

Excel for Statistics

Thomas J. Quirk

Excel 2016 in Applied Statistics for High School Students

A Guide to Solving Practical Problems



Springer

Excel for Statistics

Excel for Statistics is a series of textbooks that explain how to use Excel to solve statistics problems in various fields of study. Professors, students, and practitioners will find these books teach how to make Excel work best in their respective fields. Applications include any discipline that uses data and can benefit from the power and simplicity of Excel. Books cover all the steps for running statistical analyses in Excel 2016, Excel 2013, Excel 2010, and Excel 2007. The approach also teaches critical statistics skills, making the books particularly applicable for statistics courses taught outside of mathematics or statistics departments.

Series editor: Thomas J. Quirk

The following books are in this series:

T.J. Quirk, *Excel 2016 in Applied Statistics for High School Students: A Guide to Solving Practical Problems*, Excel for Statistics. Springer international Publishing Switzerland 2018.

T.J. Quirk, E. Rhiney, *Excel 2016 for Advertising Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2017.

T.J. Quirk, S. Cummings, *Excel 2016 for Social Work Statistics: A Guide to Solving Practical Problems*. Excel for statistics. Springer International Publishing Switzerland 2017.

T.J. Quirk, E. Rhiney, *Excel 2016 for Marketing Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk, *Excel 2016 for Business Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk. *Excel 2016 for Engineering Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk, M. Quirk, H.F. Horton, *Excel 2016 for Biological and Life Sciences Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk. *Excel 2016 for Educational and Psychological Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk, *Excel 2016 for Social Science Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk, M. Quirk, H. Horton, *Excel 2016 for Physical Sciences Statistics: A Guide to Solving Practical Problems*. Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk, S. Cummings, *Excel 2016 for Health Services Management Statistics: A Guide to Solving Practical Problems*. Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk, J. Palmer-Schuyler, *Excel 2016 for Human Resource Management Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk, M. Quirk, H.F. Horton. *Excel 2016 for Environmental Sciences Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk, M. Quirk, H.F. Horton. *Excel 2013 for Physical Sciences Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk, S. Cummings, *Excel 2013 for Health Services Management Statistics: A Guide to Solving Practical Problems*. Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk, J. Palmer-Schuyler, *Excel 2013 for Human Resource Management Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2016.

T.J. Quirk, *Excel 2013 for Business Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2015.

T.J. Quirk. *Excel 2013 for Engineering Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2015.

T.J. Quirk, M. Quirk, H.F. Horton, *Excel 2013 for Biological and Life Sciences Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2015.

T.J. Quirk. *Excel 2013 for Educational and Psychological Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2015.

T.J. Quirk, *Excel 2013 for Social Science Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2015.

T.J. Quirk, M. Quirk, H.F. Horton, *Excel 2013 for Environmental Sciences Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2015.

T.J. Quirk, M. Quirk, H.F. Horton, *Excel 2010 for Environmental Sciences Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2015.

T.J. Quirk, J. Palmer-Schuyler, *Excel 2010 for Human Resource Management Statistics: A Guide to Solving Practical Problems*, Excel for Statistics. Springer International Publishing Switzerland 2014.

Additional Statistics books by Dr. Tom Quirk that have been published by Springer

T.J. Quirk, *Excel 2010 for Business Statistics: A Guide to Solving Practical Problems*. Springer Science +Business Media 2011.

T.J. Quirk. *Excel 2010 for Engineering Statistics: A Guide to Solving Practical Problems*. Springer International Publishing Switzerland 2014.

T.J. Quirk, S. Cummings, *Excel 2010 for Health Services Management Statistics: A Guide to Solving Practical Problems*. Springer International Publishing Switzerland 2014.

T.J. Quirk, M. Quirk, H. Horton, *Excel 2010 for Physical Sciences Statistics: A Guide to Solving Practical Problems*. Springer International Publishing Switzerland 2013.

T.J. Quirk, M. Quirk, H.F. Horton, *Excel 2010 for Biological and Life Sciences Statistics: A Guide to Solving Practical Problems*. Springer Science+Business Media New York 2013.

T.J. Quirk, *Excel 2010 for Social Science Statistics: A Guide to Solving Practical Problems*. Springer Science+Business Media New York 2012.

T.J. Quirk, *Excel 2010 for Educational and Psychological Statistics: A Guide to Solving Practical Problems*. Springer Science+Business Media New York 2012.

Thomas J. Quirk

Excel 2016 in Applied Statistics for High School Students

A Guide to Solving Practical Problems

 Springer

Thomas J. Quirk
Webster University
St. Louis, MO, USA

Excel for Statistics

ISBN 978-3-319-89992-3

ISBN 978-3-319-89993-0 (eBook)

<https://doi.org/10.1007/978-3-319-89993-0>

Library of Congress Control Number: 2018940021

© Springer International Publishing AG, part of Springer Nature 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

This book is dedicated to the more than three thousand students I have taught at Webster University's campuses in St. Louis, London, and Vienna; the students at Principia College in Elsau, Illinois; and the students at the Cooperative State University of Baden-Wuerttemberg in Heidenheim, Germany. These students taught me a great deal about the art of teaching. I salute them all, and I thank them for helping me to become a better teacher.

Thomas J. Quirk

Preface

Excel 2016 in Applied Statistics for High School Students: A Guide to Solving Practical Problems helps anyone who wants to learn the basics of applying Excel's powerful statistical tools to their classes. If understanding statistics isn't your strongest suit, you are not mathematically inclined, or you are wary of computers, then this is the book for you.

You'll learn how to perform key statistical tests in Excel without being overwhelmed by statistical theory. This book clearly and logically shows how to run statistical tests to solve practical problems in several fields of study.

Excel is a widely available computer program for students and teachers. It is also an effective teaching and learning tool for quantitative analyses in statistics courses. Its powerful computational ability and graphical functions make learning statistics much easier than in years past. However, this is the first book to showcase Excel's usefulness in teaching statistics. And it focuses exclusively on this topic in order to render the subject matter applicable and practical—and easy to comprehend and apply.

Unique features of this book:

- Includes 166 color screenshots so you can be sure you are performing Excel steps correctly.
- You will be told each step of the way, not only *how* to use Excel, but also *why* you are doing each step.
- Includes specific objectives embedded in the text for each concept, so you can know the purpose of the Excel steps.
- You will learn both how to write statistical formulas using Excel and how to use Excel's drop-down menus that will create the formulas for you.
- Statistical theory and formulas are explained in clear language without bogging you down in mathematical fine points.
- Practical examples of problems are taken from several fields of study.

- Each chapter presents key steps to solve practical problems using Excel. In addition, three practice problems at the end of each chapter enable you to test your new knowledge. Answers to these problems appear in Appendix A.
- A “Practice Test” is given in Appendix B to test your knowledge at the end of the book. Answers to this test appear in Appendix C.
- This book does not come with a CD of Excel files which you can upload to your computer. Instead, you’ll be shown how to create each Excel file yourself. In a classroom situation, your teachers will not give you an Excel file. You will be expected to create your own. This book will give you ample practice in developing this important skill.
- This book is a tool that can be used either by itself or along with *any* good statistics book.

This book is appropriate for use in any statistics course—as well as for teachers/administrators who want to improve their Excel skills.

At the beginning of his academic career, Prof. Quirk spent six years in educational research at the American Institutes for Research and Educational Testing Service. He then taught social psychology, educational psychology, general psychology, accounting, management, and marketing at Principia College and is currently a Professor of Marketing in the George Herbert Walker School of Business & Technology at Webster University based in St. Louis, Missouri (USA), where he teaches marketing statistics, marketing research, and pricing strategies. He has published articles in the *Journal of Educational Psychology*, *Journal of Educational Research*, *Review of Educational Research*, *Journal of Educational Measurement*, *Educational Technology*, *The Elementary School Journal*, *Journal of Secondary Education*, *Educational Horizons*, and *Phi Delta Kappan*. In addition, he has published 20+ articles in professional journals and presented 20+ papers at professional meetings, including annual meetings of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. He holds a B.S. in mathematics from John Carroll University, both an M.A. in education and a Ph.D. in educational psychology from Stanford University, and an M.B.A. from the University of Missouri-St. Louis.

St. Louis, MO, USA

Thomas J. Quirk

Acknowledgments

Excel 2016 in Applied Statistics for High School Students: A Guide to Solving Practical Problems is the result of inspiration from three important people: my two daughters and my wife. Jennifer Quirk McLaughlin invited me to visit her M.B.A. classes several times at the University of Witwatersrand in Johannesburg, South Africa. These visits to a first-rate M.B.A. program convinced me there was a need for a book to teach students how to solve practical problems using Excel. Meghan Quirk-Horton's dogged dedication to learning the many statistical techniques needed to complete her Ph.D. dissertation illustrated the need for a statistics book that would make this daunting task more user-friendly. And Lynne Buckley-Quirk was the number-one cheerleader for this project from the beginning, always encouraging me and helping me remain dedicated to completing it.

