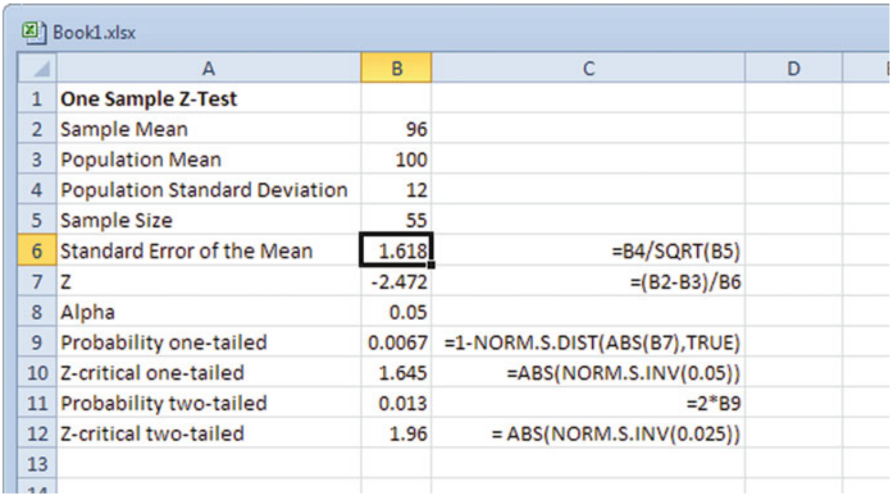
Testing for significance we use the two-tailed results since we are only interested if the means are different, not if one is larger or smaller than the other.

- The p-value of 0.003 (B11) is smaller than the alpha value of 0.05 (B8). Therefore these results are statistically significant.
- The z-critical value of 1.96 (B12) is less than the z-calculated 3.00 (B7) which also reinforces that these results are statistically significant. Remember when doing this test you need to work with absolute values.

We can use these statistically significant results as evidence to contradict the company's claim with 95 % confidence. The difference in the mean time to rent is real at the 95 % confidence level. Best to change your claim of 60 s to avoid a lawsuit.

3. Suppose that in a particular company, the overall mean and standard deviation of scores on a safety awareness test are 100 points, and 12 points, respectively. Our interest is in the scores of 55 employees in a new office that just opened in Plano, Texas. They achieved a mean score of 96. Is their mean score significantly lower than the overall company mean – that is, would the scores of the employees in this office be comparable to another simple random sample of 55 employees from the company, or are their scores surprisingly low? In other words, if we took another sample would we get similar results?

Answer:
Since the company is only interested in the sample mean score being unusually low, we are only interested in one tail.



	A	B	C	D
1	**One Sample Z-Test**			
2	Sample Mean	96		
3	Population Mean	100		
4	Population Standard Deviation	12		
5	Sample Size	55		
6	Standard Error of the Mean	1.618	=B4/SQRT(B5)	
7	Z	-2.472	=(B2-B3)/B6	
8	Alpha	0.05		
9	Probability one-tailed	0.0067	=1-NORM.S.DIST(ABS(B7),TRUE)	
10	Z-critical one-tailed	1.645	=ABS(NORM.S.INV(0.05))	
11	Probability two-tailed	0.013	=2*B9	
12	Z-critical two-tailed	1.96	= ABS(NORM.S.INV(0.025))	
13				

- The one-sided p-value of 0.007 (B9) tells us that the z value is statistically significant. Since the p-value for the one-tailed test is less than the chosen alpha value of 0.05 (B8), the results of the Plano office are real at the 0.05 level. The z-test tells us that the 55 employees of interest have an unusually low mean test score. We could also say that there is a 95 % confidence that any other random sample from this same population would generate similar results. These results are not just due to chance.
- We can also consider the z-critical values. In this case the z-critical value of 1.645 (B10) is less than the ABS z-calculated of 2.47 (B7); since we are working with absolute values, we drop the negative sign which also reinforces that these results are statistically significant.

Their mean score is statistically significantly lower than the overall company mean: that is, the scores of the employees in this office are not comparable to another simple random sample of 55 employees from the company; their scores are surprisingly low.

4. Many companies are finding that customers are using various types of online content before purchasing products. The following results are the percentages of adults and youths who use various sources of online content. Suppose the survey is based on 100 adults and 100 youths.

	Use online content	
Type of online content	Adult	Youth
Customer product ratings/reviews	74	84
For sale listings with seller ratings	72	80
For sail listings without seller ratings	61	68
Online classified ads	60	69
Message-board posts	60	74
Web blogs	58	70
Dating side profiles/personals	52	62
Peer-generated and peer-reference information	52	71
Peer-posted event listings	49	74

For “Customer product ratings/reviews” and “For sale listings with seller ratings” determine whether there is a difference between adults and youths in the proportion that use the type of online content at the 0.05 level of confidence.

Answer:

Adult		Youth	
Mean	59.77778	Mean	72.44444
Standard error	2.871207	Standard error	2.186519
Median	60	Median	71
Mode	60	Mode	74
Standard deviation	8.61362	Standard deviation	6.559556
Sample variance	74.19444	Sample variance	43.02778
Kurtosis	−0.4815	Kurtosis	0.285721
Skewness	0.63015	Skewness	0.396455
Range	25	Range	22
Minimum	49	Minimum	62
Maximum	74	Maximum	84
Sum	538	Sum	652
Count	9	Count	9

z-test: two sample for means		
Customer product ratings/reviews		
	Variable 1	Variable 2
Mean	74	84
Known variance	74.19444	43.02778
Observations	1	1
Hypothesized mean difference	0	
z	−0.92362	
P(Z <= z) one-tail	0.177841	
z critical one-tail	1.644854	
P(Z <= z) two-tail	0.355682	
z critical two-tail	1.959964	

z-test: two sample for means		
For sale listings with seller ratings		
	Variable 1	Variable 2
Mean	72	80
Known variance	74.19444	43.02778
Observations	1	1
Hypothesized mean difference	0	
z	−0.7389	
P(Z <= z) one-tail	0.229984	
z critical one-tail	1.644854	
P(Z <= z) two-tail	0.459968	
z critical two-tail	1.959964	

Based upon the online content that they used, there is no difference between what the youth and the adults used when looking at customer product ratings/ reviews or sales listings with seller ratings.

5. A company is considering an organizational change by adopting the use of self-managed work teams. To assess the attitudes of employees of the company toward this change, a sample of 400 employees is selected and asked whether they favor the institution of self-managed work teams in the organization. Three responses are permitted: favor, neutral, or oppose. The results of the survey, cross-classified by type of job and attitude toward self-managed work teams are summarized as follow:

	Self-managed work teams			
Type of job	Favor	Neutral	Oppose	Total
Hourly worker	110	49	70	229
Supervisor	20	15	29	64
Middle management	37	17	25	79
Upper management	26	10	8	44
Total	193	91	132	416

At the 0.05 level of significance, is there evidence of a relationship between attitude toward self-managed work and type of job? Use the template provided in **Chi-Square Test** file.

Chi-Square Test

	Observed frequencies			
	Column variable			
Row variable	Favor	Neutral	Oppose	**Total**
Hourly worker	110	49	70	**229**
Supervisor	20	15	29	**64**
Middle management	37	17	25	**79**
Upper management	26	10	8	**44**
Total	**193**	**91**	**132**	**416**

	Expected frequencies			
	Column variable			
Row variable	Favor	Neutral	Oppose	Total
Hourly worker	106.2428	50.09375	72.66346	229
Supervisor	29.69231	14	20.30769	64
Middle management	36.65144	17.28125	25.06731	79
Upper management	20.41346	9.625	13.96154	44
Total	193	91	132	416

Calculations

(Freq observed) − (Freq expected)		
3.757212	−1.09375	−2.66346
−9.69231	1	8.692308
0.348558	−0.28125	−0.06731
5.586538	0.375	−5.96154

((Freq observed) − (Freq expected))^2/(Freq expected)		
0.132871	0.023881	0.097629
3.16381	0.071429	3.720571
0.003315	0.004577	0.000181
1.528864	0.01461	2.545561

Data	
Level of significance	0.05
Number of rows	4
Number of columns	3
Degrees of freedom	1

Reject the null hypothesis, which means that we found a significant difference between attitude toward self-managed work and an employee’s type of job classification.

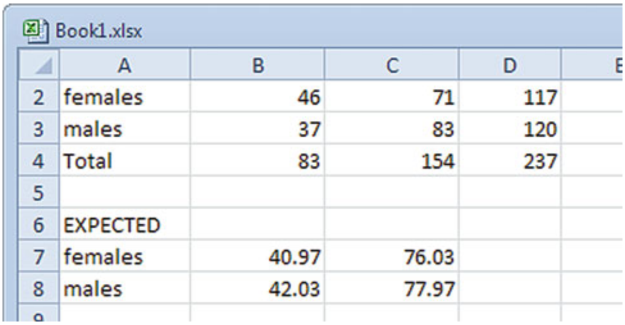
Results	
Critical value	3.841459
Chi-square test statistic	11.89531
p-value	0.000563
Reject the null hypothesis	

6. We want to know whether male or female nurses at a local hospital are more likely to gain promotions within the first 2 years of employment within a confidence level of 95 % ($\alpha = .05$). In other words are the promotions independent of gender? Is there a statistically significant relationship between gender and promotions (H_a) or are these differences simply due to chance (H_o)? The data are provided below.

	Promotion	No promotion	Total
Females	46	71	117
Males	37	83	120
Total	83	154	237

The data shows that there are 237 nurses at this hospital, 117 of whom are female while 120 are male. At first review of the data it appears that male nurses have received more promotions (37) than female nurses (46). The null hypothesis is that the two variables are independent, that the likelihood of getting promoted is the same for males and females.

Table of observed and expected values



Book1.xlsx

	A	B	C	D	E
2	females	46	71	117	
3	males	37	83	120	
4	Total	83	154	237	
5					
6	EXPECTED				
7	females	40.97	76.03		
8	males	42.03	77.97		
9					

As with all X^2 testing we need the observed and expected values. Expected values are calculated based on the row and column totals from the table.

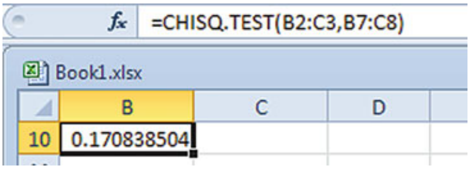
With two variables the expected value for each cell of the table can be calculated using the following formula:

$$(\text{Row Total} \times \text{Column Total})/\text{Total N count for overall table}$$

There is no Excel function that will do these calculations automatically but you can input them as formula for each cell.

By using CHISQ.TEST in Excel we generate a p-value of 0.171 (B10), but we must refer to the number of degrees of freedom to be totally correct. There is no Excel function that will automatically calculate this so you can input the formula

$$\text{df} = (\text{the number of columns} - 1) \times (\text{number of rows} - 1)$$



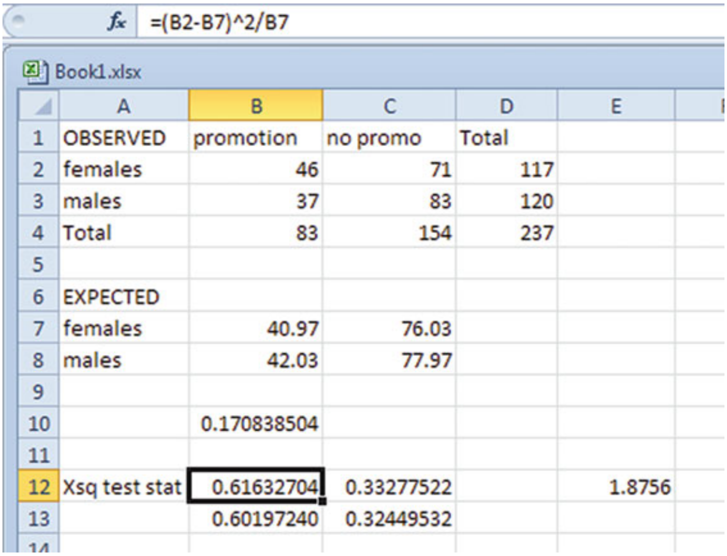
=CHISQ.TEST(B2:C3,B7:C8)

	B	C	D
10	0.170838504		

In this table, there were two rows and two columns. Therefore, the number of degrees of freedom is one. We can now state that the p-value was calculated with one degree of freedom as 0.17 (B10); this value exceeds our $p_{critical}$ of 0.05. Our results are not statistically significant; we cannot reject the null hypothesis. The relationship between gender and promotions is not significant, and so we can assume is these results are simply due to chance.

For reporting we may also want to report the actual X^2 test statistic. Unfortunately Excel does not provide this to the user in any simple way. Once again you need to input a formula:

$$X^2 = \Sigma(\text{observed value}_i - \text{expected value}_i)^2/\text{expected value}_i$$



=(B2-B7)^2/B7

	A	B	C	D	E
1	OBSERVED	promotion	no promo	Total	
2	females	46	71	117	
3	males	37	83	120	
4	Total	83	154	237	
5					
6	EXPECTED				
7	females	40.97	76.03		
8	males	42.03	77.97		
9					
10		0.170838504			
11					
12	Xsq test stat	0.61632704	0.33277522		1.8756
13		0.60197240	0.32449532		

To formally report our results, we would say "no statistically significant relationship was found between gender and promotions, $X^2(1) = 1.88$, $p = 0.17$".

Chapter 13
Multiple Regression

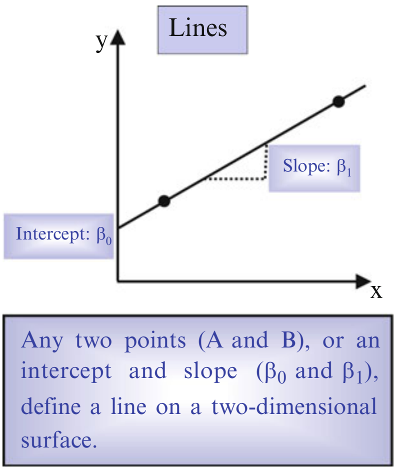
Key Concepts

Adjusted coefficient of determination, Law of parsimony, Multicollinearity, and Multiple regression.

Discussion

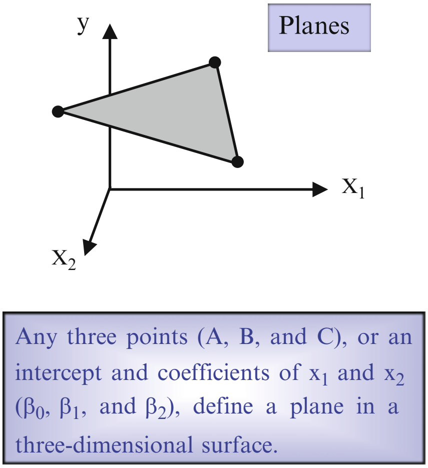
The chapter discusses **multiple regression** by building on the basic concepts of simple linear regression. Now instead of only one independent variable, we have several independent variables to predict the dependent variable. To visualize the difference between simple and multiple regression see Figs. 13.1 and 13.2. In simple linear regression, the data are fit to a straight line. In multiple regression, the data are fit to a plane.

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_13,



Any two points (A and B), or an intercept and slope (β_0 and β_1), define a line on a two-dimensional surface.

Fig. 13.1 Simple linear regression

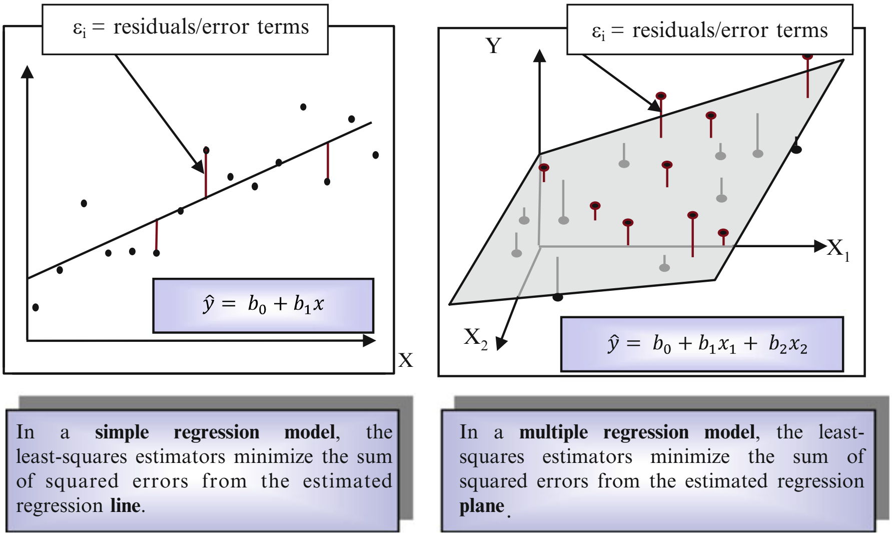


Any three points (A, B, and C), or an intercept and coefficients of x_1 and x_2 (β_0, β_1, and β_2), define a plane in a three-dimensional surface.