Thomas J. Quirk

Contents

1	Sample Size, Mean, Standard Deviation, and Standard Error of the Mean	1
1.1	Mean	1
1.2	Standard Deviation	2
1.3	Standard Error of the Mean	3
1.4	Sample Size, Mean, Standard Deviation, and Standard Error of the Mean	4
1.4.1	Using the Fill/Series/Columns Commands	4
1.4.2	Changing the Width of a Column	5
1.4.3	Centering Information in a Range of Cells	6
1.4.4	Naming a Range of Cells	8
1.4.5	Finding the Sample Size Using the =COUNT Function	9
1.4.6	Finding the Mean Score Using the =AVERAGE Function	9
1.4.7	Finding the Standard Deviation Using the =STDEV Function	10
1.4.8	Finding the Standard Error of the Mean	10
1.5	Saving a Spreadsheet	12
1.6	Printing a Spreadsheet	13
1.7	Formatting Numbers in Currency Format (Two Decimal Places)	15
1.8	Formatting Numbers in Number Format (Three Decimal Places)	17
1.9	End-of-Chapter Practice Problems	18
	Reference	22
2	Random Number Generator	23
2.1	Creating Frame Numbers for Generating Random Numbers	23
2.2	Creating Random Numbers in an Excel Worksheet	27
2.3	Sorting Frame Numbers into a Random Sequence	28

- 2.4 Printing an Excel File So That All of the Information Fits Onto One Page 32
- 2.5 End-of-Chapter Practice Problems 35
- Reference 36
- 3 Confidence Interval About the Mean Using the TINV Function and Hypothesis Testing 37**
 - 3.1 Confidence Interval About the Mean 37
 - 3.1.1 How to Estimate the Population Mean 37
 - 3.1.2 Estimating the Lower Limit and the Upper Limit of the 95% Confidence Interval About the Mean 38
 - 3.1.3 Estimating the Confidence Interval for the Chevy Impala in Miles Per Gallon 39
 - 3.1.4 Where Did the Number “1.96” Come From? 40
 - 3.1.5 Finding the Value for t in the Confidence Interval Formula 41
 - 3.1.6 Using Excel’s TINV Function to Find the Confidence Interval About the Mean 42
 - 3.1.7 Using Excel to Find the 95% Confidence Interval for a Car’s mpg Claim 42
 - 3.2 Hypothesis Testing 47
 - 3.2.1 Hypotheses Always Refer to the Population of People or Events That You Are Studying 48
 - 3.2.2 The Null Hypothesis and the Research (Alternative) Hypothesis 49
 - 3.2.3 The 7 Steps for Hypothesis-Testing Using the Confidence Interval About the Mean 52
 - 3.3 Alternative Ways to Summarize the Result of a Hypothesis Test 58
 - 3.3.1 Different Ways to Accept the Null Hypothesis 58
 - 3.3.2 Different Ways to Reject the Null Hypothesis 59
 - 3.4 End-of-Chapter Practice Problems 59
 - References 64
- 4 One-Group t-Test for the Mean 65**
 - 4.1 The 7 STEPS for Hypothesis-Testing Using the One-Group t-Test 65
 - 4.1.1 STEP 1: State the Null Hypothesis and the Research Hypothesis 66
 - 4.1.2 STEP 2: Select the Appropriate Statistical Test 66
 - 4.1.3 STEP 3: Decide on a Decision Rule for the One-Group t-Test 66
 - 4.1.4 STEP 4: Calculate the Formula for the One-Group t-Test 67

- 4.1.5 STEP 5: Find the Critical Value of t in the t-Table in Appendix E 68
- 4.1.6 STEP 6: State the Result of Your Statistical Test 69
- 4.1.7 STEP 7: State the Conclusion of Your Statistical Test in Plain English! 69
- 4.2 One-Group t-Test for the Mean 70
- 4.3 Can You Use Either the 95% Confidence Interval About the Mean OR the One-Group t-Test When Testing Hypotheses? 75
- 4.4 End-of-Chapter Practice Problems 75
- References 79
- 5 Two-Group t-Test of the Difference of the Means for Independent Groups 81**
 - 5.1 The 9 STEPS for Hypothesis-Testing Using the Two-Group t-Test 82
 - 5.1.1 STEP 1: Name One Group, Group 1, and the Other Group, Group 2 82
 - 5.1.2 STEP 2: Create a Table That Summarizes the Sample Size, Mean Score, and Standard Deviation of Each Group 82
 - 5.1.3 STEP 3: State the Null Hypothesis and the Research Hypothesis for the Two-Group t-Test 84
 - 5.1.4 STEP 4: Select the Appropriate Statistical Test 84
 - 5.1.5 STEP 5: Decide on a Decision Rule for the Two-Group t-Test 84
 - 5.1.6 STEP 6: Calculate the Formula for the Two-Group t-Test 84
 - 5.1.7 STEP 7: Find the Critical Value of t in the t-Table in Appendix E 85
 - 5.1.8 STEP 8: State the Result of Your Statistical Test 86
 - 5.1.9 STEP 9: State the Conclusion of Your Statistical Test in Plain English! 86
 - 5.2 Formula #1: Both Groups Have More Than 30 People in Them 90
 - 5.2.1 An example of Formula #1 for the Two-Group t-Test 91
 - 5.3 Formula #2: One or Both Groups Have Less Than 30 Events in Them 98
 - 5.4 End-of-Chapter Practice Problems 104
 - References 106
- 6 Correlation and Simple Linear Regression 107**
 - 6.1 What Is a “Correlation?” 107
 - 6.1.1 Understanding the Formula for Computing a Correlation 112
 - 6.1.2 Understanding the Nine Steps for Computing a Correlation, r 112

- 6.2 Using Excel to Compute a Correlation Between Two Variables 114
- 6.3 Creating a Chart and Drawing the Regression Line onto the Chart 119
 - 6.3.1 Using Excel to Create a Chart and the Regression Line Through the Data Points 121
- 6.4 Printing a Spreadsheet So That the Table and Chart Fit onto One Page 129
- 6.5 Finding the Regression Equation 131
 - 6.5.1 Installing the Data Analysis ToolPak into Excel 132
 - 6.5.2 Using Excel to Find the SUMMARY OUTPUT of Regression 135
 - 6.5.3 Finding the Equation for the Regression Line 140
 - 6.5.4 Using the Regression Line to Predict the y-Value for a Given x-Value 140
- 6.6 Adding the Regression Equation to the Chart 141
- 6.7 How to Recognize Negative Correlations in the SUMMARY OUTPUT Table 144
- 6.8 Printing Only Part of a Spreadsheet Instead of the Entire Spreadsheet 144
 - 6.8.1 Printing Only the Table and the Chart on a Separate Page 145
 - 6.8.2 Printing Only the Chart on a Separate Page 145
 - 6.8.3 Printing Only the SUMMARY OUTPUT of the Regression Analysis on a Separate Page 146
- 6.9 End-of-Chapter Practice Problems 146
- References 152
- 7 Multiple Correlation and Multiple Regression 153**
 - 7.1 Multiple Regression Equation 153
 - 7.2 Finding the Multiple Correlation and the Multiple Regression Equation 156
 - 7.3 Using the Regression Equation to Predict FROSH GPA 160
 - 7.4 Using Excel to Create a Correlation Matrix in Multiple Regression 160
 - 7.5 End-of-Chapter Practice Problems 163
 - References 168
- 8 One-Way Analysis of Variance (ANOVA) 169**
 - 8.1 Using Excel to Perform a One-Way Analysis of Variance (ANOVA) 170
 - 8.2 How to Interpret the ANOVA Table Correctly 174
 - 8.3 Using the Decision Rule for the ANOVA F-Test 174

8.4	Testing the Difference Between Two Groups Using the ANOVA t-Test	175
8.4.1	Comparing Brand A vs. Brand C in Miles Driven Using the ANOVA t-Test	176
8.5	End-of-Chapter Practice Problems	180
	References	185
	Appendices	187
	Appendix A: Answers to End-of-Chapter Practice Problems	187
	Appendix B: Practice Test	219
	Appendix C: Answers to Practice Test	230
	Appendix D: Statistical Formulas	240
	Appendix E: t-Table	242
	Index	243

Chapter 1

Sample Size, Mean, Standard Deviation, and Standard Error of the Mean



This chapter deals with how you can use Excel to find the average (i.e., “mean”) of a set of scores, the standard deviation of these scores (STDEV), and the standard error of the mean (s.e.) of these scores. All three of these statistics are used frequently and form the basis for additional statistical tests.

1.1 Mean

The *mean* is the “arithmetic average” of a set of scores. When my daughter was in the fifth grade, she came home from school with a sad face and said that she didn’t get “averages.” The book she was using described how to find the mean of a set of scores, and so I said to her:

“Jennifer, you add up all the scores and divide by the number of numbers that you have.”
She gave me “that look,” and said: “Dad, this is serious!” She thought I was teasing her.
So I said:
“See these numbers in your book; add them up. What is the answer?” (She did that.)
“Now, how many numbers do you have?” (She answered that question.)
“Then, take the number you got when you added up the numbers, and divide that number by the number of numbers that you have.”

She did that, and found the correct answer. You will use that same reasoning now, but it will be much easier for you because Excel will do all of the steps for you.

We will call this average of the scores the “mean” which we will symbolize as: \bar{X} , and we will pronounce it as: “Xbar.”

The formula for finding the mean with your calculator looks like this:

$$\bar{X} = \frac{\sum X}{n} \tag{1.1}$$

The symbol Σ is the Greek letter sigma, which stands for “sum.” It tells you to add up all the scores that are indicated by the letter X , and then to divide your answer by n (the number of numbers that you have).

Let’s give a simple example:

Suppose that you had these six test scores on an 7-item true-false quiz:

6
4
5
3
2
5

To find the mean of these scores, you add them up, and then divide by the number of scores. So, the mean is: $25/6 = 4.17$

1.2 Standard Deviation

The *standard deviation* tells you “how close the scores are to the mean.” If the standard deviation is a small number, this tells you that the scores are “bunched together” close to the mean. If the standard deviation is a large number, this tells you that the scores are “spread out” a greater distance from the mean. The formula for the standard deviation (which we will call *STDEV*) and use the letter, S , to symbolize is:

$$STDEV = S = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} \quad (1.2)$$

The formula look complicated, but what it asks you to do is this:

1. Subtract the mean from each score ($X - \bar{X}$).
2. Then, square the resulting number to make it a positive number.
3. Then, add up these squared numbers to get a total score.
4. Then, take this total score and divide it by $n - 1$ (where n stands for the number of numbers that you have).
5. The final step is to take the square root of the number you found in step 4.

You will not be asked to compute the standard deviation using your calculator in this book, but you could see examples of how it is computed in any basic statistics book. Instead, we will use Excel to find the standard deviation of a set of scores. When we use Excel on the six numbers we gave in the description of the mean above, you will find that the *STDEV* of these numbers, S , is 1.47.

1.3 Standard Error of the Mean

The formula for the *standard error of the mean* (*s.e.*, which we will use $S_{\bar{x}}$ to symbolize) is:

$$\text{s.e.} = S_{\bar{x}} = \frac{S}{\sqrt{n}} \quad (1.3)$$

To find *s.e.*, all you need to do is to take the standard deviation, STDEV, and divide it by the square root of *n*, where *n* stands for the “number of numbers” that you have in your data set. In the example under the standard deviation description above, the *s.e.* = 0.60. (You can check this on your calculator.)

If you want to learn more about the standard deviation and the standard error of the mean, see Agresti and Franklin (2013).

Now, let’s learn how to use Excel to find the sample size, the mean, the standard deviation, and the standard error of the mean using a geometry test given to a class of eight 9th graders at the end of the first term of the school year (50 points possible). The hypothetical data appear in Fig. 1.1.

Fig. 1.1 Worksheet Data for a Geometry Test (Practical Example)

	A	B	C
1			
2			
3	Student	Geometry Test Score	
4	1	10	
5	2	10	
6	3	12	
7	4	16	
8	5	22	
9	6	29	
10	7	39	
11	8	47	
12			

1.4 Sample Size, Mean, Standard Deviation, and Standard Error of the Mean

Objective: To find the sample size (n), mean, standard deviation (STDEV), and standard error of the mean (s.e.) for these data

Start your computer, and click on the Excel 2016 icon to open a blank Excel spreadsheet.

Click on: Blank Workbook

Enter the data in this way:

A3: Student

B3: Geometry Test Score

A4: 1

1.4.1 Using the Fill/Series/Columns Commands

Objective: To add the student numbers 2–8 in a column underneath student #1

Put pointer in A4

Home (top left of screen)

Important note: The “Paste” command should be on the top of your screen on the far left of the screen.

Important note: Notice the Excel commands at the top of your computer screen:

File → **Home** → **Insert** → **Page Layout** → **Formulas** etc.

If these commands ever “disappear” when you are using Excel, you need to click on “Home” at the top left of your screen to make them reappear!

Fill (top right of screen: click on the down arrow; see Fig. 1.2)

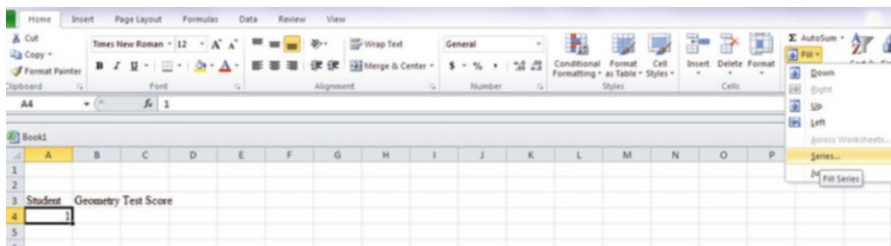


Fig. 1.2 Home/Fill/Series commands

Series

Columns

Step value: 1

Stop value: 8 (see Fig. 1.3)

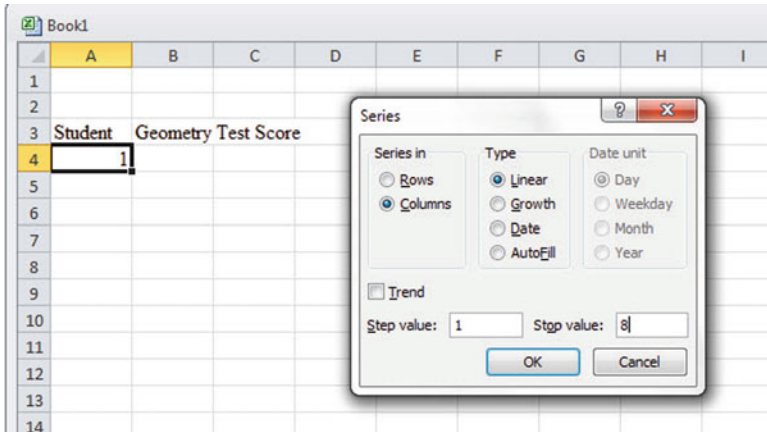


Fig. 1.3 Example of Dialogue Box for Fill/Series/Columns/Step Value/Stop Value commands

OK

The student numbers should be identified as 1–8, with 8 in cell A11.

Now, enter the Geometry Test Scores in cells B4: B11.

Since your computer screen shows the information in a format that does not look professional, you need to learn how to “widen the column width” and how to “center the information” in a group of cells. Here is how you can do those two steps:

1.4.2 Changing the Width of a Column

Objective: To make a column width wider so that all of the information fits inside that column

If you look at your computer screen, you can see that Column B is not wide enough so that all of the information fits inside this column. To make Column B wider:

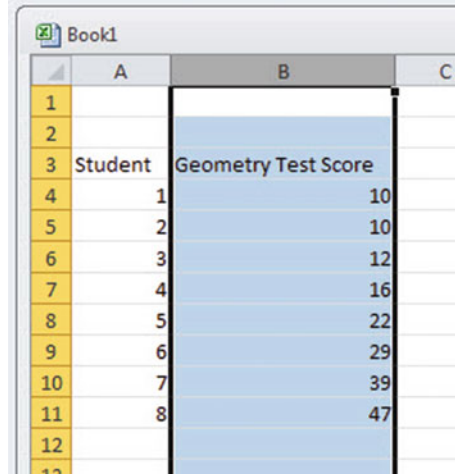
Click on the letter, B, at the top of your computer screen

Place your mouse pointer at the far right corner of B until you create a “cross sign” on that corner

Left-click on your mouse, hold it down, and move this corner to the right until it is “wide enough to fit all of the data”

Take your finger off the mouse to set the new column width (see Fig. 1.4)

Fig. 1.4 Example of How to Widen the Column Width



Then, click on any empty cell (i.e., any blank cell) to “deselect” column B so that it is no longer a darker color on your screen.

When you widen a column, you will make all of the cells in all of the rows of this column that same width.

Now, let’s go through the steps to center the information in both Column A and Column B.

1.4.3 Centering Information in a Range of Cells

Objective: To center the information in a group of cells

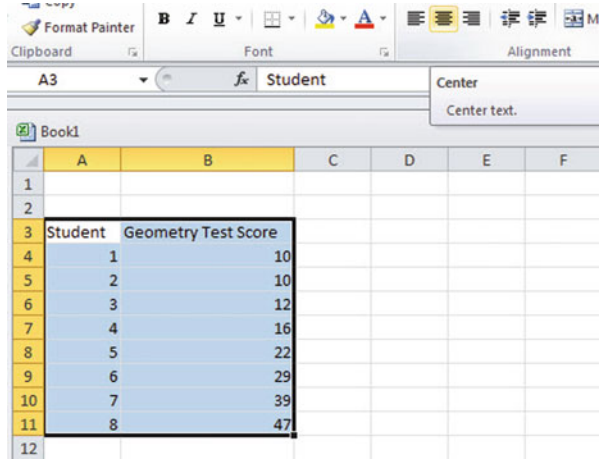
In order to make the information in the cells look “more professional,” you can center the information using the following steps:

Left-click your mouse on A3 and drag it to the right and down to highlight cells A3:B11 so that these cells appear in a darker color

Home

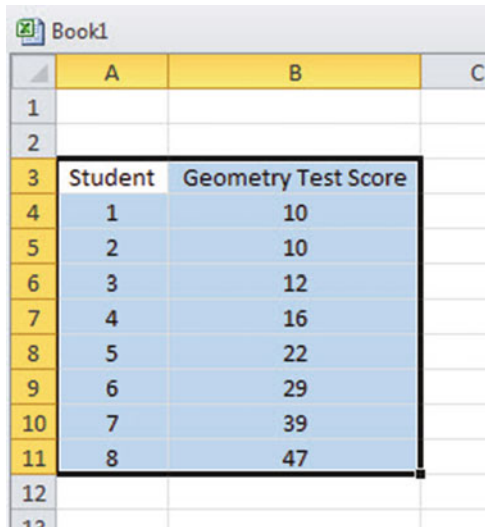
At the top of your computer screen, you will see a set of “lines” in which all of the lines are “centered” to the same width under “Alignment” (it is the second icon at the bottom left of the Alignment box; see Fig. 1.5)

Fig. 1.5 Example of How to Center Information Within Cells



Click on this icon to center the information in the selected cells (see Fig. 1.6)

Fig. 1.6 Final Result of Centering Information in the Cells



Since you will need to refer to the Geometry Test Scores in your formulas, it will be much easier to do this if you “name the range of data” with a name instead of having to remember the exact cells (B4:B11) in which these figures are located. Let’s call that group of cells: Geometry, but we could give them any name that you want to use.

1.4.4 Naming a Range of Cells

Objective: To name the range of data for the test scores with the name: Geometry

Highlight cells B4: B11 by left-clicking your mouse on B4 and dragging it down to B11

Formulas (top left of your screen)

Define Name (top center of your screen)

Geometry (type this name in the top box; see Fig. 1.7)

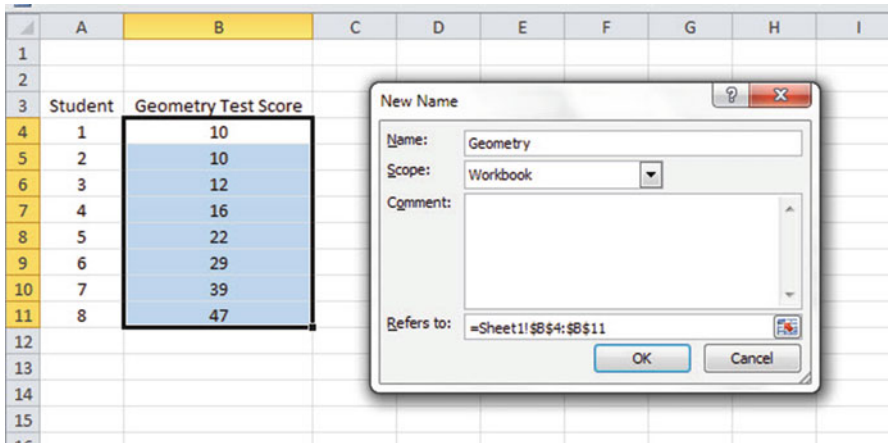


Fig. 1.7 Dialogue box for “naming a range of cells” with the name: Geometry

OK

Then, click on any cell of your spreadsheet that does not have any information in it (i.e., it is an “empty cell”) to deselect cells B4:B11

Now, add the following terms to your spreadsheet:

E6: n

E9: Mean

E12: STDEV

E15: s.e. (see Fig. 1.8)

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F
1						
2						
3	Student	Geometry Test Score				
4	1	10				
5	2	10				
6	3	12			n	
7	4	16				
8	5	22			Mean	
9	6	29				
10	7	39				
11	8	47			STDEV	
12						
13						
14						
15					s.e.	
16						
17						

Fig. 1.8 Example of Entering the Sample Size, Mean, STDEV, and s.e. Labels

Note: Whenever you use a formula, you must add an equal sign (=) at the beginning of the name of the function so that Excel knows that you intend to use a formula.

1.4.5 Finding the Sample Size Using the =COUNT Function

Objective: To find the sample size (n) for these data using the =COUNT function

F6: =COUNT(Geometry)

This command should insert the number 8 into cell F6 since there are eight students in this class.