Fig. 13.2 Multiple regression

Figures 13.1 and 13.2 display the basic models. Excel cannot draw a multi-dimensional graph; if you actually need to see a multiple regression graph, more sophisticated software is required.

Figure 13.3 compares how we fit a simple regression model and a multiple regression model by minimizing the residuals or the error terms. The simple linear regression data is fit as a line while the multiple regression data is fit to a plane.


ε_i = residuals/error terms
$\hat{y} = b_0 + b_1x$
X
ε_i = residuals/error terms
Y
X_1
X_2
$\hat{y} = b_0 + b_1x_1 + b_2x_2$

In a **simple regression model**, the least-squares estimators minimize the sum of squared errors from the estimated regression **line**.

In a **multiple regression model**, the least-squares estimators minimize the sum of squared errors from the estimated regression **plane**.

Fig. 13.3 Minimizing the residuals

The **multiple regression** equation combines a calculated amount of each of the x variables with a y intercept. The **regression coefficients** indicate how much of each x variable to include in your model.

Consider the multiple regression equation $\mathbf{y = 2x_1 + 5.2x_2 - .78x_3 + 22}$
In this case, the regression coefficients are 2, 5.2, and −.78. There are three independent variables (x_1, x_2, x_3) and the y-intercept is indicated as the "b_0" term (22, in this case).

Adjusted Coefficient of Determination Remember, in simple linear regression, the coefficient of determination describes how much of the variation in y can be explained by one x. The **adjusted coefficient of determination** describes how much of the variation in y can be explained by all of the x's. The term "adjusted" is included to reflect the fact there is more than one x in the model, and we want to statistically "adjust out" or remove the effect of simply adding more independent variables to the model when Excel calculates the coefficient of determination.

Law of Parsimony This law states that the optimum model is one with the fewest number of independent variables. The more variables you have the more difficult it becomes to collect good data for all of the variables each time. You can gain insight on which independent variable(s) should be deleted by completing correlation tests and significance tests.

The first step is to run a correlation matrix to identify which of the x variables have high correlations with the other x variables. When two or more of the independent variables have a strong coefficient of correlation we refer to this as **multicollinearity**. In other words, multicollinear variables are considered redundant and should be eliminated. High x to x correlations suggest the x's are contributing similar effects to the model and you probably only need one of them.

The second step is to consider the associated significance tests which can indicate which of the independent variables are not worth keeping in the model as robust predictors of the dependent variable.

Excel

In this first example we will test if sales can be predicted by knowing how much money has been spent on radio advertising and newspaper ads. Because there are two x's we need to run a multiple regression model rather a simple linear regression model.

Step 1: Fit the Model with Selected Independent Variables

$$\mathrm{y} = \mathrm{x}_1(\text{radio advertising \$}) + \mathrm{x}_2(\text{newspaper advertising \$}) + \mathrm{b}$$

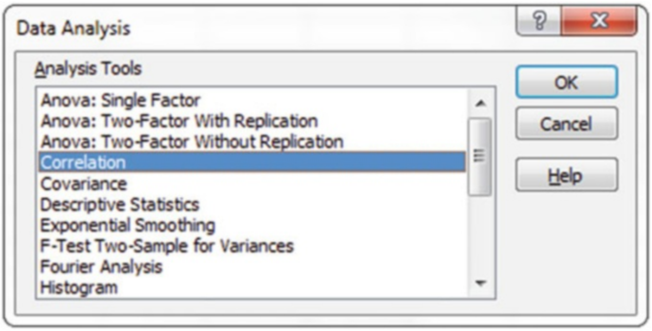
Where y $\rightarrow$ sales, $\mathrm{x}_1 \rightarrow$ radio ads, $\mathrm{x}_2 \rightarrow$ newspaper ads

Step 2: Does Multicollinearity Exist? Run a Correlation Matrix

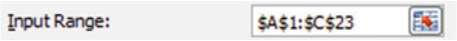
- ➢ Input data from Fig. 13.4
- ➢ Click on **Data** tab
- ➢ Select **Data Analysis** function
- ➢ Select **Correlation** from the list of Analysis Tools

Sales	Radio	Newspaper
973	0	40
1119	0	40
875	25	25
625	25	25
910	30	30
971	30	30
931	35	35
1177	35	35
882	40	25
982	40	25
1628	45	45
1577	45	45
1044	50	0
914	50	0
1329	55	25
1330	55	25
1405	60	30
1436	60	30
1521	65	35
1741	65	35
1866	70	40
1717	70	40

Fig. 13.4 Input data



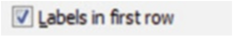
- ➢ Click **OK**
- ➢ Click inside the **Input Range** text box and then highlight all columns for the x variables and the y variable.



Input Range: $A$1:$C$23

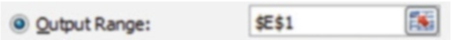
- ◈ Remember to include labels so you can keep track of which variables correlate to each other.

- ➢ Check the **Labels in first row** box



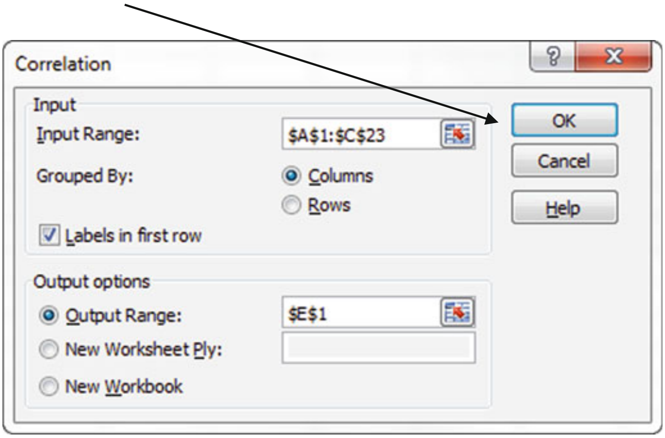
☑ Labels in first row

- ➢ Select **Output Range**, then click inside the text box to input the output range



● Output Range: $E$1

- ➢ Click **OK**



Correlation

Input

Input Range: $A$1:$C$23

Grouped By: ● Columns ○ Rows

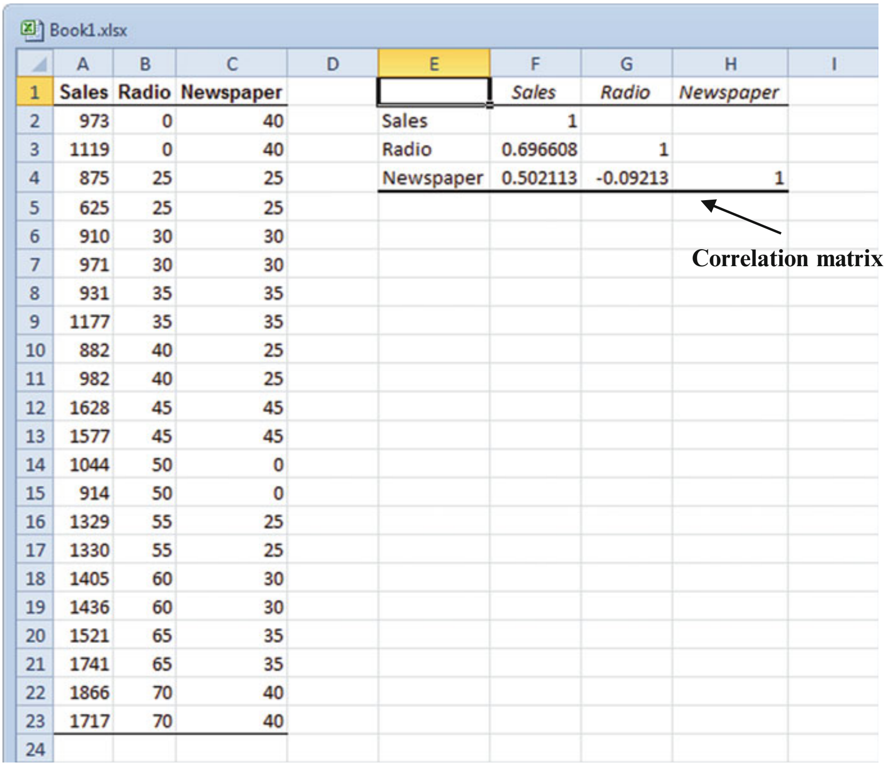
☑ Labels in first row

Output options

● Output Range: $E$1

○ New Worksheet Ply:

○ New Workbook



	A	B	C	D	E	F	G	H	I
1	Sales	Radio	Newspaper			*Sales*	*Radio*	*Newspaper*	
2	973	0	40		Sales	1			
3	1119	0	40		Radio	0.696608	1		
4	875	25	25		Newspaper	0.502113	-0.09213	1	
5	625	25	25						
6	910	30	30						
7	971	30	30						
8	931	35	35						
9	1177	35	35						
10	882	40	25						
11	982	40	25						
12	1628	45	45						
13	1577	45	45						
14	1044	50	0						
15	914	50	0						
16	1329	55	25						
17	1330	55	25						
18	1405	60	30						
19	1436	60	30						
20	1521	65	35						
21	1741	65	35						
22	1866	70	40						
23	1717	70	40						
24									

Correlation matrix

The rule of thumb suggests that r values (correlation coefficients) greater than +0.6 or less than −0.6 between **x variables** should be flagged as possible multicollinear variables. In this example, the relationship between **Radio** (x_1) and **Newspaper** (x_2) is only −0.09 (G4), which is within the acceptable level. Therefore, no multicollinearity is noted so no need to delete any of the x variables in this model.

But we also would consider how strongly each x variable correlates with y (sales). Radio is stronger with an r of .7, while newspaper only achieves an r value of .5. We need to flag the newspaper variable for further consideration before deleting from the model.

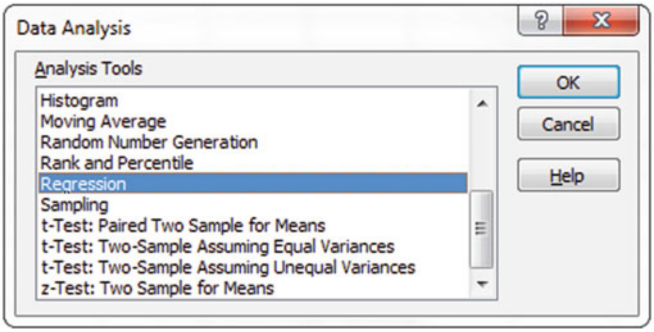
Step 3: Run Regression Model

◇ **Note:** Columns of data much be adjacent to one another.

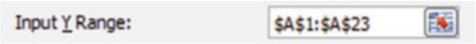
➢ Click on **Data** tab

➢ Select **Data Analysis** function

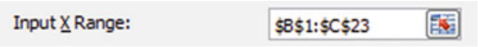
➢ Select **Regression** from the list of Analysis Tools


Data Analysis
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

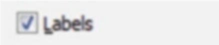
- **Input Y Range:** you need to highlight the column of y data


Input Y Range: $A$1:$A$23

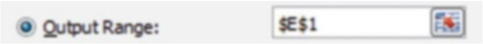
- **Input X Range:** you need to highlight **<u>all</u>** columns of **<u>all x variables</u>**


Input X Range: $B$1:$C$23

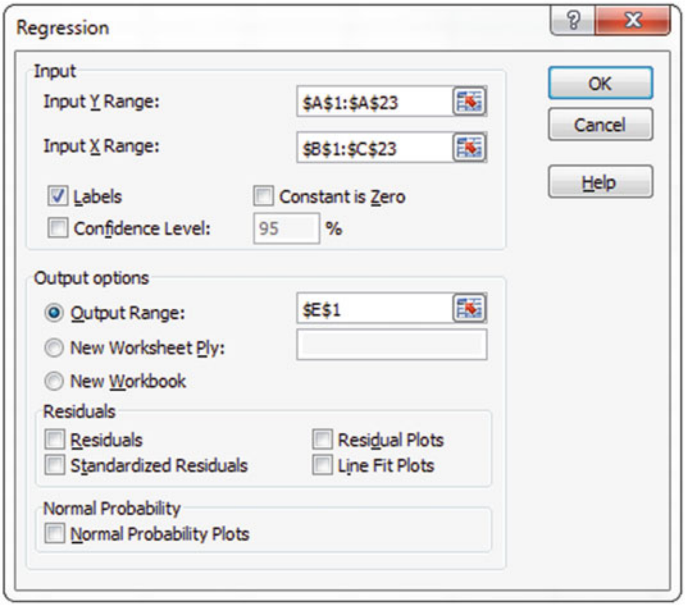
- Check the **Labels** box, if the first row of your data includes labels


☑ Labels

- Open a new work page by clicking on **Sheet 2** at the bottom of the page.
- Select the **Output Range** option and click on the cell where you want the output displayed


● Output Range: $E$1

➢ Click **OK**



Regression

Input

Input Y Range: $A$1:$A$23

Input X Range: $B$1:$C$23

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: $E$1

○ New Worksheet Ply:

○ New Workbook

Residuals

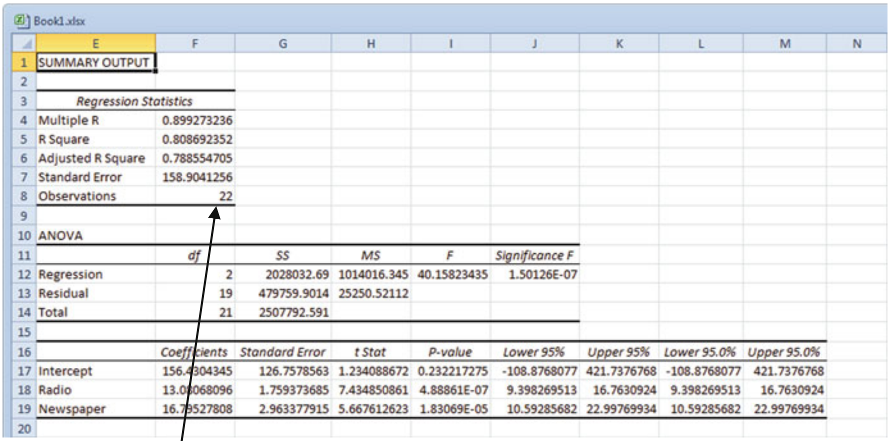
☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

Results:



Book1.xlsx

	E	F	G	H	I	J	K	L	M	N
1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple R	0.899273236								
5	R Square	0.808692352								
6	Adjusted R Square	0.788554705								
7	Standard Error	158.9041256								
8	Observations	22								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*Significance F*				
12	Regression	2	2028032.69	1014016.345	40.15823435	1.50126E-07				
13	Residual	19	479759.9014	25250.52112						
14	Total	21	2507792.591							
15										
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	
17	Intercept	156.4304345	126.7578563	1.234088672	0.232217275	-108.8768077	421.7376768	-108.8768077	421.7376768	
18	Radio	13.08068096	1.759373685	7.434850861	4.88861E-07	9.398269513	16.7630924	9.398269513	16.7630924	
19	Newspaper	16.79527808	2.963377915	5.667612623	1.83069E-05	10.59285682	22.99769934	10.59285682	22.99769934	
20										

◈ Remember to check that all your data have been included by looking at the **Observation**s count.

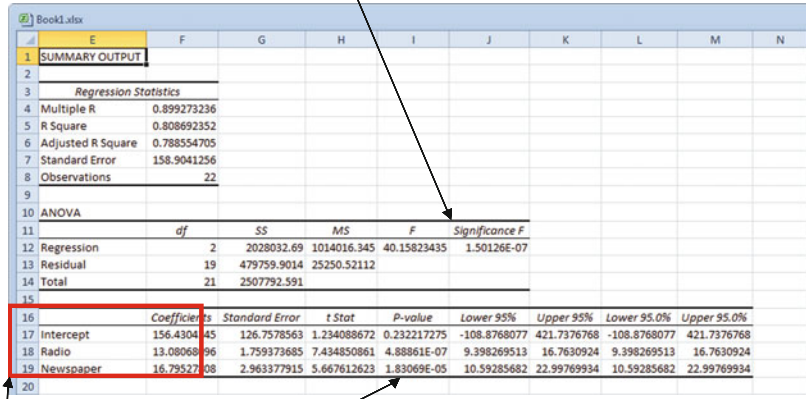
Step 4: Are the Assumptions of Regression Satisfied?