1.4.6 Finding the Mean Score Using the =AVERAGE Function

Objective: To find the mean test score figure using the =AVERAGE function

F9: =AVERAGE(Geometry)

This command should insert the number 23.125 into cell F9.

1.4.7 Finding the Standard Deviation Using the =STDEV Function

Objective: To find the standard deviation (STDEV) using the =STDEV function

F12: =STDEV(Geometry)

This command should insert the number 14.02485 into cell F12.

1.4.8 Finding the Standard Error of the Mean

Objective: To find the standard error of the mean using a formula for these eight data points

F15: =F12/SQRT(8)

This command should insert the number 4.958533 into cell F15 (see Fig. 1.9).

The screenshot shows an Excel spreadsheet with the following data and calculations:

	A	B	C	D	E	F	G
1							
2							
3	Student	Geometry Test Score					
4	1	10					
5	2	10					
6	3	12			n	8	
7	4	16					
8	5	22					
9	6	29			Mean	23.125	
10	7	39					
11	8	47					
12					STDEV	14.02485	
13							
14							
15					s.e.	4.958533	
16							
17							

Fig. 1.9 Example of Using Excel Formulas for Sample Size, Mean, STDEV, and s.e.

Important note: Throughout this book, be sure to double-check all of the figures in your spreadsheet to make sure that they are in the correct cells, or the formulas will not work correctly!

1.4.8.1 Formatting Numbers in Number Format (Two Decimal Places)

Objective: To convert the mean, STDEV, and s.e. to two decimal places

Highlight cells F9:F15

Home (top left of screen)

Look under “Number” at the top center of your screen. In the bottom right corner, gently place your mouse pointer on you screen at the bottom of the .00 .0 until it says: “Decrease Decimal” (see Fig. 1.10)

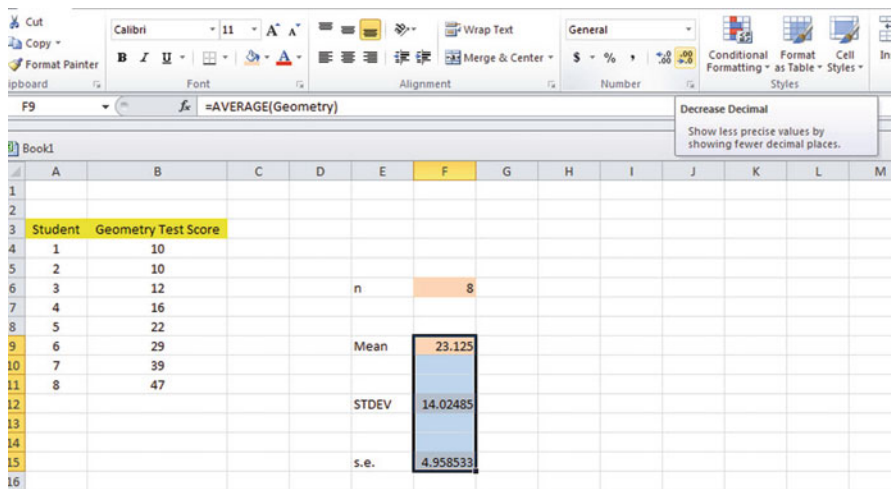


Fig. 1.10 Using the “Decrease Decimal Icon” to convert Numbers to Fewer Decimal Places

Click on this icon *once* and notice that the cells F9:F15 are now all in just two decimal places (see Fig. 1.11)

	A	B	C	D	E	F	G
1							
2							
3	Student	Geometry Test Score					
4	1	10					
5	2	10					
6	3	12			n	8	
7	4	16					
8	5	22					
9	6	29			Mean	23.13	
10	7	39					
11	8	47					
12					STDEV	14.02	
13							
14							
15					s.e.	4.96	
16							
17							

Fig. 1.11 Example of Converting Numbers to Two Decimal Places

Now, click on any “empty cell” on your spreadsheet to deselect cells F9:F15.

1.5 Saving a Spreadsheet

Objective: To save this spreadsheet with the name: Geometry3

In order to save your spreadsheet so that you can retrieve it sometime in the future, your first decision is to decide “where” you want to save it. That is your decision and you have several choices. If it is your own computer, you can save it onto your hard drive (you need to ask someone how to do that on your computer). Or, you can save it onto a “CD” or onto a “flash drive.” You then need to complete these steps:

File

Save as

(select the place where you want to save the file by scrolling either down or up the bar on the left, and click on the place where you want to save the file; for example: This PC: My Documents location)

File name: Geometry3 (enter this name to the right of File name; see Fig. 1.12)

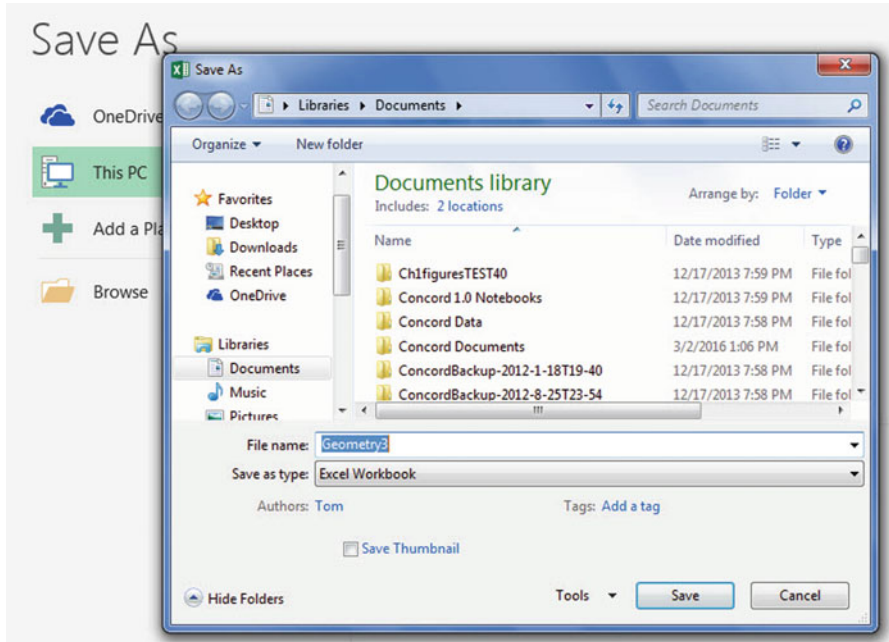


Fig. 1.12 Dialogue Box of Saving an Excel Workbook File as “Geometry3” in My Documents location

Save

Important note: Be very careful to save your Excel file spreadsheet every few minutes so that you do not lose your information!

1.6 Printing a Spreadsheet

Objective: To print the spreadsheet

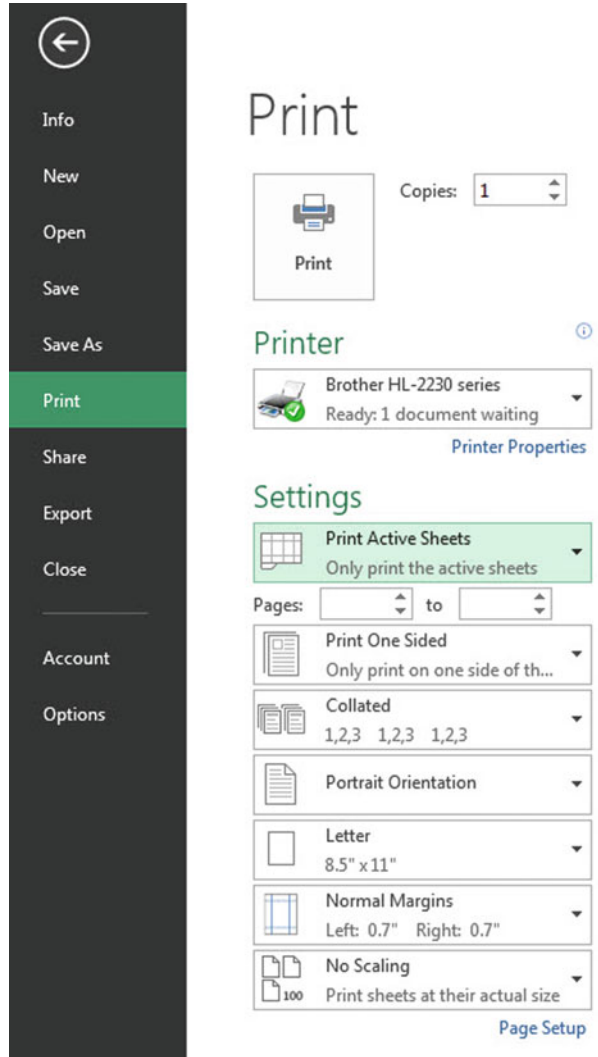
Use the following procedure when printing any spreadsheet.

File

Print

Print Active Sheets (see Fig. 1.13)

Fig. 1.13 Example of How to Print an Excel Worksheet Using the File/Print/Print Active Sheets Commands



Print (top of your screen)

The final spreadsheet is given in Fig 1.14

	A	B	C	D	E	F	G
1							
2							
3	Student	Geometry Test Score					
4	1	10					
5	2	10					
6	3	12			n	8	
7	4	16					
8	5	22					
9	6	29			Mean	23.13	
10	7	39					
11	8	47					
12					STDEV	14.02	
13							
14							
15					s.e.	4.96	
16							
17							

Fig. 1.14 Final Result of Printing an Excel Spreadsheet

Before you leave this chapter, let’s practice changing the format of the figures on a spreadsheet with two examples: (1) using two decimal places for figures that are dollar amounts, and (2) using three decimal places for figures.

Close your spreadsheet by: File/Save, then close your spreadsheet by: File/Close/, and open a blank Excel spreadsheet by using:

File/New/Blank Workbook (on the top left of your screen).