Check the model in Fig. 13.5 to see if the assumptions are satisfied.

Step 5: Test Overall Model Significance (F-Test)

The significance test results are included in the **Regression Output.** Remember to check the **Significance F** for the overall model significance is less than 0.05 (for 95 % confidence level etc.). Also, check the **p-values** are less than 0.05 (for 95 % confidence level etc.) for each of the independent variables as discussed in the previous chapters on Significance Testing.

In this example, the **Significance F** is very small, much less than the threshold of 0.05, so this is good. If it is greater than or equal to 0.05, you need to go back and review the x variables you have included in the model. Make sure you included all columns of x data when rerunning the model.



Book1.xlsx

	E	F	G	H	I	J	K	L	M	N
1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple R	0.899273236								
5	R Square	0.808692352								
6	Adjusted R Square	0.788554705								
7	Standard Error	158.9041256								
8	Observations	22								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*Significance F*				
12	Regression	2	2028032.69	1014016.345	40.15823435	1.50126E-07				
13	Residual	19	479759.9014	25250.52112						
14	Total	21	2507792.591							
15										
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	
17	Intercept	156.4304[illegible]45	126.7578563	1.234088672	0.232217275	-108.8768077	421.7376768	-108.8768077	421.7376768	
18	Radio	13.08068[illegible]96	1.759373685	7.434850861	4.88861E-07	9.398269513	16.7630924	9.398269513	16.7630924	
19	Newspaper	16.79527[illegible]08	2.963377915	5.667612623	1.83069E-05	10.59285682	22.99769934	10.59285682	22.99769934	
20										

Step 6: Check p-Values for Independent Variables Meet Significance Criteria (t-Test)

Identify the weak x's by considering the p-values greater than 0.05. Sometimes the **p-values** can be greater than 0.05, because there is a low sample size for the variable relative to the other variable sample sizes. If you know the variable is important to the model, you may decide to include it despite the fact that it does not pass the significance test. It is important to remove one x at a time when cleaning up the model. Rerun your model if you need to remove non-significant x variables. In this example the **p-values** are all smaller than 0.05, so there is no need to delete any x data at this point. No need to rerun the model.

- No need to look at the p value for the y-intercept

Step 7: Run Model for Prediction and Estimation

The resulting equation can now be used to predict new values of y by inputting data for each or the significant x variables. Consider the resulting model:

$$y = 13.08\ x_1 + 16.80\ x_2 + 156.43$$

This model suggests that about 79 % of sales can be explained by knowing the spend on radio and newspaper advertising. As you gain more and more data it is always recommended to rerun the regression analysis and recalibrate the model.

Figure 13.5 summarizes the step by step road map for developing the best fitting regression model. Remember if there is only one x variable a simple linear regression can be used. Multiple regression models are used when there is more than one x variable.

Run Regression Model

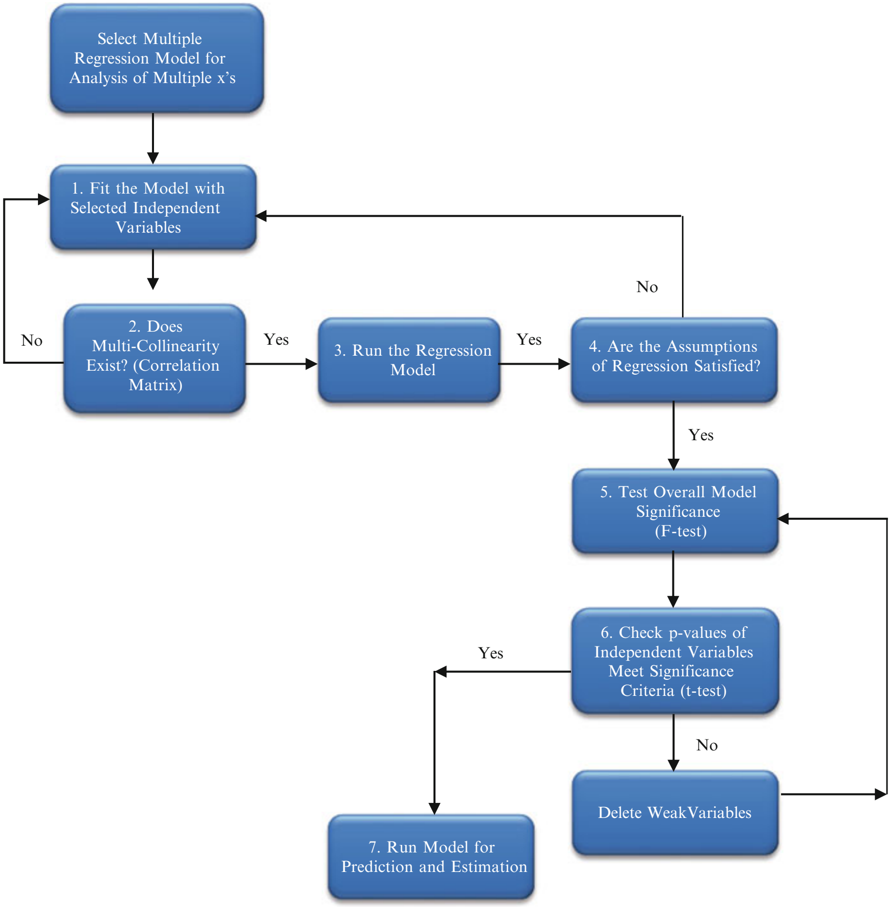

Select Multiple Regression Model for Analysis of Multiple x's
1. Fit the Model with Selected Independent Variables
2. Does Multi-Collinearity Exist? (Correlation Matrix)
No
Yes
3. Run the Regression Model
Yes
4. Are the Assumptions of Regression Satisfied?
No
Yes
5. Test Overall Model Significance (F-test)
6. Check p-values of Independent Variables Meet Significance Criteria (t-test)
Yes
No
Delete WeakVariables
7. Run Model for Prediction and Estimation

Fig. 13.5 Road map for selecting the best fitting regression model

Final Thoughts and Activities

Practice Problems

1. Generate a multiple regression equation for the following data where y is sales (\$000), x_1 is money spent on decorating the store (\$000), and x_2 is area of the store (hundreds of square feet). The owner is trying to determine how much he can expect in sales from a store that is 4,000 sq. ft. and will require \$4,000 worth of decorating. Can you help him out?

Data		
Costs of decorating (\$000) (x_1)	Area (00 ft^2) (x_2)	Sales (\$000) (y)
0.4	19.7	19.7
2.8	19.1	19.3
4.0	18.2	18.6
6.0	5.2	7.9
1.1	4.3	4.4
2.6	9.3	9.6
7.1	3.6	8.0
5.3	14.8	15.7
9.7	11.9	15.4
3.1	9.3	9.8
9.9	2.8	10.3
5.3	9.9	11.2
6.7	15.4	16.8
4.3	2.7	5.1
6.1	10.6	12.2
9.0	16.6	18.9
4.2	11.4	12.2
4.5	18.8	19.3
5.2	15.6	16.5
4.3	17.9	18.4

2. A survey of recently sold single-family houses in a small city is selected. The houses in the city were reassessed 6-months prior to the study. Develop a model to predict the selling price (in thousands of dollars), using the assessed value (in thousands of dollars), as well as the time period when sold (in months since reassessment). Use the results in the file **Housing** for your analysis.

   (a) Is there multicollinearity?

   (b) Determine whether there is a significant relationship between selling price and the two dependent variables (assessed value and time period) at the 0.05 level of confidence.

   (c) Determine the adjusted R^2.

Discussion Boards

1. When a university accepts a new student, it is basically predicting that the student can be successful in his or her studies. What predictor variables do admissions offices use? Some studies suggest SAT scores are not useful. Discuss.
2. Why is "parsimony" important in multiple regression?

Parting Thought

The 50-50-90 rule states that anytime you have a 50-50 chance of getting something right, there is a 90 % probability you will get it wrong.

Problem Solutions

1. Generate a multiple regression line for the following data where y is sales ($000), x_1 is money spent on decorating the store ($000), and x_2 is area of the store (hundreds of square feet). The owner is trying to determine how much he can expect in sales from a store that is 4,000 ft$^{\mathbf{2}}$. and will require $4,000 worth of decorating. Can you help him out?

Costs of decorating ($000) ($x_1$)	Area (00 ft^2) (x_2)	Sales ($000) (y)
0.4	19.7	19.7
2.8	19.1	19.3
4.0	18.2	18.6
6.0	5.2	7.9
1.1	4.3	4.4
2.6	9.3	9.6
7.1	3.6	8.0
5.3	14.8	15.7
9.7	11.9	15.4
3.1	9.3	9.8
9.9	2.8	10.3
5.3	9.9	11.2
6.7	15.4	16.8
4.3	2.7	5.1
6.1	10.6	12.2
9.0	16.6	18.9
4.2	11.4	12.2
4.5	18.8	19.3
5.2	15.6	16.5
4.3	17.9	18.4

Answer:

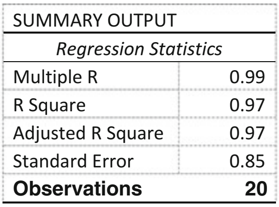
Step 1: Fit the model with selected independent variables
Step 2: Check for multicollinearity

	Costs of Decorating ($000)	Area (00ft^2)	Sales ($000)
Costs of Decorating ($000)	1		
Area (00ft^2)	-0.21838896	1	
Sales ($000)	0.035218554	0.95487407	1

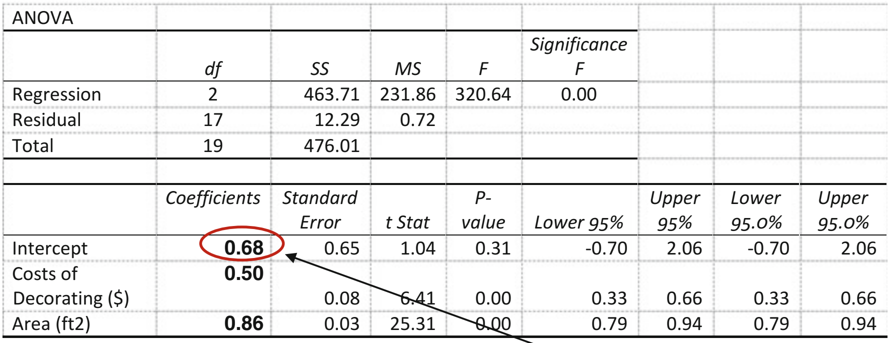
The r value of x_1 and x_2 is less than the threshold value of ±0.6, thus indicating there is no multicollinearity in this model.

Step 3: Run Multiple Regression Model

The **Adjusted R square value** is nice and high at 0.97. This means that 97 % of the change in sales can be explained by the cost of decorating (x_1) and the area of the store (x_2). When we check the **Observations**, we see that all of our input data has been included (20).



SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.99
R Square	0.97
Adjusted R Square	0.97
Standard Error	0.85
Observations	**20**



ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	463.71	231.86	320.64	0.00
Residual	17	12.29	0.72		
Total	19	476.01			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	**0.68**	0.65	1.04	0.31	-0.70	2.06	-0.70	2.06
Costs of Decorating ($)	**0.50**	0.08	6.41	0.00	0.33	0.66	0.33	0.66
Area (ft2)	**0.86**	0.03	25.31	0.00	0.79	0.94	0.79	0.94

The resulting equation for this model is $y = .5x_1 + .86x_2 + .68$

Step 4: Test Assumptions of Regression

Make sure all the assumptions for regression are met**.**

Step 5: Review F-tests

The overall model is significant with a **Significance F** value of 0.00

Step 6: Review Significance t-tests

Both x_1 and x_2 have significant **p-values** of 0.00

Step 7: Run Model for Prediction and Estimation

This multiple regression model seems to be a reasonable model for predicting sales using the cost of decorating and the area of the store as input data. Now plug in $x_1 = \$4{,}000$(input as 4) and $x_2 = 4{,}000\ \text{ft}^2$ (input as 40) to the equation and the equation is ready to use.

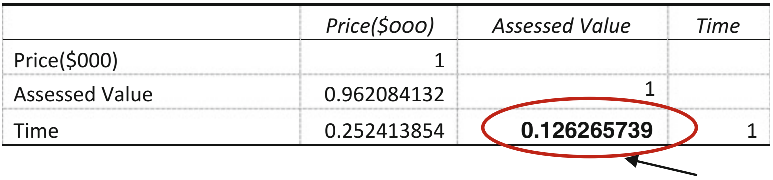
◇ Be very careful with the input units

$$y = .5(4) + .86\ (40) + .68 = \$37080$$

In this example, the value for predicted sales (y) is $37,080.

2. A survey of recently sold single-family houses in a small city is selected. Develop a model to predict the selling price (in thousands of dollars), using the assessed value (in thousands of dollars) as well as time period when sold (in months since reassessment). The houses in the city have been reassessed at full value a year prior to the study. Use the results in the file **Housing** for your analysis.

(a) Is there an issue with multicollinearity?



	Price($ooo)	*Assessed Value*	*Time*
Price($000)	1		
Assessed Value	0.962084132	1	
Time	0.252413854	**0.126265739**	1

There is no multicollinearity between the two x variables (assessed value and time). The r value of x_1 and x_2 is less than the threshold value of ±0.6.

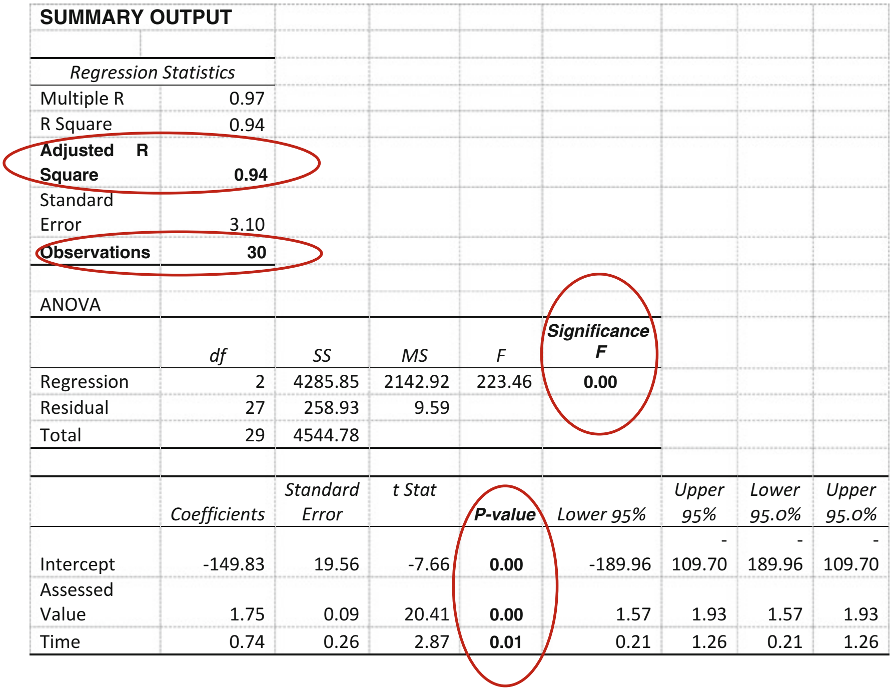
(b) Determine whether there is a significant relationship between selling price and the two dependent variables (assessed value and time period) at the 0.05 level of confidence. Note although number of foreclosures in the area were provided, the model did not request that data be included.

Yes the relationship is significant. The **p-values** of the x variables are 0.00 and 0.01 which are both significant. The overall model is also significant with a **Significance F** of 0.00.

(c) Determine the adjusted R^2.

0.94. This means that 94 % of the variation in y (selling price) can be explained by considering these the two x variables (assessed value and time period).

Output Table.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.97
R Square	0.94
Adjusted R Square	**0.94**
Standard Error	3.10
Observations	**30**

ANOVA

	df	*SS*	*MS*	*F*	***Significance F***
Regression	2	4285.85	2142.92	223.46	**0.00**
Residual	27	258.93	9.59		
Total	29	4544.78			

	Coefficients	*Standard Error*	*t Stat*	***P-value***	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-149.83	19.56	-7.66	**0.00**	-189.96	-109.70	-189.96	-109.70
Assessed Value	1.75	0.09	20.41	**0.00**	1.57	1.93	1.57	1.93
Time	0.74	0.26	2.87	**0.01**	0.21	1.26	0.21	1.26

Chapter 14
Non-linear Regression

Key Concepts

Exponential function, Logarithmic function, Polynomial function, and Power function.