1.7 Formatting Numbers in Currency Format (Two Decimal Places)

Objective: To change the format of figures to dollar format with two decimal places

- A3: Price
- A4: 1.25
- A5: 3.45
- A6: 12.95

Highlight cells A4:A6 by left-clicking your mouse on A4 and dragging it down so that these three cells are highlighted in a darker color

Home

Number (top center of screen: click on the down arrow on the right; see Fig. 1.15)

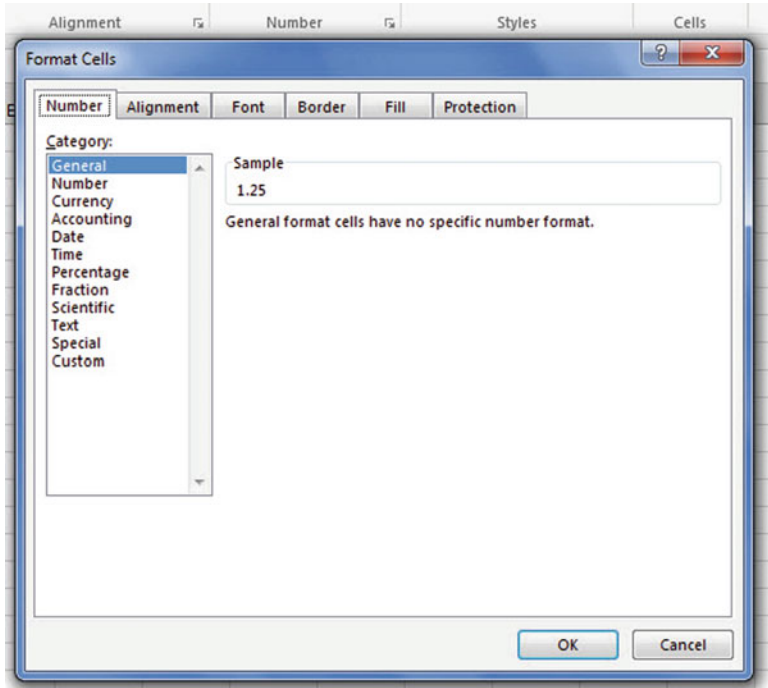


Fig. 1.15 Dialogue Box for Number Format Choices

Category: Currency
Decimal places: 2 (then see Fig. 1.16)

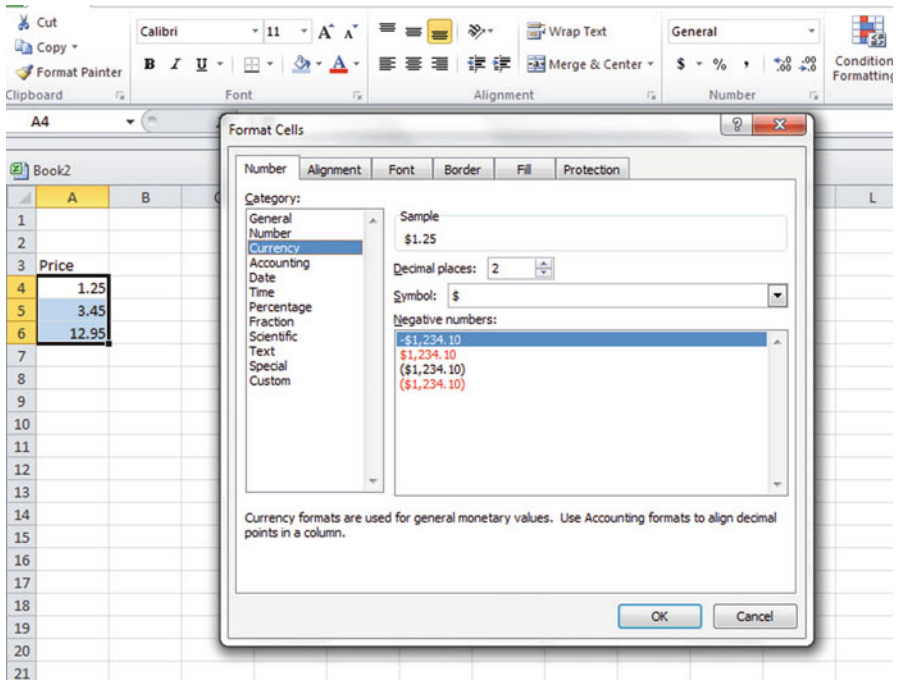


Fig. 1.16 Dialogue Box for Currency (two decimal places) Format for Numbers

OK

The three cells should have a dollar sign in them and be in two decimal places. Next, let's practice formatting figures in number format, three decimal places.

1.8 Formatting Numbers in Number Format (Three Decimal Places)

Objective: To format figures in number format, three decimal places

Home

Highlight cells A4:A6 on your computer screen

Number (click on the down arrow on the right)

Category: number

At the right of the box, change two decimal places to three decimal places by clicking on the "up arrow" once

OK

The three figures should now be in number format, each with three decimals.

Now, click on any blank cell to deselect cells A4:A6. Then, close this file by File/Close/Don't Save (since there is no need to save this practice problem).

You can use these same commands to format a range of cells in percentage format (and many other formats) to whatever number of decimal places you want to specify.

1.9 End-of-Chapter Practice Problems

1. Suppose that you work for an advertising firm that does research about potential television commercials by having members of a panel view the commercials and comment on how effective the commercials are in encouraging them to purchase the product that is described in the ad. Note that a “panel” is a group of people who have agreed to participate in research studies over the Web. There are different panels for different target market segments. Suppose that you have been asked to analyze the data for a possible TV ad for a new product based on the survey responses of male college students (ages 18–24) in the panel. The survey has ten items in it, but you have decided to create some hypothetical data for just Item #10 which asks about purchase intent based on the TV ad. These hypothetical data appear in Fig. 1.17.

TV ADVERTISING PILOT TEST							
Panel of male college students (ages 18-24)							
Item #10:	Based on the TV commercial that you just saw, how likely are you to purchase the advertised product?						
	1	2	3	4	5	6	7
	Very Unlikely						Very Likely
		RATING					
		3					
		4					
		2					
		6					
		3					
		5					
		4					
		3					
		6					
		2					
		1					
		2					
		1					
		3					
		4					
		3					
		2					
		4					
		1					
		2					
		3					

Fig. 1.17 Worksheet Data for Chap. 1: Practice Problem #1

- (a) Use Excel to the right of the table to find the sample size, mean, standard deviation, and standard error of the mean for these data. Label your answers, and round off the mean, standard deviation, and standard error of the mean to two decimal places; use number format for these three figures.
 - (b) Print the result on a separate page.
 - (c) Save the file as: TVad4
2. Suppose you work for Ford Motor Company and you have been asked to do data analysis for a panel of female college students (ages 18–24) to determine their importance of possible features for a new Ford Focus automobile that is on the drawing board. You want to test your Excel skills and so you have created a table of hypothetical data for Item #12 from the survey. These data are given in Fig. 1.18.

FORD MOTOR COMPANY							
Survey of new-car features							
Panel of female college students (ages 18-24)							
Question #12:	If you were to purchase a new car today, how important to you is the feature that "the car parallel parks itself to the curb" by using a computer?						
	1	2	3	4	5	6	7
	Not Important						Very Important
		<u>RATING</u>					
		5					
		6					
		4					
		3					
		7					
		6					
		5					
		7					
		6					
		7					
		4					
		3					
		1					
		7					
		6					
		4					
		5					

Fig. 1.18 Worksheet Data for Chap. 1: Practice Problem #2

- (a) Use Excel to create a table of these ratings, and at the right of the table use Excel to find the sample size, mean, standard deviation, and standard error of the mean for these data. Label your answers, and round off the mean, standard deviation, and standard error of the mean to three decimal places using number format.
 - (b) Print the result on a separate page.
 - (c) Save the file as: CAR12A
3. A “spokesperson” is someone who participates in an advertisement (TV, Radio, Internet, Direct Mail, Magazines, etc.) who tries to convince you that the product/service that this spokesperson is representing is one that you should purchase. In order for the spokesperson to be effective, the target market of potential customers needs to believe that the spokesperson is trustworthy, informative, sincere, qualified, and so forth, in these and in other dimensions so that target market finds this spokesperson credible in his or her claims about the product/service.

An important part of advertising research involving spokespersons is to study the effectiveness of the spokesperson on the target market’s perceptions of both the spokesperson and what the ad is saying about the product/service.

One type of measurement scale that is frequently used in this type of advertising research is called the semantic differential scale. This type of scale is frequently described as containing bipolar adjectives as the end-points of a 7-point rating scale. For example, the target market might be asked to describe what part of the rating scale the spokesperson fell on items such as trustworthy/untrustworthy, informative/uninformative, sincere/insincere, qualified/unqualified, and so forth.

Suppose that you have been asked to analyze the data from a target market that consisted of female undergraduate college students (ages 18–24) who were shown a possible television ad for a new type of cosmetic product that is being developed by a major manufacturer of these types of products. Before diving into the data, you want to try your Excel skills on a set of hypothetical data for Item #12 of this survey to make sure that you can do this type of analysis correctly. The data appear in Fig. 1.19.

TELEVISION SPOKESPERSON RATINGS BY FEMALE COLLEGE STUDENTS						
COMMERCIAL No. 512						
Item #12:	Think about the spokesperson you just saw in the commercial. How trustworthy do you think this spokesperson is in terms of the product that was advertised?					
7	6	5	4	3	2	1
Very Trustworthy						Very Untrustworthy
		RATING				
		3				
		4				
		6				
		2				
		1				
		3				
		5				
		1				
		7				
		6				
		5				
		3				
		4				
		2				
		6				
		5				
		4				

Fig. 1.19 Worksheet Data for Chap. 1: Practice Problem #3

- (a) Use Excel to create a table for these data, and at the right of the table, use Excel to find the sample size, mean, standard deviation, and standard error of the mean for these data. Label your answers, and round off the mean, standard deviation, and standard error of the mean to two decimal places using number format.
- (b) Print the result on a separate page.
- (c) Save the file as: Trust23

Reference

Agresti, A.A. and Franklin, C.F. Statistics: The Art and Science of Learning from Data (3rd ed.) Boston, MA: Pearson Education, Inc. 2013.

Chapter 2

Random Number Generator



Suppose that a local school superintendent asked you to take a random sample of 5 of an elementary school’s 32 teachers using Excel so that you could interview these five teachers about their job satisfaction at their school.

To do that, you need to define a “sampling frame.” A sampling frame is a list of people from which you want to select a random sample. This frame starts with the identification code (ID) of the number 1 that is assigned to the name of the first teacher in your list of 32 teachers in this school. The second teacher has a code number of 2, the third a code number of 3, and so forth until the last teacher has a code number of 32.

Since this school has 32 teachers, your sampling frame would go from 1 to 32 with each teacher having a unique ID number.

We will first create the frame numbers as follows in a new Excel worksheet:

2.1 Creating Frame Numbers for Generating Random Numbers

Objective: To create the frame numbers for generating random numbers

A3: FRAME NO.