Discussion

Since not all x,y relationships are linear, this chapter will discuss some of the most common non-linear regression tools: power, polynomial, exponential, logarithmic. Excel can provide a graphical display with the associated trendline equation and R^2 for single independent variable models. The best way to get started is to insert various forms of trendlines on a scatterplot and observe the resulting R^2 values. Remember, in all trend fitting, we want the model that best fits the data with the least number of variables or terms.

Power

The formula for this trendline is:

$$y = a * x^b$$

Example $y = 3.2x^{1.67}$

This means that the independent variable x is raised to a power. The value for the power must be positive, and does not have to be a whole number. There is only one regression co-efficient in this case and that is shown in front of the x. As we increase the power, the graph climbs more steeply as shown in Fig. 14.1.

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_14,

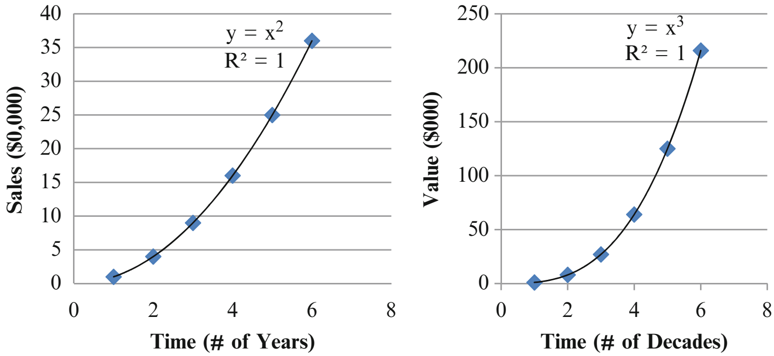


Fig. 14.1 Increasing power levels

Polynomial

The formula for this trendline requires you to state the highest power you want x to be raised to:

For power (order) of 2 $y = a * x^2 + b * x + c$

Example $y = 3.2x^2 + 5x + 10$

For power (order) of 3 $y = d * x^3 + a * x^2 + b * x + c$

Example $y = 6x^3 + 3.2x^2 + 5x + 10$

This means that the independent variable x is raised to several powers. The convention that Excel follows is the power level(s) as the order, so a third order polynomial has x raised to the highest power of 3. A second order polynomial would have x raised to the highest power of 2. The value for the power must always be positive, and will always include values as whole numbers in sequential order from the highest power you select. If you selected the highest power of 4, Excel would provide a trendline equation with x^4, x^3, x^2, and x. The values for x can be positive or negative.

When data seem to fit a simple curve or bulge you might want to try a second order polynomial. But in cases where you have an "s" shape, in other words, a peak and a trough in your data, you should try a third order polynomial. See Fig. 14.2. As you add more powers, the shape of the model is wavier.

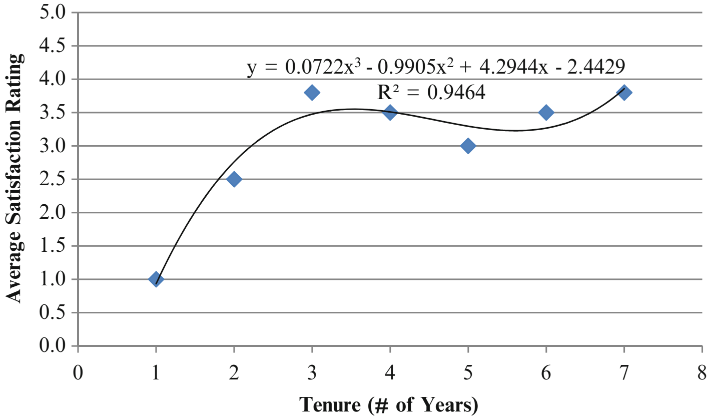


Fig. 14.2 Third order (cubic) polynomial

◇ But remember each time you add a higher power the more complicated your model. Try to get the lowest order model with a reasonable R^2.

Exponential

The formula for this trendline is:

$y = a * e^{bx}$
Example $y = 3.2e^{1.67x}$

This means that the independent variable x is now showing up in the power. The value for "e" is always a numerical constant that is equal to 2.71828. The independent variable x, must have only positive values, and does not have to be a whole number. There are two regression co-efficients in this case, one in front of the e (3.2) and one in front of the x (1.67).

Examples of a negative power and a positive power are provided in Figs. 14.3 and 14.4.

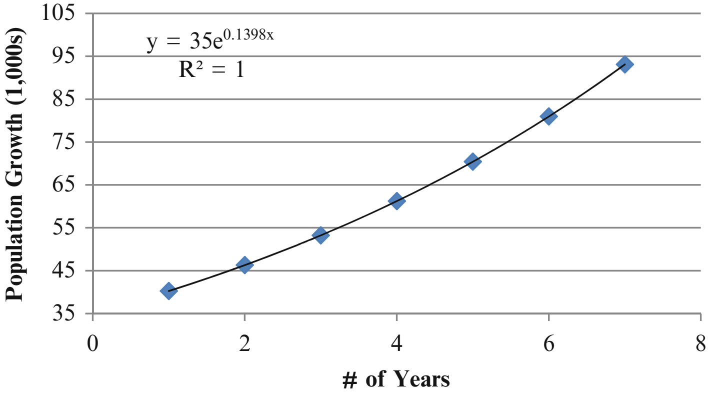


Fig. 14.3 Positive exponential model

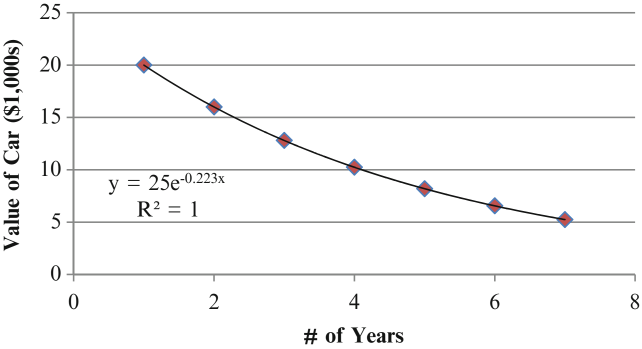


Fig. 14.4 Negative exponential model

Logarithmic

The formula for this trend line is (Fig. 14.5):

$y = a * \ln(x) + b$

Example: $y = 6.7 \ln(1) + 1$

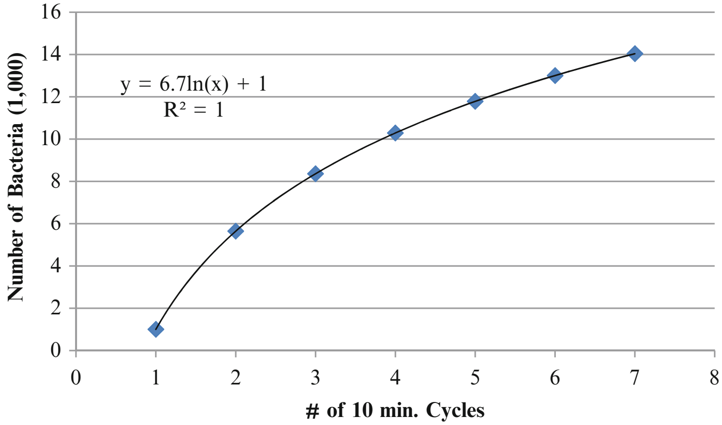


Fig. 14.5 Logarithmic function

The natural logarithm of a number *x,* written as *ln(x)* is the power to which *e* would have to be raised to equal *x*. Here are some important relationships:

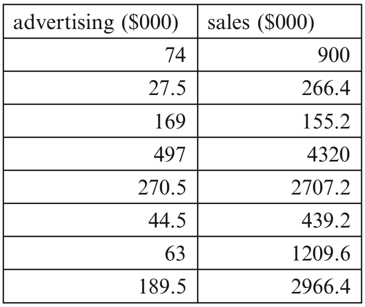
- *ln (7.389 . . .)* is 2, because $e^2 = 7.389\ldots$.
- The natural log of *e*, *ln(e)*, is 1, since $e^1 = e$.
- The natural logarithm of 1, *ln(1)* is 0, since $e^0 = 1$

Although logarithms are a more complicated function type to understand, there are still many places where they are used. These include:

- Working with the data obtained from magnitude scales used to measure earthquakes
- Compounding interest calculations
- Calculating how long it will take a bank deposit at a set interest rate to reach a specified higher amount
- Understanding global warming by calculating alterations in atmospheric CO_2
- Calculating radioactive decay-dating estimations

Excel

The input data set below is used for exploring non-linear regression models in this chapter. We will consider the relationship between the independent variable, *advertising dollars spent*, and the dependent variable, *resulting sales* (Fig. 14.6).



advertising ($000)	sales ($000)
74	900
27.5	266.4
169	155.2
497	4320
270.5	2707.2
44.5	439.2
63	1209.6
189.5	2966.4

Fig. 14.6 Input data set

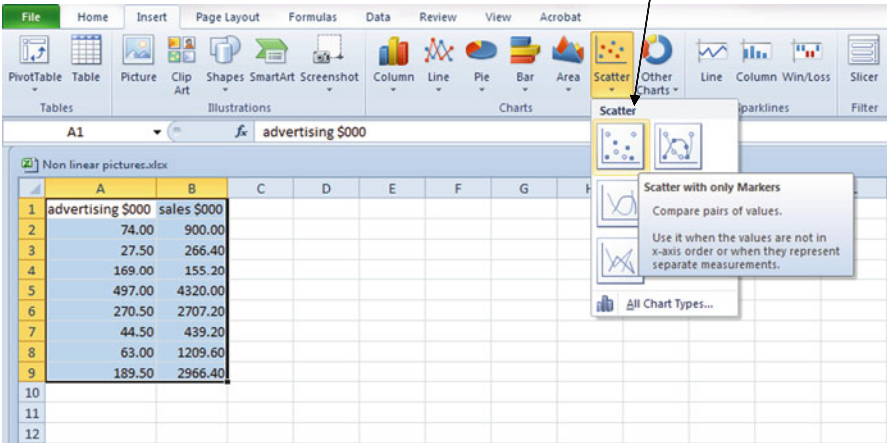
If we want to see the trendline graph output, we can only have one “x” and one “y” in our models. Excel cannot create multi-dimensional graphs. If you want to have a visual output of a multiple non-linear regression model, you need to use more sophisticated modeling software such as SAS or XGobi. In this chapter, we will consider simple (only one x) non-linear models, thereby allowing us to show trendlines on the scatterplots for each of the model types which include

- Power
- Polynomial
- Exponential
- Logarithmic

Create the Trendline Graphs

- ➢ Highlight x and y data on your spreadsheet with x as the first column followed by the y data in the next column
- ➢ Select **Insert** from the top ribbon
- ➢ Select the **Scatter** icon

➢ From the drop down menu, highlight the upper left hand corner display that shows markers without lines



A1 advertising $000

Non linear pictures.xlsx

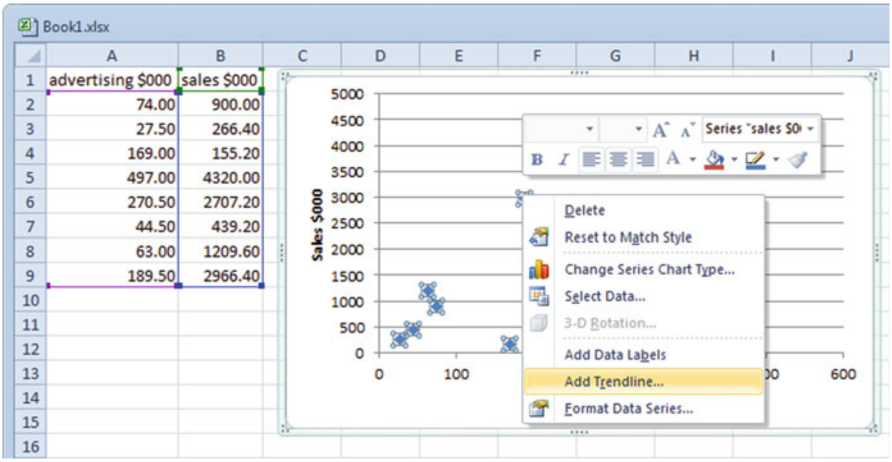
	A	B
1	advertising $000	sales $000
2	74.00	900.00
3	27.50	266.40
4	169.00	155.20
5	497.00	4320.00
6	270.50	2707.20
7	44.50	439.20
8	63.00	1209.60
9	189.50	2966.40

Scatter with only Markers

Compare pairs of values.

Use it when the values are not in x-axis order or when they represent separate measurements.

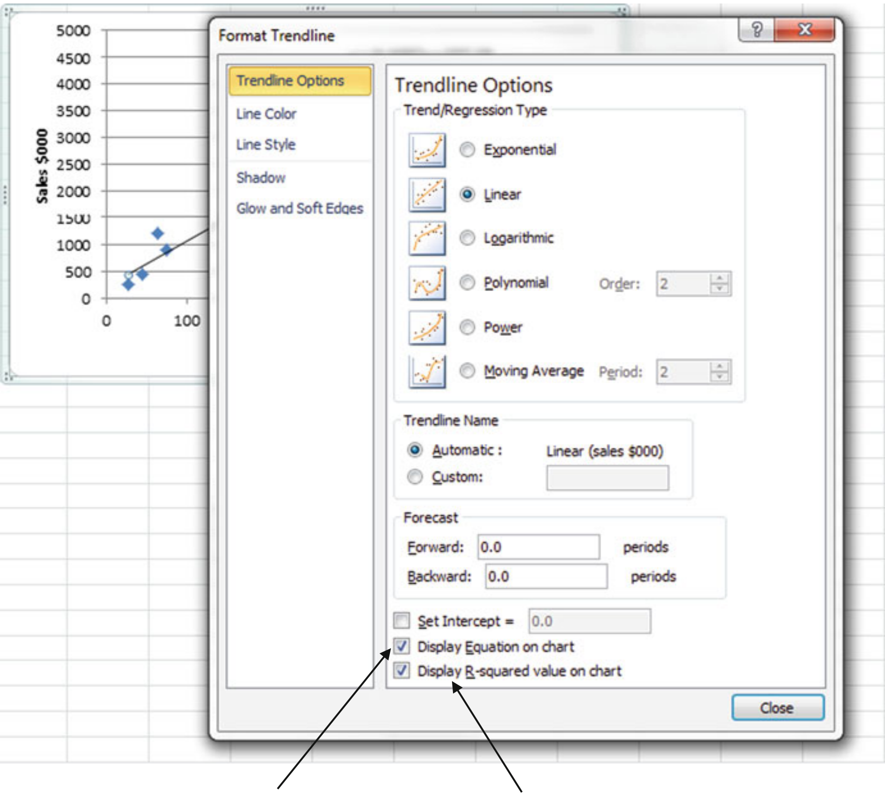
➢ Right click on any data point and select the **Add Trendline** option



Book1.xlsx

	A	B
1	advertising $000	sales $000
2	74.00	900.00
3	27.50	266.40
4	169.00	155.20
5	497.00	4320.00
6	270.50	2707.20
7	44.50	439.20
8	63.00	1209.60
9	189.50	2966.40

➢ Select the **Trendline Options** you want by clicking on the corresponding box



Format Trendline

Trendline Options

Trend/Regression Type

○ Exponential

● Linear

○ Logarithmic

○ Polynomial Order: 2

○ Power

○ Moving Average Period: 2

Trendline Name

● Automatic : Linear (sales $000)

○ Custom:

Forecast

Forward: 0.0 periods

Backward: 0.0 periods

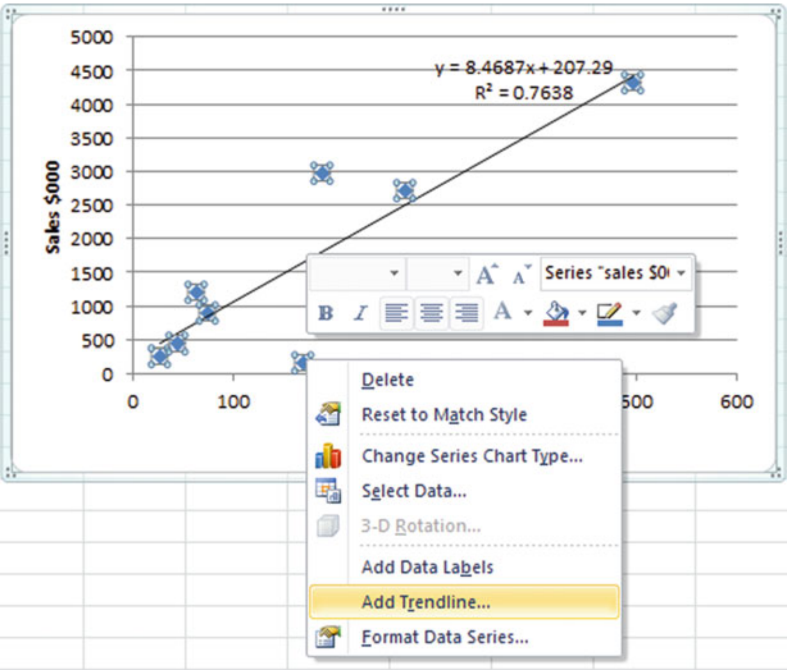
□ Set Intercept = 0.0

☑ Display Equation on chart

☑ Display R-squared value on chart

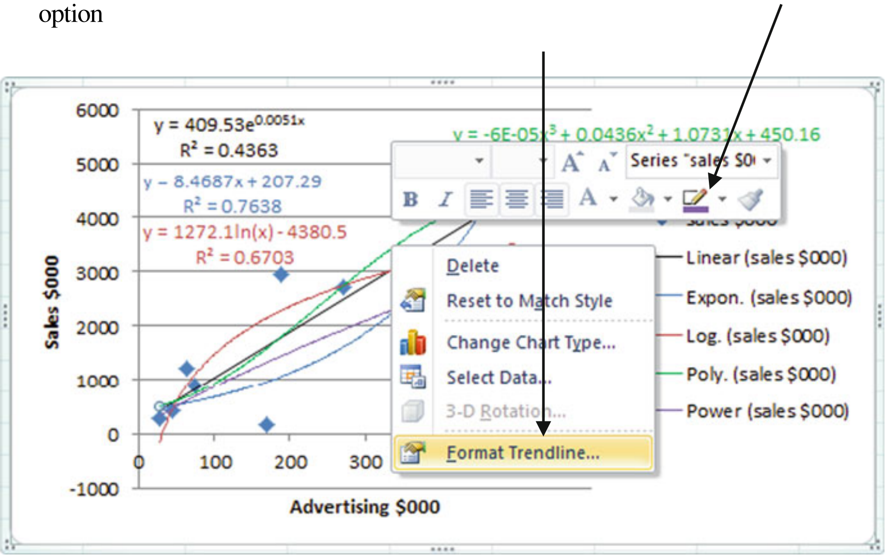
➢ Also check **Display Equation on chart** and **Display R-squared value on chart**

- ◇ You can add several different trendlines on the same graph to see which one looks like it fits best. Just keeping right clicking on a data point to choose a new type of trend line.

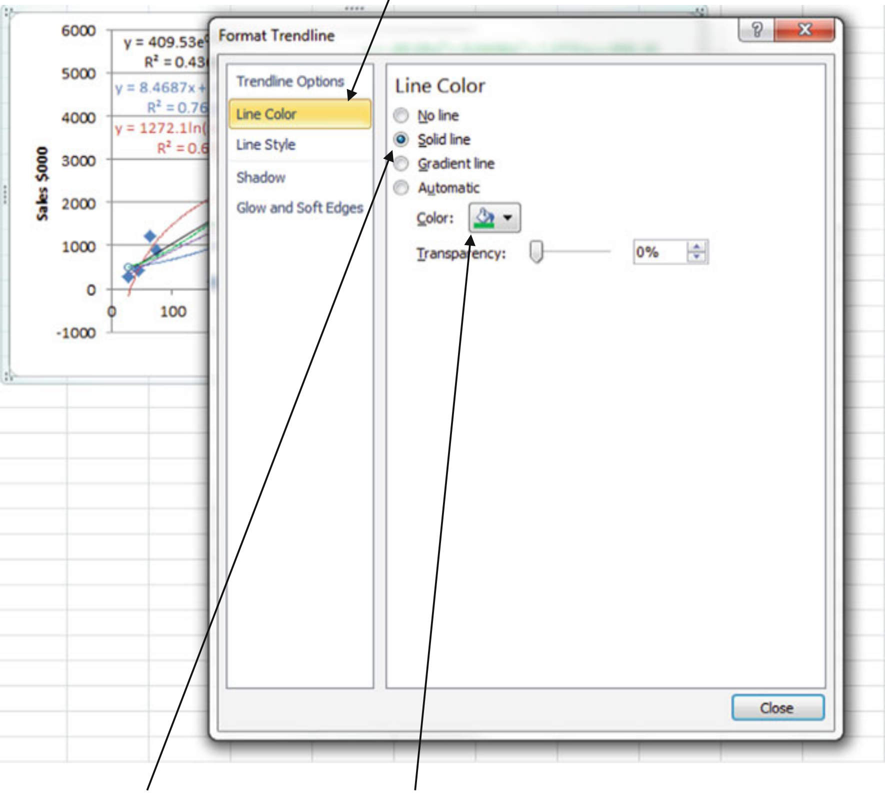


To be able to differentiate the various trendlines you might want to change the color of each of the lines. To do this:

➢ Right click on any trendline and select **Format Trendline** or select the **Outline Color** option



➢ If you **Format Trendline**, select **Line Color**


Format Trendline

Line Color

○ No line
● Solid line
○ Gradient line
○ Automatic

Color:

Transparency: 0%

➢ Select **Solid line** and choose the corresponding color

➢ Click **Close**

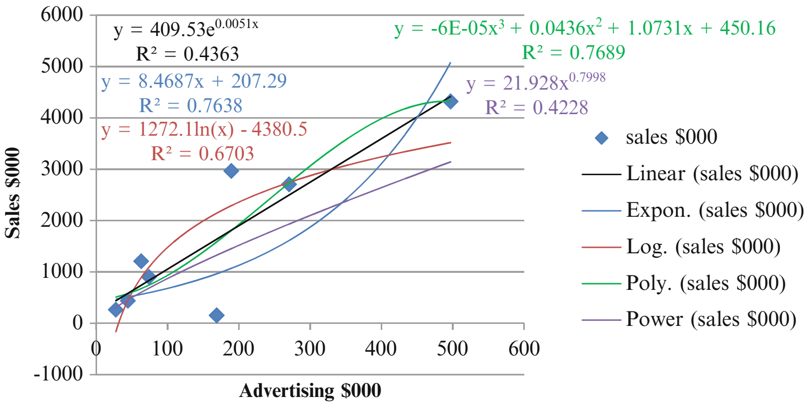


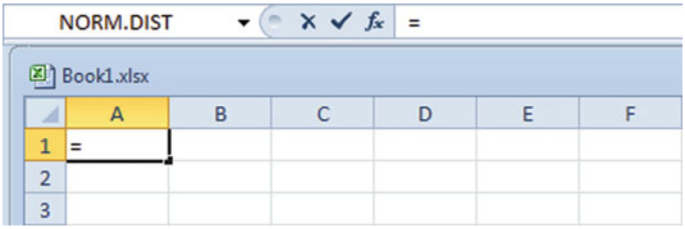
Fig. 14.7 Output data

Figure 14.7 illustrates the various output trendlines. Note we have used different colors to more easily associate the equations with their corresponding trendlines. We can compare the R^2 values to see which model explains more of the variation in the y variable. In this case it seems that either the simple linear or the third order polynomial models have R^2 of 77 %. After considering the R^2 it is usually best to select the simplest model; in this case we would select the simple linear regression model.

Using the Non-linear Regression Trendline for Prediction

To use these models we need to input the resulting trendline equations as formulas for each of these functions.

- ➢ Highlight a cell on the page
- ➢ In the formula ribbon area, type: =



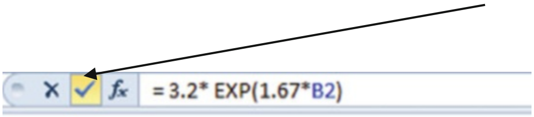
=

Now you can begin typing your actual equation. Here's how to type in the equations for each trendline model with **x** being located in cell **B2.**

Trendline type	Example of type equation	Input to excel	Let B2 have a value of 3 for x. So y equals
Power	$y = 3.2x^{1.67}$	= 3.2 * **B2** ^ 1.67	20.04
Polynomial	$y = 6x^3 + 3.2x^2 + 5x + 10$	= 6 * **B2**^3 + 3.2 * **B2** ^ 2 + 5 * **B2** + 10	215.80
Exponential	$y = 3.2e^{1.67x}$	= 3.2 * EXP(1.67 * **B2**)	479.70
Logarithmic	$y = 3.2 \text{ Ln } (x) + 5$	= 3.2 * LN(**B2**) + 5	8.52

Check that you know how to input these equations by typing a value of 3 into cell B2. When you input the equations from the third column of the table above, you should get the answers as indicated in the fourth column from the above table.

After you have typed in the formula click the green **checkmark**

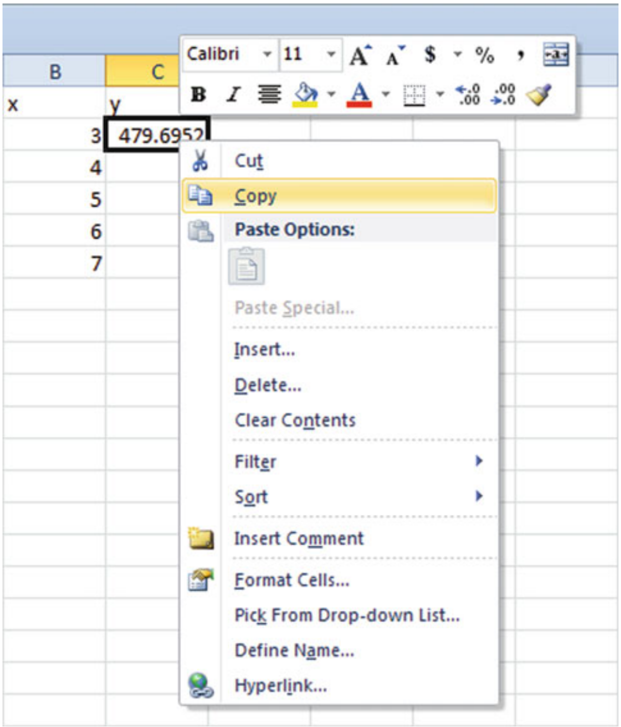


=3.2* EXP(1.67*B2)

This formula will be used to calculate the y value corresponding to the x value you input from cell B2.

To calculate other y values for given x values

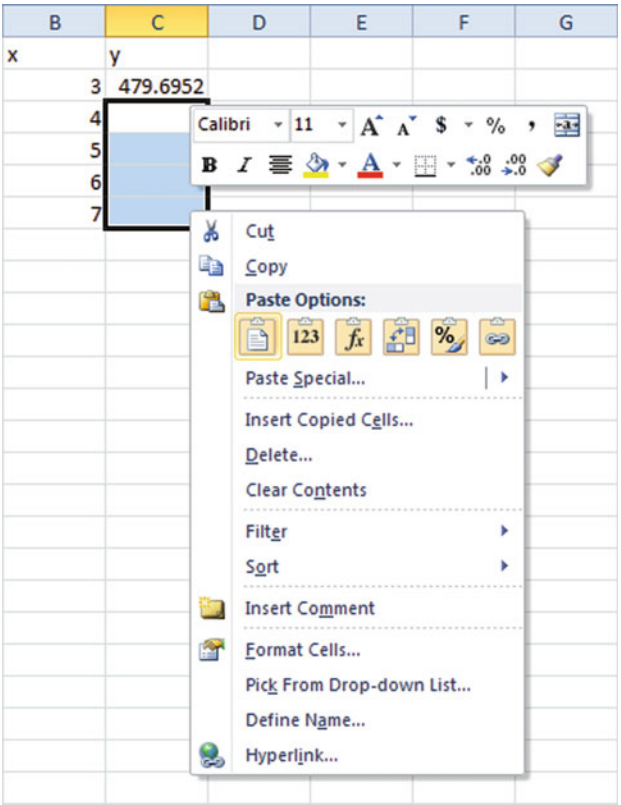
➢ Click on the cell where you first used the formula
➢ Select **Copy**



B	C
x	y
3	479.6952
4	
5	
6	
7	

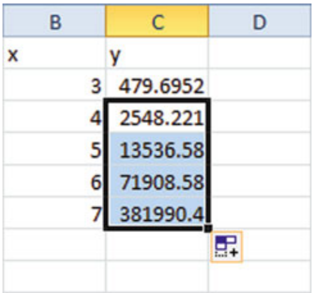
➢ Highlight the other empty y cells and type **Paste**

Before **Paste**:



B	C	D	E	F	G
x	y				
3	479.6952				
4					
5					
6					
7					

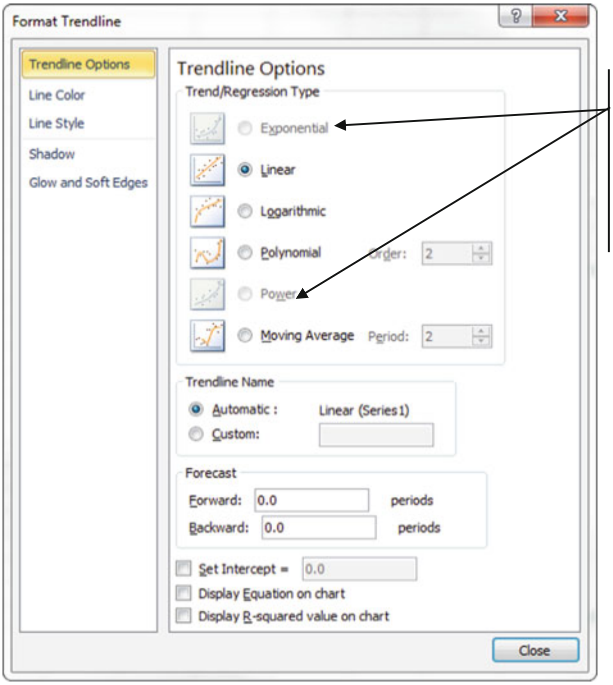
After **Paste**:



B	C	D
x	y	
3	479.6952	
4	2548.221	
5	13536.58	
6	71908.58	
7	381990.4	

Common Excel Pitfalls

- All x values have to be positive to run a power or exponential model. If you do have some negative values or values equal to zero, you can simply add a constant to all of the x data values. This will not affect the trendline fit.



Format Trendline

Trendline Options

Trend/Regression Type

○ Exponential
● Linear
○ Logarithmic
○ Polynomial Order: 2
○ Power
○ Moving Average Period: 2

Trendline Name

● Automatic : Linear (Series1)
○ Custom:

Forecast

Forward: 0.0 periods
Backward: 0.0 periods

□ Set Intercept = 0.0
□ Display Equation on chart
□ Display R-squared value on chart

Note: Excel will gray out trendlines that are not feasible. For example, it cannot run exponential or power functions when negative numbers are present.

- Make sure you know what form of logarithm you need to work with. We have only used the natural logarithm in this chapter. If you want to use a base other than the constant e, you must use a different form of this function

y = LN(number)	Natural log
y = LOG (number,base)	Any base
y = LOG10 (number)	Base 10

- When adding in trend lines, all the lines will appear in the default color, unless you format the color of the line.
- Make sure you have an equal amount of "(" as you do ")". If you do not have enough close parentheses, Excel will give you an error or may not calculate your equation correctly.

Final Thoughts and Activities

Practice Problems and Case Studies

1. A technology manager of a small company would like to better understand the trends in laptop depreciation prior to purchasing a new set of laptops for the employees. She reviews laptop prices over the course of 9 months with the following results:

# of months	Price ($)
1	2,750
2	2,000
3	1,550
4	1,250
5	950
6	750
7	625
8	500
9	375

   (a) Construct a scatter plot.
   (b) Determine the appropriate non-linear function that matches the data.
   (c) What does the model tell you about the relationship between laptop price and time?

2. A travel expense coordinator at a hotel needs to understand the fluctuation of gas prices over the course of a year to better plan the department's travel budget. After doing some research, he acquired the following data:

Month	# of months	Average gas price ($)
January	1	3.07
February	2	3.20
March	3	3.55
April	4	3.77
May	5	3.95
June	6	3.72
July	7	3.61
August	8	3.58
September	9	3.50
October	10	3.40
November	11	3.30
December	12	3.27

(a) Create a frequency plot.
(b) Determine the appropriate non-linear function that matches the data.
(c) What does the model tell about the relationship between average gas prices over the course of 12 months?

3. An ad agency wanted to understand how the number of ads shown in one day related to the number of people who were aware of the advertisement. The data they gathered is displayed below:

Number of ads/day	# of people aware (0000s)
2	1.00
3	1.25
4	2.25
5	6.25
6	22.25
7	86.25
8	342.25

(a) Construct a scatter plot.
(b) Determine the appropriate non-linear function that best matches the data.
(c) What does the model tell you about the relationship the number of ads shown in one day related to the number of people who were aware of the advertisement?

Discussion Boards

1. The pricing of oil is a complex process. Consumers are often confused that the "price at the pump" does not seem to track with corporate profits or the discovery of new oil fields. Collect the price at the pump data and plot it against the corporate net profit figures for one of the major oil companies over the past 24 months. Are you surprised at the shape of the resulting graph?
2. Most statistical factors, relationships and processes involving people have a behavioral component. This means they do not follow a linear model. Why is that?

Group Activity

Change management processes are often evaluated using linear models. In other words as change is implemented, performance improvement is expected to follow. In fact performance may decline, flatten or improve. The punctuated equilibrium model describes periods of inertia followed by periods of positive energy. Choose a recent change initiative in your own organization. Which of the non-linear models might best fit the process and why?

Parting Thought

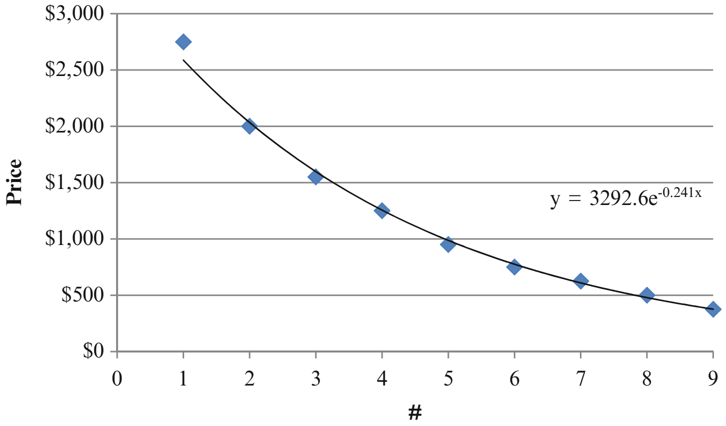
Statistically speaking, in China, even if you are a one in a million kind of guy, there are a thousand more just like you.

Problem Solutions

1. A technology manager of a small company would like to better understand the trends in laptop depreciation prior to purchasing a new set of laptops for the employees. She reviews laptop prices over the course of 9 months with the following results:

Month	Price($)
1	2,750
2	2,000
3	1,550
4	1,250
5	950
6	750
7	625
8	500
9	375

(a) Construct a scatter plot



(b) Determine the appropriate non-linear function that matches the data.

A negative exponential function fits the chart best.

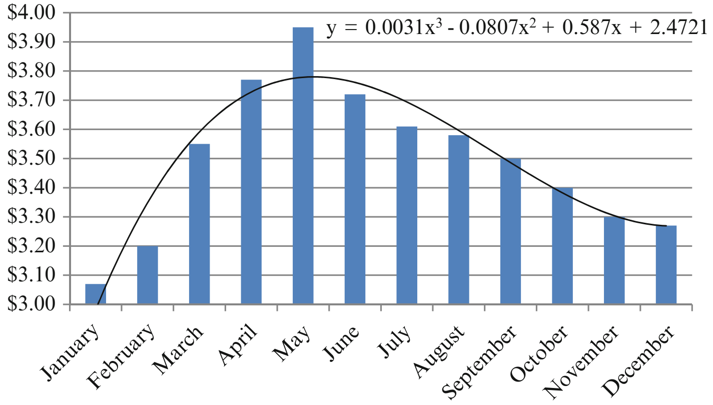
(c) What does the model tell about the relationship between laptop price depreciation over time?

The data shows that as time progresses the depreciation of value for a laptop is exponentially negative, or drops exponentially.

2. A travel expense coordinator at a hotel needs to understand the fluctuation of gas prices over the course of a year to better plan the department's travel budget. After doing some research, he acquired the following data:

Month	# of months	Average gas price ($)
January	1	3.07
February	2	3.20
March	3	3.55
April	4	3.77
May	5	3.95
June	6	3.72
July	7	3.61
August	8	3.58
September	9	3.50
October	10	3.40
November	11	3.30
December	12	3.27

(a) Plot the data



(b) Determine the appropriate non-linear function that matches the data

A polynomial trendline best fits the data.

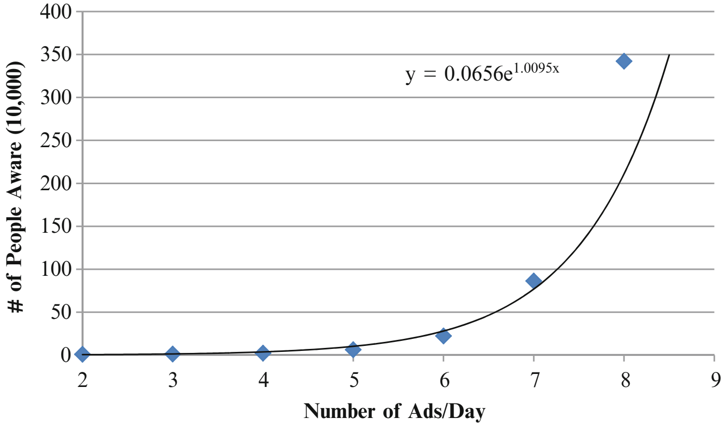
(c) What does the model tell about the relationship between average gas prices over the course of 1 year?

The trendline shows that over the course of 1 year gas prices fluctuate and reach a peak in the summer months but decrease for the fall and winter.

3. An ad agency wanted to understand how the number of ads shown in one day related to the number of people who were aware of the advertisement. The data they gathered is displayed below:

Number of ads/day	# of people aware (0000s)
2	1.00
3	1.25
4	2.25
5	6.25
6	22.25
7	86.25
8	342.25

(a) Construct a scatter plot



(b) Determine the nonlinear function that matches the data.

$$y = 0.0656e^{1.0095x}$$

(c) What does the model tell about the relationship between that the number of ads and the number of people who are aware of the ad?

The chart shows that the number of ads is positively exponentially related to the number of people who are aware of the ad.

Chapter 15
Survey Reports

Case Study: Creating an Impactful Report for Infinity Auto Insurance

To better understand the Hispanic market in Houston, Texas, Infinity Auto Insurance sampled 400 respondents in Houston, Texas area. The sample design and specifications of the study are summarized in the table below.

Design	Specification
Methodology	Telephone survey
Population	Residents of the City of Houston, Texas
Sample size	200

It's just our way of saying thanks ...
INFINITY DriverClub®
TOWING
JUMP-START
TIRE CHANGE
LOCKOUT
REFUELING
read conmigo
Sponsored by
INFINITY Auto Insurance
INFINITY Auto Insurance

As a thank you for signing up for our Read Conmigo program, please enjoy a **FREE lifetime membership in the Infinity DriverClub®**. Signing up is easy! Just call 800-516-6894 or go to InfinityDriverClub.com. We believe that no one should be stuck on the side of the road. Join Infinity DriverClub® and receive access to 24/7 Nationwide Roadside Assistance. You are covered:

- Any time, in any car
- Pay only for the services you use
- No membership fees
- Coverage available everywhere in the USA

(continued)

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3_15,

Case Study: (continued)

To help Infinity Auto Insurance better understand the gathered data and using your new found knowledge of statistics and survey reporting, create a report using the data provided below. With what you have learned from this book, write a report that includes this survey's sampling methodology, survey creation, results, conclusions, and recommendations.

Own vehicle

	Frequency	Valid %
Yes	179	89.5
No	21	10.5
Total	200	100.0

Own insurance

	Frequency	Valid %
Yes	150	75
No	50	25
Total	200	100.0

Switch insurance

	Frequency	Valid %
Not at all likely	92	61.3
Not likely	9	6.0
Somewhat likely	29	19.3
Likely	13	8.7
Extremely likely	7	4.7
Total	150	100.0

Note: This question is only asked of people do own vehicle insurance.

Heard of Fu Auto Insurance

	Frequency	Valid %
Yes	150	75.0
No	50	25.0
Total	200	100.0

Heard of ES Auto Insurance

	Frequency	Valid %
Yes	50	25.0
No	150	75.0
Total	200	100.0

Heard of Infinity Auto

	Frequency	Valid %
Yes	100	50.0
No	100	50.0
Total	200	100.0

(continued)

Case Study: (continued)

Seen Infinity Auto Ad

	Frequency	Valid %
Yes	86	86.0
No	14	14.0
Total	100	100.0

Seen Fu Auto Insurance Ad

	Frequency	Valid %
Yes	125	83.3
No	25	16.7
Total	150	100.0

Seen ES Auto Insurance Ad

	Frequency	Valid%
Yes	30	60.0
No	20	40.0
Total	50	100.0

Fu Auto Insurance Favorability

	Frequency	Valid %
Very unfavorable	3	2
Unfavorable	6	4
Somewhat favorable	32	21.3
Favorable	51	34
Very favorable	58	38.7
Total	150	100.0

Fu Auto Insurance Favorability Mean: 4.03

ES Auto Insurance Favorability

	Frequency	Valid %
Very unfavorable	2	4
Unfavorable	3	6
Somewhat favorable	17	34
Favorable	16	32
Very favorable	12	24
Total	50	100.0

ES Auto Insurance Favorability Mean: 3.66

Infinity Auto Favorability

	Frequency	Valid %
Very unfavorable	9	9.0
Unfavorable	7	7.0

(continued)

Case Study: (continued)

Infinity Auto Favorability

	Frequency	Valid %
Somewhat favorable	22	22.0
Favorable	32	32.0
Very favorable	30	30.0
Total	100	100.0

Infinity Auto Favorability Mean: 3.67
Demographics

Spoken language

	Frequency	Valid %
English	32	16.0
Spanish	168	84.0
Total	200	100.0

Ad Language preference

	Frequency	Valid %
English-language advertisement	29	14.5
Spanish-language advertisement	171	85.5
Total	200	100.0

Gender

	Frequency	Valid %
Male	97	48.5
Female	103	51.5
Total	200	100.0

Key Concepts

Cross-tabulation, Conclusions and Recommendations, Executive summary, Multi-analysis report, Overall report, and Pilot testing

Discussion

The earlier chapters introduced the basic business statistics and this final chapter will integrate those concepts in the discussion of survey reports and presentations.

We often hear that a picture is worth a thousand words. In this chapter, we focus our attention on how to use Excel to visually enhance our reports. The audience is often critical about how much information is necessary to display and report. For example, XYZ, Inc. has 30,000 stores worldwide. Operational results can vary from one part of the world to another: a store manager in New York City is probably not concerned about the operational results from a store located in Milan, Italy. Therefore, we must target our results to our audience of interest.

Once all the data has been collected and the statistical analysis conducted, the research is usually not complete until a final report is written. The report should clearly explain to the intended audience the essential findings of the study/research. There are many ways to approach the survey report process. The following fundamentals of a standard survey report are discussed in this chapter.

I. Title Page
II. Table of Contents
III. Executive Summary
IV. Background and Objectives
V. Research Methodology
VI. Overview of Results
VII. Conclusions and Recommendations
VIII. Bibliography
IX. Appendices

When your survey and analysis has been completed, the final step in the survey process is to present your findings, which involves the creation of a research report. This report should include a background of why you conducted the survey, a breakdown of the results, and conclusions and recommendations supported by this material.

Title Page

The title for a report should be simple yet attention-grabbing. Generally, the title will give the recipient of the report an overview or highlight of its contents. The title should be long enough to relay the point across but short enough to hold the recipient's interest. An example of an attention grabbing title would be: "Los Angeles: Extremely Loyal Market Place with Company Reputation."

Include the names of those who prepared the report, to whom it will be presented, and the submission date of the report.

Table of Contents

List the sections in your report. Here is where you give a high-level overview of the topics to be discussed in the order they are presented in the report. Depending on the length of your report, you should consider including a listing of all charts and graphs so that your audience can quickly locate them.

Executive Summary

The Executive Summary is located at the beginning of the report, following the Table of Contents. The Executive Summary provides the recipient of the report with a high-level overview of the study. This section usually encompasses no more than two pages, preferably only a few paragraphs. It includes a quick summary of the survey's background, objective, results, and sometimes a few key conclusions or recommendations. The outline explains what the recipient will be seeing on the pages to follow.

Example of a Key Finding in the Executive Summary *This study shows that over the past three years, people's impression of the Company has steadily increased. The Company has the highest rate of favorability compared to its competitors in the Los Angeles and Orange County area.*

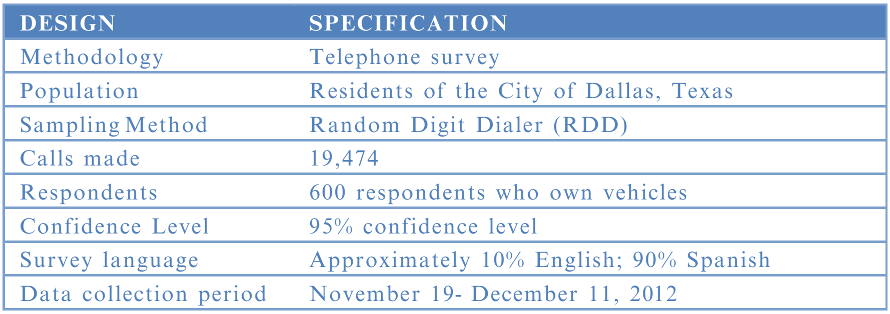
Background and Objectives

The Background section explains what led up to conducting the study and its purpose, whereas the Objectives section explains the survey study's goals. The Background section details to the recipient, who may or may not have any prior knowledge of the study, who requested the study, why the study was requested, the purpose of the study, what previous studies may have been completed, and what the study measured.

The Objectives section describes the main purpose or goal of the study and the supporting rationale for doing the study. Often times, survey studies have multiple objectives, or secondary objectives. Each of the objectives should be stated or outlined, and each separate objective should be explained as necessary, along with any goals or expected outcomes for the objectives.

Example of an Objective

Objective: Taking the research and coalescing the information together in an annual report will integrate all the avenues of research and create a more meaningful picture regarding the organization's target market areas and population.



DESIGN	SPECIFICATION
Methodology	Telephone survey
Population	Residents of the City of Dallas, Texas
Sampling Method	Random Digit Dialer (RDD)
Calls made	19,474
Respondents	600 respondents who own vehicles
Confidence Level	95% confidence level
Survey language	Approximately 10% English; 90% Spanish
Data collection period	November 19- December 11, 2012

Fig. 15.1 Example summary table

Research Methodology

The Research Methodology details every step taken to prepare, execute, and analyze this study. It explains the process of the survey study and provides the audience with a general timeline of events. Include a copy of the survey in the appendix section of the report.

The first step detailed in the Research Methodology section is usually the sampling plan. The sampling plan should explain the rational for the sample size as well as discuss the sample representativeness.

The next section should describe the type of survey and the procedures taken to create, administer, and collect the survey and data. When explaining the survey creation, specify who participated in the creation of the survey, what type of questions were included, and who approved the final survey. **Pilot testing** involves sending out the survey to “test the waters”; it is an opportunity to get feedback on the survey to allow for improvements before the final survey goes out. The report should include comments on any pilot testing that was completed. The survey administration and data collection section illuminates the procedures taken before and during the data collection process. It can include pilot test data information, how the survey was distributed or collected (via mail, email, telephone, etc.), and the basic details of the data collection process, including any challenges that were overcome.

A quick summary of the sample methodology at the end of the Research Methodology section can prove helpful. Placing the sample methodology into a table shares the necessary information in a concise manner. Figure 15.1 below is an example of a summary table.

Results

The Results section gives the audience the basic results of the study. In this section, the survey results are explained in detail, usually in words rather than in charts and figures. For readability purposes, the results usually follow the order in which they appear in the survey. Categorizing the sections of the survey helps us group the results into more meaningful themes and concepts. The rationale for why the results for a specific question should be explained in this section. By explaining the results, we help the audience understand not just the basic numbers but what the numbers actually mean. This is the section where you can discuss the demographics (measurable characteristics) of the sample.