A4: 1

Now, create the frame numbers in column A with the Home/Fill commands that were explained in the first chapter of this book (see Sect. 1.4.1) so that the frame numbers go from 1 to 32, with the number 32 in cell A35. If you need to be reminded about how to do that, here are the steps:

Click on cell A4 to select this cell

Home

Fill (then click on the “down arrow” next to this command and select)

Series (see Fig. 2.1)

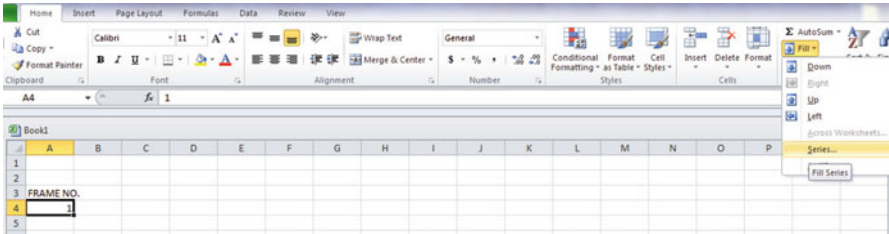


Fig. 2.1 Dialogue Box for Fill/Series Commands

Columns

Step value: 1

Stop value: 32 (see Fig. 2.2)

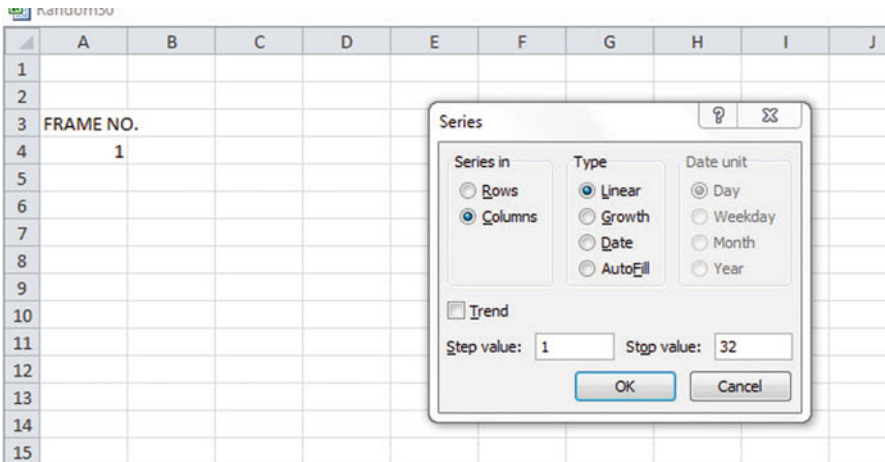


Fig. 2.2 Dialogue Box for Fill/Series/Columns/Step value/Stop value Commands

OK

Then, save this file as: Random29. You should obtain the result in Fig. 2.3.

Fig. 2.3 Frame Numbers from 1 to 32

FRAME NO.
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

Now, create a column next to these frame numbers in this manner:

B3: DUPLICATE FRAME NO.

B4: 1

Next, use the Home/Fill command again, so that the 32 frame numbers begin in cell B4 and end in cell B35. Be sure to widen the columns A and B so that all of the information in these columns fits inside the column width. Then, center the information inside both Column A and Column B on your spreadsheet. You should obtain the information given in Fig. 2.4.

Fig. 2.4 Duplicate Frame Numbers from 1 to 32

FRAME NO.	DUPLICATE FRAME NO.
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13
14	14
15	15
16	16
17	17
18	18
19	19
20	20
21	21
22	22
23	23
24	24
25	25
26	26
27	27
28	28
29	29
30	30
31	31
32	32

Save this file as: Random30

You are probably wondering why you created the same information in both Column A and Column B of your spreadsheet. This is to make sure that before you sort the frame numbers that you have exactly 32 of them when you finish sorting them into a random sequence of 32 numbers.

Now, let's add a random number to each of the duplicate frame numbers as follows:

2.2 Creating Random Numbers in an Excel Worksheet

C3: RANDOM NO.

(then widen columns A, B, C so that their labels fit inside the columns; then center the information in A3:C35)

C4: =RAND()

Next, hit the Enter key to add a random number to cell C4.

Note that you need *both* an open parenthesis *and* a closed parenthesis after =RAND(). The RAND command “looks to the left of the cell with the RAND() COMMAND in it” and assigns a random number to that cell.

Now, put the pointer using your mouse in cell C4 and then move the pointer to the bottom right corner of that cell until you see a “plus sign” in that cell. Then, click and drag the pointer down to cell C35 to add a random number to all 32 ID frame numbers (see Fig. 2.5).

Fig. 2.5 Example of Random Numbers Assigned to the Duplicate Frame Numbers

FRAME NO.	DUPLICATE FRAME NO.	RANDOM NO.
1	1	0.003582027
2	2	0.782788863
3	3	0.579542097
4	4	0.194826007
5	5	0.180624088
6	6	0.11111051
7	7	0.841557731
8	8	0.43839362
9	9	0.039665164
10	10	0.950687104
11	11	0.701382569
12	12	0.340485937
13	13	0.622362199
14	14	0.823691903
15	15	0.196989157
16	16	0.644490495
17	17	0.818952316
18	18	0.198204913
19	19	0.703849331
20	20	0.412769208
21	21	0.356822645
22	22	0.906835718
23	23	0.770448533
24	24	0.22095226
25	25	0.141239005
26	26	0.035501923
27	27	0.11336016
28	28	0.055916598
29	29	0.368006445
30	30	0.133117174
31	31	0.736884083
32	32	0.702218141

Then, click on any empty cell to deselect C4:C35 to remove the dark color highlighting these cells.

Save this file as: Random31

Now, let's sort these duplicate frame numbers into a random sequence:

2.3 Sorting Frame Numbers into a Random Sequence

Objective: To sort the duplicate frame numbers into a random sequence

Highlight cells B3:C35 (include the labels at the top of columns B and C)

Data (top of screen)

Sort (click on this word at the top center of your screen; see Fig. 2.6)

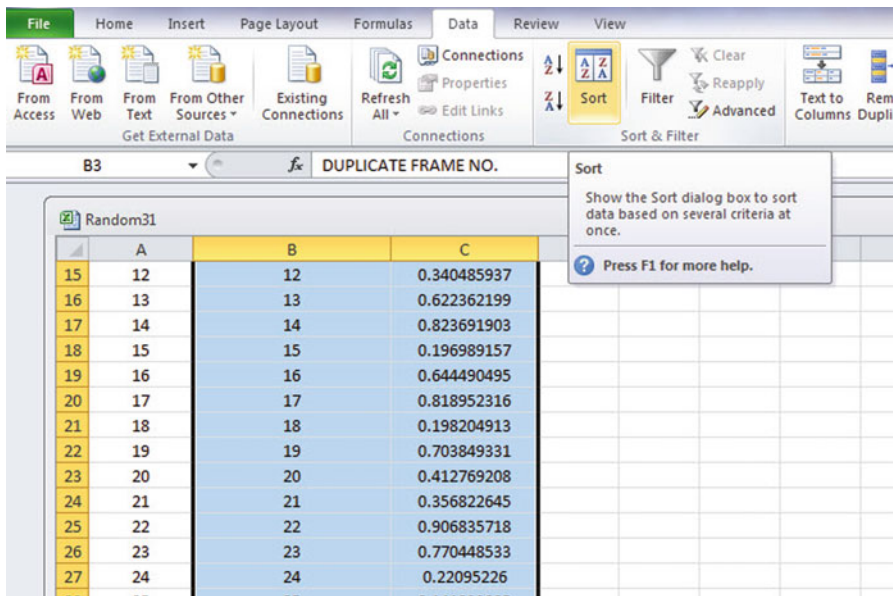


Fig. 2.6 Dialogue Box for Data/Sort Commands

Sort by: RANDOM NO. (click on the down arrow)

Smallest to Largest (see Fig. 2.7)

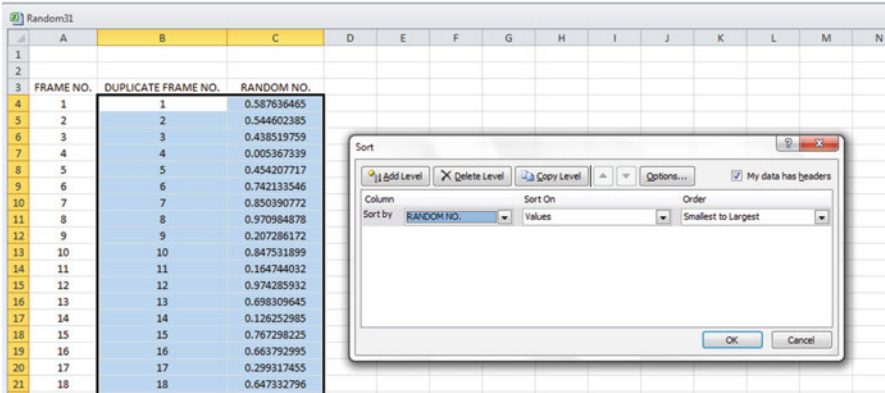


Fig. 2.7 Dialogue Box for Data/Sort/RANDOM NO./Smallest to Largest Commands

OK

Click on any empty cell to deselect B3:C35.

Save this file as: Random32

Print this file now.

These steps will produce Fig. 2.8 with the DUPLICATE FRAME NUMBERS sorted into a random order:

Fig. 2.8 Duplicate Frame Numbers Sorted by Random Number

FRAME NO.	DUPLICATE FRAME NO.	RANDOM NO.
1	4	0.048919918
2	32	0.715166245
3	19	0.734859307
4	31	0.457996527
5	14	0.837332575
6	25	0.509257077
7	11	0.71473365
8	20	0.297513233
9	29	0.455511676
10	9	0.398344533
11	28	0.074506919
12	17	0.822659637
13	26	0.680326909
14	30	0.586949154
15	3	0.685958841
16	5	0.801702294
17	24	0.042432294
18	27	0.913192281
19	2	0.911877184
20	1	0.000107945
21	22	0.660085877
22	18	0.782601369
23	16	0.222790879
24	13	0.735556843
25	6	0.849257376
26	15	0.232958712
27	10	0.506254928
28	7	0.092256678
29	23	0.990635679
30	21	0.767847232
31	8	0.004497343
32	12	0.748968302

Important note: Because Excel randomly assigns these random numbers, your Excel commands will produce a different sequence of random numbers from everyone else who reads this book!