Example of Results *The sample of 600 respondents is 51 % male. Ninety percent of respondents primarily speak Spanish at home and 86 % prefer advertisements in Spanish as opposed to in English.*

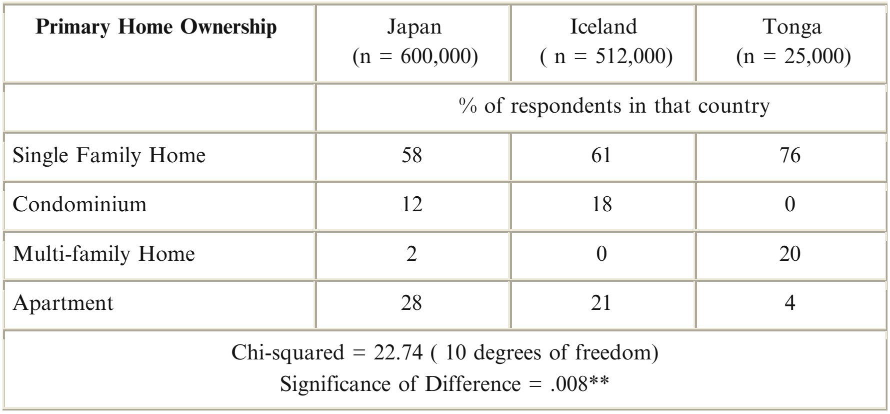
Cross-Tabulate Relevant Pairs of Questions

Most questions don't have much meaning on their own, so you need to compare or **cross-tabulate** them with other questions. If the answers to a cross-tab are not statistically significant but nonetheless are important to the organization, you may still choose to report the results for completeness, i.e. "Question 3 produced no statistically significant differences at the .05 level for different genders, age groups, or geographical areas."

When putting tables into a report, remember than you can only squeeze about 8 columns across a page, and about 50 rows down. If there are too many different answers, you'll need to recode the question to reduce the number of columns (or, sometimes, rows).

Make sure that readers cannot misinterpret the percentages. The normal convention is that percentages add to 100 downwards. If they add across (usually done to fit the table onto a single page) you must show this very clearly. See Fig. 15.2.

Note in Fig. 15.2 that the sample sizes and the overall significance results are provided. For a review of significance results see Chaps. 11 and 12.



Primary Home Ownership	Japan (n = 600,000)	Iceland (n = 512,000)	Tonga (n = 25,000)
	% of respondents in that country		
Single Family Home	58	61	76
Condominium	12	18	0
Multi-family Home	2	0	20
Apartment	28	21	4
Chi-squared = 22.74 (10 degrees of freedom) Significance of Difference = .008**			

Fig. 15.2 Cross tab example

Conclusions and Recommendations

The Conclusions and Recommendations section is the main emphasis of the report. Often the Conclusions and Recommendations are grouped together. The Conclusions should summarize and explain the results, where the Recommendations give these results applications and meanings. This section does not include graphs, charts, or tables. It emphasizes the purpose of the report and what it means. The conclusions should explain what inferences can be drawn from the data. The Recommendations should draw upon the inferences from the Conclusions and provide limitations of the study with possible improvements for future work. When writing recommendations, always draw on the information gathered from the survey and restate it in the Conclusions. We need to ensure that a link exists between the actual data and the recommendations.

Example of a Conclusion and Recommendation: *Because the Los Angeles/Orange County market area is doing well, we suggest that the Company conduct focus groups in that area to see what is working best and what trends have had the most impact on that area. These focus groups would not only help Los Angeles sustain its favorability rating but also give other marketing areas insights on how to improve their own visibility and favorability.*

Bibliography

The Bibliography lists the sources you consulted during your survey and any additional sources you may have referred to in the report. The Bibliography section is necessary if we refer to different studies or references in journals, or if we've used information from external sources, such as newspapers, the internet, etc., in our report. For example, we want to explain our rationale for using both landlines and mobile numbers for our telephone survey. From our internet research we found a journal article stating what percentage of people use their mobile phone as their main phone number. If we refer to the percentage in the journal article, we will need to site it in our bibliography.

Example of a Bibliography Entry *Blumberg, S. J., & Luke, J. V. (2009). Reevaluating the need for concern regarding noncoverage bias in landline surveys. American Journal of Public Health, 99(10), 1806–10. doi: 10.2105/AJPH.2008.152835*

Appendices

The Appendix section should include any supplementary information that will help your audience understand the survey study better, such as any additional graphs, charts, and tables, a copy of the actual survey, the invitation email, etc. This is a good place to put transcripts of historic documents and other lengthy bits of documentation that do not fit comfortably in the main body of your report.

Useful Hints and Phrases for the Report

Make sure you clearly note on the cover page of the report the level of privacy associated with the content. The more common labels for privacy include:

- **Confidential:** Private information not for general distribution, usually limited to only approved individuals, often limited to just those in the presentation room
- **For Management Review Only:** Distribution limited to management levels, not for distribution to employees below the rank of manager

It is common for certain phases to get repeated over and over throughout the report. Here are several common phases with examples of how to say the same thing.

- **To introduce:** *The purpose/aim of this report, As requested, This survey was carried out/ conducted by means of. . ., The questionnaire consisted of etc.*
- **To generalize:** *In general, Generally, On the whole, etc.*
- **To refer to a fact:** *The fact is that. . ., In fact, In practice, etc.*
- **To conclude/summarize:** *In conclusion, All things considered, To sum up, all in all, It is not easy to reach any definite conclusions, If any conclusions may be drawn from the data, It is clear that, The survey shows/indicates/demonstrates, etc.*

Present tenses, reported speech and an impersonal style should be used in survey reports. Use a variety of reporting verbs such as claim, state, report, agree, complain, suggest, etc.

Effective PowerPoint Presentations with Excel

Often the delivery of the report content is also requested as a PowerPoint presentation. Excel slides can be too complex and confusing for the audience to easily follow, so here are some useful hints to share the Excel data in a more "audience-friendly" way.

1. Use charts and graphs to illustrate big ideas. Keep it simple, and don't give too much detail.
2. If reproduction will be in color then use contrasting colors. However, if any copies may get made in black and white, use different black and white patterns. Check if the audience or client has brand colors that are important to utilize.
3. Use graphics for instructional purposes or reasons. Don't add clip-art just to add art that can confuse or distract the audience.

4. Don’t crowd your reports with too much information. Make sure to use appropriate amounts of white space.
5. Make good use of the space available in the slide. Enlarge the graphs and have the text large enough that it is easy to read from across a room.
6. One graph per slide is ideal. Two graphs maximum per slide. This will make your data easier to understand. If you must have two visuals, make sure the accompanying text is simple.
7. Use fonts appropriately. Serif fonts are typically easier to read. However, the audience or client may have a font preference. Reports should not have more than two fonts throughout the report. Experiment with type styles, sizes, and colors. Don’t be afraid to bold text, underline or italicize if you are trying to emphasize a point.
8. Keep titles short. About five to seven words will get your point across.
9. Avoid using busy slide backgrounds. Multiple colors or gradients can make text hard to read.
10. Avoid regurgitating a lot of data without interpretation of that data. The details are appropriate for the Appendix of the report, not for the presentation or main body of the report.
11. Make sure your presentation “stands on its own”. Some people may forward the presentation to other interested parties i.e. those employees who missed the initial presentation; make sure they will be able to understand your work and recommendations by reading the slides.

Executive Summary

1. Background

   Infinity Auto Insurance is looking to expand their customer base starting with large metropolitan areas across the United States. Because the Latino/Hispanic population is the largest growing population in the United States, Infinity Auto Insurance wishes to better understand this market segment and expand its advertising campaigns to include this population. They have decided to concentrate their strategic efforts to large metropolitan areas in 4 different states: California, Arizona, Texas, and Florida. For initial study, they sampled 200 people in Houston, TX.
2. Methodology

Design	Specification
Methodology	Telephone survey
Population	Residents of the City of Houston, Texas
Sample size	200

3. Results
   (a) Over 85 % of respondents own a vehicle.
   (b) Roughly 75 % of respondents own auto insurance.
   (c) Most respondents who own auto insurance do not plan to switch insurance in the next 12 months.
   (d) About 50 % of people have heard of Infinity.

Methodology

Sampling Plan and Survey Creation

Infinity Property and Casualty Corporation (Infinity Auto Insurance) headquartered in Birmingham, Alabama provides personal automobile insurance with a concentration on nonstandard auto insurance. Nonstandard insurance serves individuals unable to obtain coverage through standard insurance companies, which can be due to a driving record with accidents and/or tickets, prior DUI, the driver's age, vehicle type, etc. Infinity Auto Insurance's products include personal automobile insurance for individuals, commercial vehicle insurance for businesses and classic collector insurance for individuals with classic and antique automobiles. Infinity Auto Insurance distributes its products primarily through independent agencies and brokers. Infinity Auto Insurance, a top-performing Infinity brand, provides nonstandard car insurance through more than 12,500 independent agents. Infinity Auto Insurance utilizes Internet-based software applications to provide many of its agents with real-time underwriting, claims and policy information. The Company is licensed to write insurance in all 50 states and the District of Columbia.

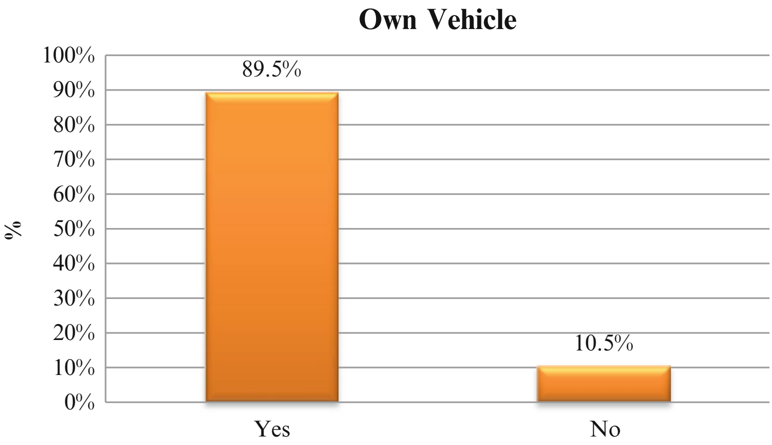
Infinity Auto Insurance is looking to expand their customer base starting with large metropolitan areas across the United States. Because the Latino/Hispanic population is the largest growing population in the United States, Infinity Auto Insurance wishes to better understand this market segment and expand its advertising campaigns to include this population. They have decided to concentrate their strategic efforts to large metropolitan areas in 4 different states: California, Arizona, Texas, and Florida. For this initial study, they sampled 200 people in Houston, TX.

Roughly 2.1 Million Hispanics live in Houston, Texas. Infinity Auto Insurance has asked you to help them create a market research survey for the Houston, Texas Area. The Infinity marketing group wants to better understand the Hispanic automobile insurance market in this large metropolitan area. Infinity Auto Insurance has decided that a telephone survey would reach their target population (less acculturated Hispanics) better than any other data collection method. To get a representative sample, Infinity Auto Insurance has decided that 200 respondents would be sufficient to draw some conclusions.

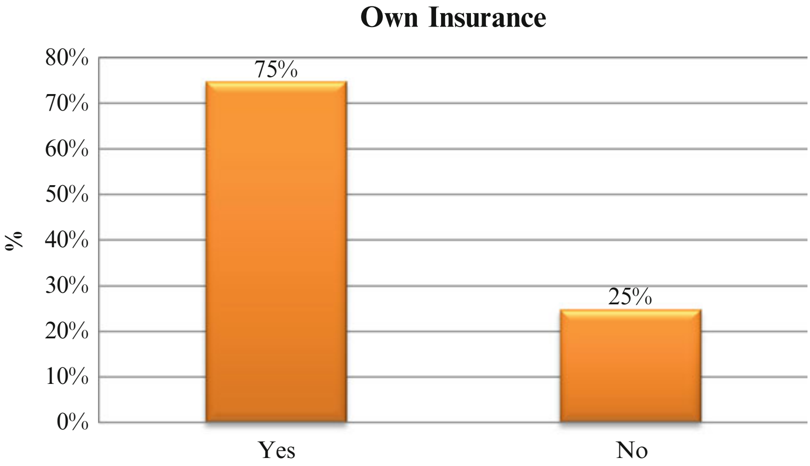
Infinity Auto Insurance wanted to better understand some of the following information: what percentage of Hispanics are insured, what companies are Hispanics insured with, how willing are Hispanics to buy or switch insurance companies, etc. Finally, Infinity Auto Insurance would like to know what attributes keep Hispanics buying auto insurance in the Houston area.

After speaking with Subject Matter Experts (SMEs) and doing research on the Houston area, we created a survey based on Infinity Auto Insurance's needs.

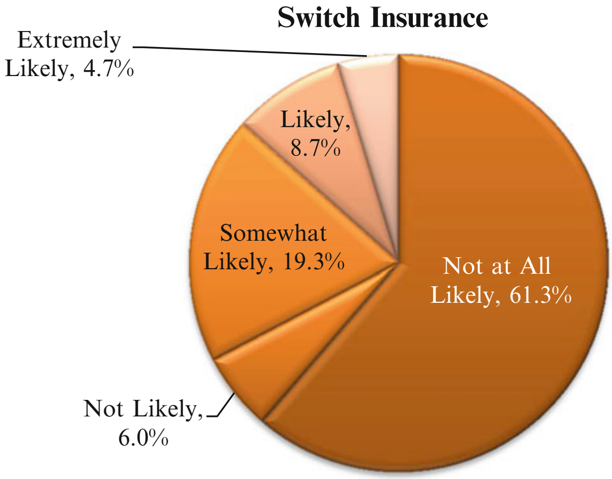
Results



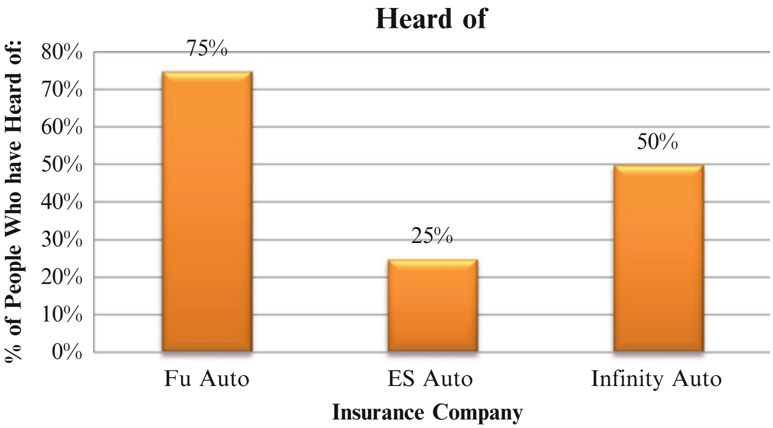
The above figure depicts the responses for those who own a vehicle. Of the sample of 200 respondents, 89.5 % of people own a vehicle.



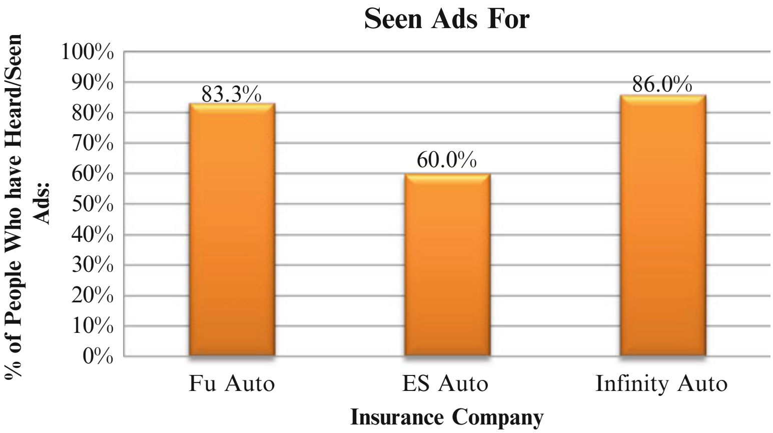
The above figure depicts the responses for those who own a vehicle. Of the sample of 200 respondents, 75 % of people own auto insurance.



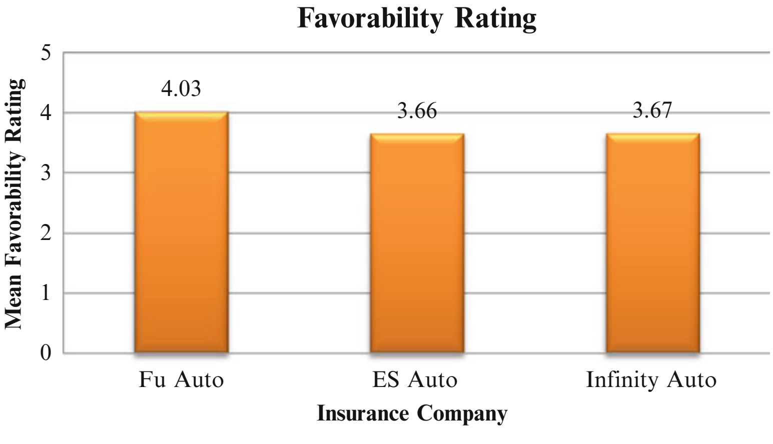
The figure above shows the responses to the question: "How likely or unlikely would you be to switch auto insurance companies in the next 12 months?" When insured respondents were asked whether or not they would switch auto insurance companies, 61.3 % of respondents said that they were "Not at All Likely" to switch.