Because your objective at the beginning of this chapter was to select randomly 5 of this school's 32 teachers for a personal interview, you now can do that by selecting the *first five ID numbers* in DUPLICATE FRAME NO. column after the sort.

Although your first five random numbers will be different from those we have selected in the random sort that we did in this chapter, we would select these five IDs of teachers to interview using Fig. 2.9.

4, 32, 19, 31, 14

Fig. 2.9 First Five Teachers Selected Randomly

FRAME NO.	DUPLICATE FRAME NO.	RANDOM NO.
1	4	0.048919918
2	32	0.715166245
3	19	0.734859307
4	31	0.457996527
5	14	0.837332575
6	25	0.509257077
7	11	0.71473365
8	20	0.297513233
9	29	0.455511676
10	9	0.398344533
11	28	0.074506919
12	17	0.822659637
13	26	0.680326909
14	30	0.586949154
15	3	0.685958841
16	5	0.801702294
17	24	0.042432294
18	27	0.913192281
19	2	0.911877184
20	1	0.000107945
21	22	0.660085877
22	18	0.782601369
23	16	0.222790879
24	13	0.735556843
25	6	0.849257376
26	15	0.232958712
27	10	0.506254928
28	7	0.092256678
29	23	0.990635679
30	21	0.767847232
31	8	0.004497343
32	12	0.748968302

Save this file as: Random33

Remember, your five ID numbers selected after your random sort will be different from the five ID numbers in Fig. 2.9 because Excel assigns a different random number *each time the =RAND() command is given.*

Before we leave this chapter, you need to learn how to print a file so that all of the information on that file fits onto a single page without “dribbling over” onto a second or third page.

2.4 Printing an Excel File So That All of the Information Fits Onto One Page

Objective: To print a file so that all of the information fits onto one page

Note that the three practice problems at the end of this chapter require you to sort random numbers when the files contain 63 Honda dealers, 114 counties of the state of Missouri, and 76 key accounts, respectively. These files will be “too big” to fit onto one page when you print them unless you format these files so that they fit onto a single page when you print them.

Let’s create a situation where the file does not fit onto one printed page unless you format it first to do that.

Go back to the file you just created, Random 33, and enter the name: *Jennifer* into cell: A50.

If you printed this file now, the name, *Jennifer*, would be printed onto a second page because it “dribbles over” outside of the page range for this file in its current format.

So, you would need to change the page format so that all of the information, including the name, Jennifer, fits onto just one page when you print this file by using the following steps:

Page Layout (top left of the computer screen)

(Notice the “Scale to Fit” section in the center of your screen; see Fig. 2.10)

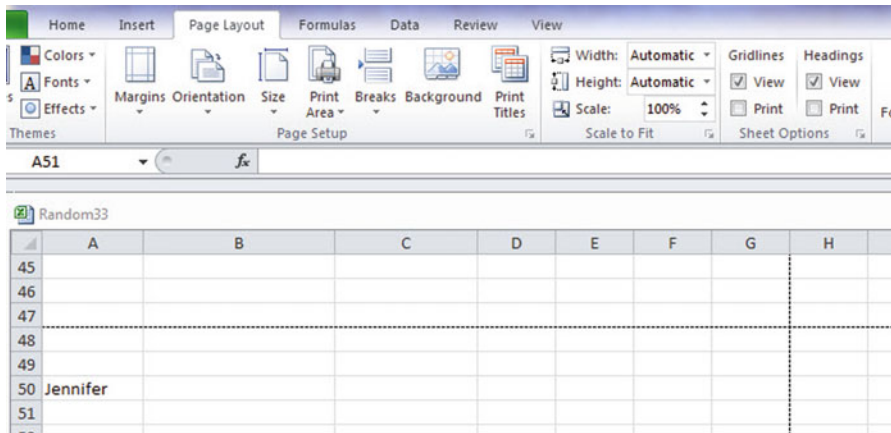


Fig. 2.10 Dialogue Box for Page Layout/Scale to Fit Commands

Hit the down arrow to the right of 100% *once* to reduce the size of the page to 95%

Now, note that the name, Jennifer, is still on a second page on your screen because her name is below the horizontal dotted line on your screen in Fig. 2.11 (the dotted lines tell you outline dimensions of the file if you printed it now).

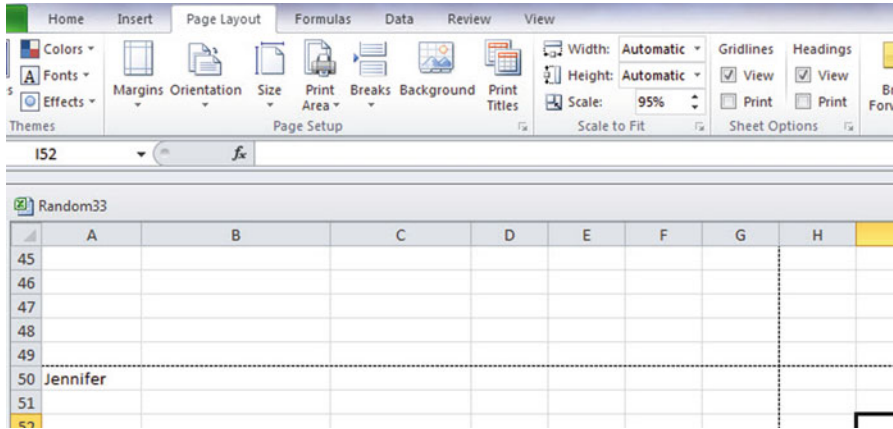


Fig. 2.11 Example of Scale Reduced to 95% with “Jennifer” to be Printed on a Second Page

So, you need to repeat the “scale change steps” by hitting the down arrow on the right once more to reduce the size of the worksheet to 90% of its normal size.

Notice that the “dotted lines” on your computer screen in Fig. 2.12 are now below Jennifer’s name to indicate that all of the information, including her name, is now formatted to fit onto just one page when you print this file.

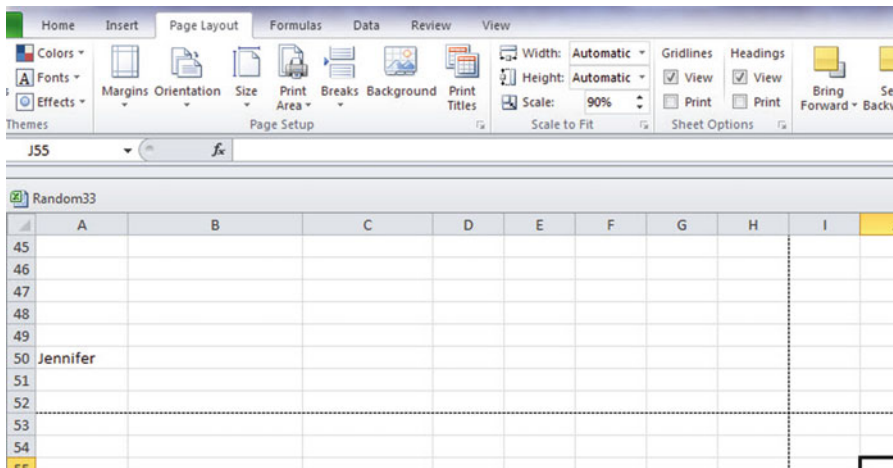


Fig. 2.12 Example of Scale Reduced to 90% with “Jennifer” to be printed on the first page (note the dotted line below Jennifer on your screen)

Save the file as: Random34

Print the file. Does it all fit onto one page? It should (see Fig. 2.13).

2.5 End-of-Chapter Practice Problems

1. Suppose that your advertising agency represents 63 Honda dealers in your state and that you have been asked to perform a “dealer satisfaction phone survey” of 15 of 63 car dealers to obtain their ideas about how your agency can advertise Hondas more effectively.
 - (a) Set up a spreadsheet of frame numbers for these dealers with the heading: FRAME NUMBERS using the Home/Fill commands.
 - (b) Then, create a separate column to the right of these frame numbers which duplicates these frame numbers with the title: Duplicate frame numbers
 - (c) Then, create a separate column to the right of these duplicate frame numbers and use the =RAND() function to assign random numbers to all of the frame numbers in the duplicate frame numbers column, and change this column format so that three decimal places appear for each random number
 - (d) Sort the duplicate frame numbers and random numbers into a random order
 - (e) Print the result so that the spreadsheet fits onto one page
 - (f) Circle on your printout the I.D. number of the first 15 dealers that you would call in your phone survey
 - (g) Save the file as: RAND9

Important note: Note that everyone who does this problem will generate a different random order of dealer ID numbers since Excel assigns a different random number each time the RAND() command is used. For this reason, the answer to this problem given in this Excel Guide will have a completely different sequence of random numbers from the random sequence that you generate. This is normal and what is to be expected.

2. Suppose that you wanted to do a random sample of 10 of the 114 counties in the state of Missouri as requested by a political pollster who wants to select registered voters by county in Missouri for a phone survey of their voting preferences in the next election. You know that there are 114 counties in Missouri because you have accessed the Web site for the U.S. census (U.S. Census Bureau, 2000). For your information, the United States has a total of 3140 counties in its 50 states (U.S. Census Bureau 2000).
 - (a) Set up a spreadsheet of frame numbers for these counties with the heading: FRAME NO.
 - (b) Then, create a separate column to the right of these frame numbers which duplicates these frame numbers with the title: Duplicate frame no.
 - (c) Then, create a separate column to the right of these duplicate frame numbers entitled “Random number” and use the =RAND() function to assign random numbers to all of the frame numbers in the duplicate frame numbers column. Then, change this column format so that three decimal places appear for each random number

- (d) Sort the duplicate frame numbers and random numbers into a random order
 - (e) Print the result so that the spreadsheet fits onto one page
 - (f) Circle on your printout the I.D. number of the first ten counties that the pollster would call in his phone survey
 - (g) Save the file as: RANDOM6
3. Suppose that a Sales department at a company wants to do a “customer satisfaction survey” of 20 of this company’s 76 “key accounts.” Suppose, further, that the Sales Vice-President has defined a key account as a customer who purchased at least \$30,000 worth of merchandise from this company in the past 90 days.
- (a) Set up a spreadsheet of frame numbers for these customers with the heading: FRAME NUMBERS.
 - (b) Then, create a separate column to the right of these frame numbers which duplicates these frame numbers with the title: Duplicate frame numbers
 - (c) Then, create a separate column to the right of these duplicate frame numbers entitled “Random number” and use the =RAND() function to assign random numbers to all of the frame numbers in the duplicate frame numbers column. Then, change this column format so that three decimal places appear for each random number
 - (d) Sort the duplicate frame numbers and random numbers into a random order
 - (e) Print the result so that the spreadsheet fits onto one page
 - (f) Circle on your printout the I.D. number of the first 20 customers that the Sales Vice-President would call for his phone survey.
 - (g) Save the file as: RAND5

Reference

U.S. Census Bureau Census 2000 PHC-T-4. Ranking tables for counties 1990 and 2000. Retrieved from <http://www.census.gov/population/www/cen2000/briefs/phc-t4/tables/tab01.pdf>

Chapter 3

Confidence Interval About the Mean Using the TINV Function and Hypothesis Testing



This chapter focuses on two ideas: (1) finding the 95% confidence interval about the mean, and (2) hypothesis testing.