The figure above shows the responses to the question: "Have you heard of [Fu Auto, ES Auto, Infinity Auto] Insurance?" When asked if respondents have heard of Fu Auto Insurance, ES Auto Insurance and Infinity Auto, most respondents have heard of Fu Auto Insurance.



The figure above shows the responses to the question: "Have you seen ads for [Fu Auto, ES Auto, Infinity Auto] Insurance?" When asked of respondents have heard of Fu Auto Insurance, ES Auto Insurance and Infinity Auto whether they have seen advertisements for those companies, most respondents have heard of Infinity Auto Insurance.

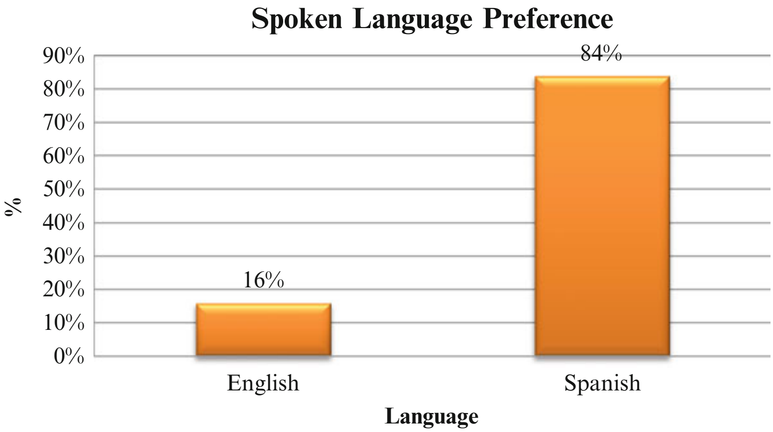


The figure above shows the responses to the question: "How favorable or unfavorable would you rate [Fu Auto, ES Auto, Infinity Auto Insurance] based on their reputation?"

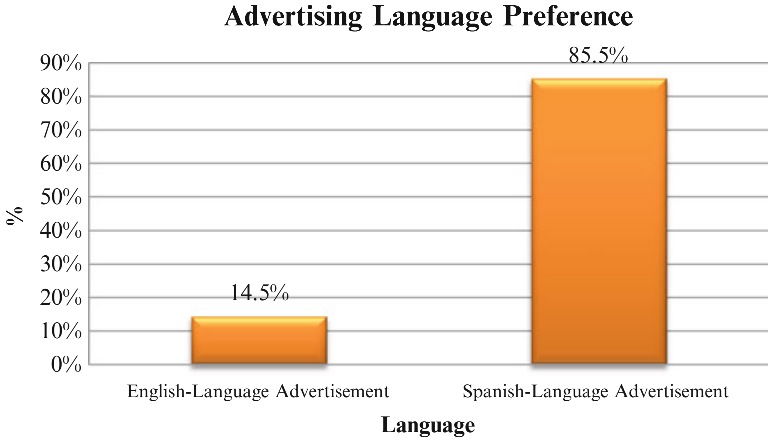
Scale:

1 = "Very Unfavorable" 2 = "Unfavorable" 3 = "Somewhat Favorable"
4 = "Favorable" 5 = "Very Favorable."

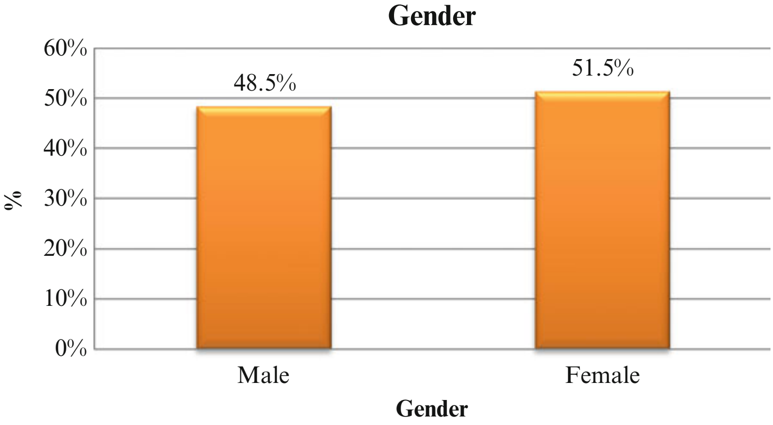
When asked of respondents have heard of Fu Auto Insurance, ES Auto Insurance and Infinity Auto Insurance how favorable respondents would rate those companies, Fu Auto Insurance ranked the highest with a 4.03 out of 5 favorability rating.



The figure above shows the responses to the question: “What language do you primarily speak in your home?” Most respondents (84 %) stated that they spoke Spanish at home.



The figure above shows the responses to the question: “Do you prefer English-language advertising or Spanish-language advertising?” Most respondents (85.5 %) stated that they prefer Spanish-language advertising.



The figure above shows the responses to the question: "What gender are you?" 51.5 % of respondents denoted that they are female.

Conclusions and Recommendations

- This is the first study done in the Houston, TX area. We suggest annually conducting the same random telephone interviews to keep track of Infinity Auto Insurance's brand reputation in the Houston area. Each market area has some individual differences and should be monitored as significant advertising changes are made in the Houston market.
- The Houston market area is relatively brand loyal with only about 33 % of respondents willing to switch auto insurance companies in the next 12 months. Infinity Marketing could coordinate a few insight sessions Hispanic customers which could reveal interesting auto insurance preferences. These insight sessions can help build on how to get Hispanic customers to switch, attract (advertise), and retain future Infinity Auto Insurance.

Final Thoughts and Activities

Practice Problems and Case Studies

1. Review the attached report and identify areas of inaccuracy and areas for improvement.

++

Tiga Consulting
Lunch Room Report for EIGC

To: S.M. Riesling, General Manager
From: Paula Simms
Date: June 11
Subject: Lunch Room Vending Machine

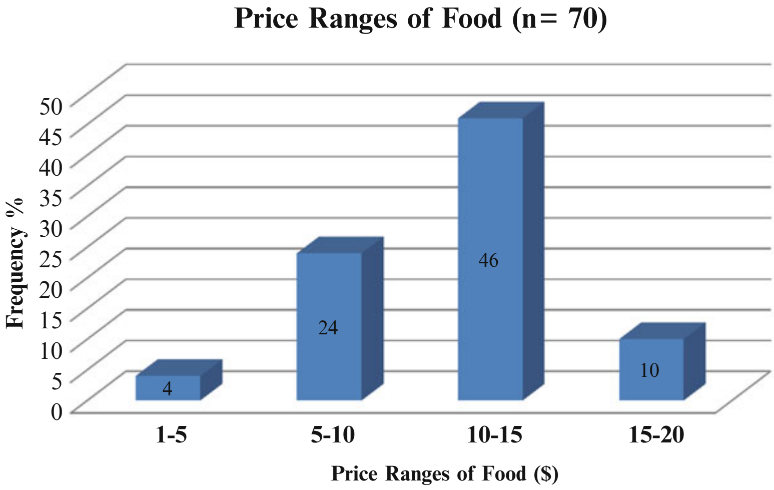
1. At the monthly staff meeting on May 21 20XX, you requested our company provide a report about staff satisfaction with the new high end vending machine. The aim of this report is to present this information with recommendations.
2. Since the move to the new office location in Brownsville, employees have had difficulty in finding a nearby place to buy lunch. The new office is in a factory area and restaurants are limited. A new state of the art bistro vending machine was purchased. However some employees expressed dissatisfaction with this vending machine.
3. All 122 employees were surveyed by an online questionnaire on May 21 and May 22.
4. In general, staff members are not satisfied with the prices and food choices in the vending machine.
5. Some were dissatisfied with the location of the vending machine, but most were dissatisfied with its reliability.
6. On analyzing the data, two distinct groups of staff emerged. The first includes 68 staff that had usually eaten in the office when the office was in West Grove. The second group of 40 staff had usually eaten outside the facility in West Grove.
7. The overall response rate was 90 % with 108 responses.

Table 1 Staff satisfaction with the vending machine

	Staff satisfaction		
Feature	Group A: Usually ate in the office	Group B: Usually ate outside the office	Both groups
Price	55 %	35 %	48 %
Food choices	30 %	25 %	28 %
Location	70 %	75 %	71 %
Reliability	95 %	95 %	95 %
Overall avg	62.5 %	57.5 %	60.6 %

8. 25 % of Group B staff said that the vending machine had adequate food choices. Less than a third of Group A staff were satisfied, giving a total for both groups of less than 30 %.
9. About 65 % (100 % − 35 %) of Group B staff said that the prices were too high and were not satisfied with the prices. For example, one member of staff noted that a Caesar salad was $12, which feels like airport prices. However, slightly more than half of Group A staff (55 %) indicated satisfaction with the current prices, giving a total satisfaction for both groups of just under 50 %.
10. Both groups of staff thought that the location of the vending machine was acceptable. Only 29 % of the staff wanted a better position for the vending machine.
11. 95 % were satisfied with the reliability, meaning when they put their money in they got the correct product, and if appropriate they received the correct change.
12. The findings show that staff members, especially staff in Group B, were not satisfied with prices and food selections. We should therefore consider buying a larger vending machine or another separate vending machine so there are more selections at broader range of pricing.
13. In addition to the bistro vending machine. The company should consider providing a microwave and a blending station. This would allow the staff to prepare their own lunches with food from home.
14. A frequency analysis of pricing within the current machine indicates that most (56 %) of items are priced above $10.00. (See graph that follows).

Graph 1



15. A second more detailed survey should be undertaken to better understand the acceptable price points and food preferences for this group of employees.

Discussion Boards

1. Reports need to avoid quantitative terms as most clients don't really understand statistics.
2. Sample size can make or break your final report.
3. Executive reports should not be more than two pages.

Group Activity

Find a report online for the financial investment industry, the wine industry and the Environmental Protection Agency (EPA). What are some of the commonalities and differences in how these reports are organized? Consider the mission and motivations of these groups. Do these factors influence what is included in the reports and how it is presented? Give specific examples.

Parting Thought

Statistics means never having to say you're certain.

Problem Solutions

1. The errors and areas for improvement include:
    - Missing Title Page
    - Missing Table of Contents; but since the report is so short, one is not really necessary
    - Complete date missing in header
    - Reports need separate headings for each section
    - Executive Summary missing
    - Graph not numbered
    - Frequencies in Graph 1 don't add up to 100 %
    - No sample size provided in Table 1
    - Categories overlap in Graph 1
    - Page Numbers missing
    - No copy of the actual survey attached. What are the questions? Did all questions get answered or were some confusing?
    - Was this anonymous/confidential? If so, make sure you mention this
    - No discussion of the average data in Table 1
    - Survey source and development details missing
    - Statistical significance discussion missing

The corrected report follows:

Tiga Consulting
Lunch Room Report for EIGC
June 11 20–

Consultant: Dr. Paula Simms
888-745-4444
psimms@gazelle.nz

To: S.M. Riesling, General Manager
From: Paula Simms
Date: June 11/20–
Subject: Lunch Room Vending Machine

Executive Summary

Employees are not satisfied with the current lunchroom situation. This survey was conducted to explore details on the reasons for this dissatisfaction. The lunchroom has been outfitted with a new bistro style vending machine, as the only source of purchasing food in the area. Outside the actual office is an industrial area with no available restaurants. Staff are not as dissatisfied with the location of the machine and its reliability. The issues primarily focus on the price of food and the limited food choices.

The recommendations we suggest based on this exploratory survey include buying an additional vending machine with lower priced but healthy foods. To better understand what the employees want, a follow-up survey is recommended that specifically addresses price points and food choices.

Background and Objectives

At the monthly staff meeting on May 21 20–, you requested our company provide a report about staff satisfaction with the new high end vending machine. The aim of this report is to present this information with recommendations.

Since the move to the new office location in Brownsville, employees have had difficulty in finding a nearby place to buy lunch. The new office is in a factory area and restaurants are limited. A new state of the art bistro vending machine was purchased. However some employees expressed dissatisfaction with this vending machine.

Research Methodology

All 122 employees were surveyed by an online questionnaire on May 21 and May 22/ 20–. The resulting response rate was 89 %, representing 110 completed surveys. The survey included 28 questions with a Likert scale that went from 1 to 5, with 5 being the most satisfied and 1 being the most dissatisfied. The survey was developed by Tiga Consulting based on initial interviews with EIGC employees. The survey is included in Appendix 1.

The data were analyzed in Excel using the frequency analysis tool to show the results graphically. There were no outliers in the data that required deletion.

Overview of Results

In general, staff are not satisfied with the prices and food choices in the vending machine. A minority was dissatisfied with the location of the vending machine, but most were satisfied with its reliability.

On analyzing the data, two distinct groups of staff emerged. The first includes 68 staff who had usually lunched in the office when the office was in West Grove (Group A). The second group of 40 staff had usually lunched outside in West Grove (Group B).

Table 1.0 Staff satisfaction with the vending machine

	Staff satisfaction (n = 108)		
Feature	Group A Usually lunched in the office (n = 68)	Group B Usually lunched outside the office (n = 40)	Both groups
Price	55 %	35 %	48 %
Food choices	30 %	25 %	28 %
Location	70 %	75 %	71 %
Reliability	95 %	95 %	95 %
Overall avg	62.5 %	57.5 %	60.6 %

Twenty-five percent of Group B staff said that the vending machine had adequate food choices. Less than a third of Group A staff were satisfied with the food choices, giving a total for both groups of less than 30 %.

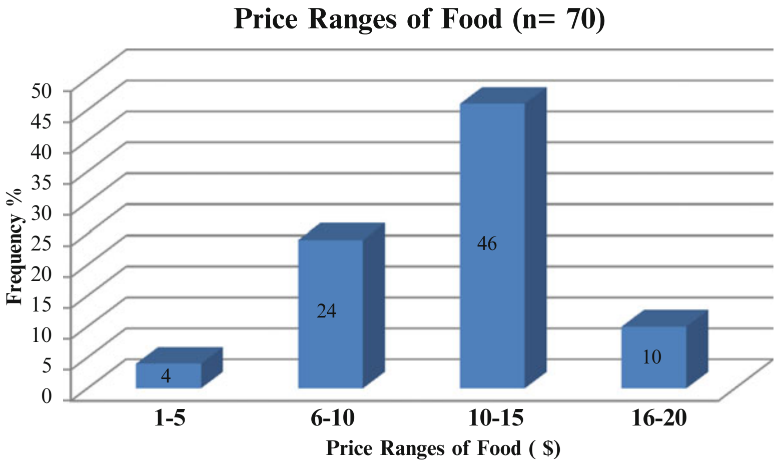
About 65 % of Group B staff said that the prices were too high. For example, one member of staff noted that a Caesar salad was $12, which feels like airport prices. However, slightly more than half of Group A staff indicated satisfaction with the current prices, giving a total satisfaction for both groups of just under 50 %.

Both groups of staff thought that the location of the vending machine was acceptable. Only 29 % of the staff wanted a better position for the vending machine.

Ninety-five percent were satisfied with the reliability, meaning when they put their money in they got the correct product, and if appropriate they received the correct change.

The findings show that staff members, especially staff in Group B, were not satisfied with prices and food selections.

A frequency analysis of pricing within the current machine indicates that most (56 %) of items are priced above $10.00.



Graph 1.0 Price ranges of food

Conclusions and Recommendations

We should therefore consider buying a larger vending machine or another separate vending machine so there are more selections with a broader range of pricing.

A second more detailed survey should be undertaken to better understand the acceptable price points and food preferences for this group of employees.

In addition to the bistro vending machine, the company should consider providing a microwave and a blending station. This would allow the staff to prepare their own lunches with food from home. Perhaps the microwave and blending machines may be more appealing than the vending machines. Future surveys should address how demand for vending machines may decline if other equipment such as microwave and blending machines were installed. These results may impact the decision to purchase a second vending machine or even to retain the existing vending machine.

Bibliography

Freedman, D., Pisani, R., Purves, R., 1998. Statistics (third edition). Norton and Company, NY, USA.

Nelson, D., Quick, J., 2006. Organizational Behavior (fifth edition). Thomson/ Southwestern, Ohio, USA

Appendix 1

Include the actual survey questions here

Index

L. Herkenhoff and J. Fogli, *Applied Statistics for Business and Management using Microsoft Excel*, DOI 10.1007/978-1-4614-8423-3,