Let's talk about the confidence interval first.

3.1 Confidence Interval About the Mean

In statistics, we are always interested in *estimating the population mean*. How do we do that?

3.1.1 How to Estimate the Population Mean

Objective: To estimate the population mean, μ

Remember that the population mean is the average of all of the people in the target population. For example, if we were interested in how well adults ages 25–44 liked a new flavor of Ben & Jerry's ice cream, we could never ask this question of all of the people in the U.S. who were in that age group. Such a research study would take way too much time to complete and the cost of doing that study would be prohibitive.

So, instead of testing *everyone* in the population, we take a sample of people in the population and use the results of this sample to estimate the mean of the entire population. This saves both time and money. When we use the results of a sample to estimate the population mean, this is called "*inferential statistics*" because we are inferring the population mean from the sample mean.

When we study a sample of people in business research, we know the size of our sample (n), the mean of our sample (\bar{X}), and the standard deviation of our sample (STDEV). We use these figures to estimate the population mean with a test called the “confidence interval about the mean.”

3.1.2 *Estimating the Lower Limit and the Upper Limit of the 95% Confidence Interval About the Mean*

The theoretical background of this test is beyond the scope of this book, and you can learn more about this test from studying any good statistics textbook (e.g. Levine 2011) but the basic ideas are as follows.

We assume that the population mean is somewhere in an interval which has a “lower limit” and an “upper limit” to it. We also assume in this book that we want to be “95% confident” that the population mean is inside this interval somewhere. So, we intend to make the following type of statement:

“We are 95% confident that the population mean in miles per gallon (mpg) for the Chevy Impala automobile is between 26.92 miles per gallon and 29.42 miles per gallon.”

If we want to create a billboard for this car that claims that this car gets 28 miles per gallon (mpg), we can do that because 28 is *inside the 95% confidence interval* in our research study in the above example. We do not know exactly what the population mean is, only that it is somewhere between 26.92 and 29.42 mpg, and 28 is inside this interval.

But we are only 95% confident that the population mean is inside this interval, and 5% of the time we will be wrong in assuming that the population mean is 28 mpg.

But, for our purposes in business research, we are happy to be 95% confident that our assumption is accurate. We should also point out that 95% is an arbitrary level of confidence for our results. We could choose to be 80% confident, or 90% confident, or even 99% confident in our results if we wanted to do that. But, in this book, *we will always assume that we want to be 95% confident of our results*. That way, you will not have to guess on how confident you want to be in any of the problems in this book. We will always want to be 95% confident of our results in this book.

So how do we find the 95% confidence interval about the mean for our data?

In words, we will find this interval this way:

“Take the sample mean (\bar{X}), *and add to it* 1.96 times the standard error of the mean (s.e.) to get the upper limit of the confidence interval. Then, take the sample mean, *and subtract from it* 1.96 times the standard error of the mean to get the lower limit of the confidence interval.”

You will remember (See Sect. 1.3) that the standard error of the mean (s.e.) is found by dividing the standard deviation of our sample (STDEV) by the square root of our sample size, n .

In mathematical terms, the formula for the 95% confidence interval about the mean is:

$$\bar{X} \pm 1.96 \text{ s.e.} \quad (3.1)$$

Note that the “ \pm sign” stands for “plus or minus,” and this means that you first add 1.96 times the s.e. to the mean to get the upper limit of the confidence interval, and then subtract 1.96 times the s.e. from the mean to get the lower limit of the confidence interval. Also, the symbol 1.96 s.e. means that you multiply 1.96 times the standard error of the mean to get this part of the formula for the confidence interval.

Note: We will explain shortly where the number 1.96 came from.

Let’s try a simple example to illustrate this formula.

3.1.3 Estimating the Confidence Interval for the Chevy Impala in Miles Per Gallon

Let’s suppose that you asked owners of the Chevy Impala to keep track of their mileage and the number of gallons used for two tanks of gas. Let’s suppose that 49 owners did this, and that they average 27.83 miles per gallon (mpg) with a standard deviation of 3.01 mpg. The standard error (s.e.) would be 3.01 divided by the square root of 49 (i.e., 7) which gives a s.e. equal to 0.43.

The 95% confidence interval for these data would be:

$$27.83 \pm 1.96 (0.43)$$

The *upper limit of this confidence interval* uses the plus sign of the \pm sign in the formula. Therefore, the upper limit would be:

$$27.83 + 1.96(0.43) = 27.83 + 0.84 = 28.67 \text{ mpg}$$

Similarly, the *lower limit of this confidence interval* uses the minus sign of the \pm sign in the formula. Therefore, the lower limit would be:

$$27.83 - 1.96(0.43) = 27.83 - 0.84 = 26.99 \text{ mpg}$$

The result of our research study would, therefore, be the following:

“We are 95% confident that the population mean for the Chevy Impala is somewhere between 26.99 mpg and 28.67 mpg.”

If we were planning to create a billboard that claimed that this car got 28 mpg, we would be able to do that based on our data, since 28 is inside of this 95% confidence interval for the population mean.

You are probably asking yourself: “Where did that 1.96 in the formula come from?”

3.1.4 Where Did the Number “1.96” Come From?

A detailed mathematical answer to that question is beyond the scope of this book, but here is the basic idea.

We make an assumption that the data in the population are “normally distributed” in the sense that the population data would take the shape of a “normal curve” if we could test all of the people in the population. The normal curve looks like the outline of the Liberty Bell that sits in front of Independence Hall in Philadelphia, Pennsylvania. The normal curve is “symmetric” in the sense that if we cut it down the middle, and folded it over to one side, the half that we folded over would fit perfectly onto the half on the other side.

A discussion of integral calculus is beyond the scope of this book, but essentially we want to find the lower limit and the upper limit of the population data in the normal curve so that 95% of the area under this curve is between these two limits. *If we have more than 40 people in our research study*, the value of these limits is plus or minus 1.96 times the standard error of the mean (s.e.) of our sample. The number 1.96 times the s.e. of our sample gives us the upper limit and the lower limit of our confidence interval. If you want to learn more about this idea, you can consult a good statistics book (e.g. Salkind 2010).

The number 1.96 would change if we wanted to be confident of our results at a different level from 95% as long as we have more than 40 people in our research study.

For example:

1. If we wanted to be 80% confident of our results, this number would be 1.282.
2. If we wanted to be 90% confident of our results, this number would be 1.645.
3. If we wanted to be 99% confident of our results, this number would be 2.576.

But since we always want to be 95% confident of our results in this book, we will always use 1.96 in this book whenever we have more than 40 people in our research study.

By now, you are probably asking yourself: “Is this number in the confidence interval about the mean always 1.96?” The answer is: “No!”, and we will explain why this is true now.

3.1.5 Finding the Value for t in the Confidence Interval Formula

Objective: To find the value for t in the confidence interval formula

The correct formula for the confidence interval about the mean for different sample sizes is the following:

$$\bar{X} \pm t \text{ s.e.} \quad (3.2)$$

To use this formula, you find the sample mean, \bar{X} , and add to it the value of t times the s.e. to get the upper limit of this 95% confidence interval. Also, you take the sample mean, \bar{X} , and subtract from it the value of t times the s.e. to get the lower limit of this 95% confidence interval. And, you find the value of t in the table given in Appendix E of this book in the following way:

Objective: To find the value of t in the t-table in Appendix E

Before we get into an explanation of what is meant by “the value of t ,” let’s give you practice in finding the value of t by using the t-table in Appendix E.

Keep your finger on Appendix E as we explain how you need to “read” that table.

Since the test in this chapter is called the “confidence interval about the mean test,” you will use the first column on the left in Appendix E to find the critical value of t for your research study (note that this column is headed: “sample size n ”).

To find the value of t , you go down this first column until you find the sample size in your research study, and then you go to the right and read the value of t for that sample size in the “critical t column” of the table (note that this column is the column that you would use for the 95% confidence interval about the mean).

For example, if you have 14 people in your research study, the value of t is 2.160.

If you have 26 people in your research study, the value of t is 2.060.

If you have more than 40 people in your research study, the value of t is always 1.96.

Note that the “critical t column” in Appendix E represents the value of t that you need to use to obtain to be 95% confident of your results as “significant” results.

Throughout this book, we are assuming that you want to be 95% confident in the results of your statistical tests. Therefore, the value for t in the t-table in Appendix E tells you which value you should use for t when you use the formula for the 95% confidence interval about the mean.

Now that you know how to find the value of t in the formula for the confidence interval about the mean, let’s explore how you find this confidence interval using Excel.

3.1.6 Using Excel's TINV Function to Find the Confidence Interval About the Mean

Objective: To use the TINV function in Excel to find the confidence interval about the mean

When you use Excel, the formulas for finding the confidence interval are:

$$\text{Lower limit:} = \bar{X} - TINV(1 - 0.95, n - 1) * s.e. \text{ (no spaces between these symbols)} \quad (3.3)$$

$$\text{Upper limit:} = \bar{X} + TINV(1 - 0.95, n - 1) * s.e. \text{ (no spaces between these symbols)} \quad (3.4)$$

Note that the “* symbol” in this formula tells Excel to use the multiplication step in the formula, and it stands for “times” in the way we talk about multiplication.

You will recall from Chap. 1 that n stands for the sample size, and so $n - 1$ stands for the sample size minus one.

You will also recall from Chap. 1 that the standard error of the mean, s.e., equals the STDEV divided by the square root of the sample size, n (See Sect. 1.3).

Let's try a sample problem using Excel to find the 95% confidence interval about the mean for a problem.

Suppose that General Motors wanted to claim that its Chevy Impala gets 28 miles per gallon (mpg), and that it wanted to advertise on a billboard in St. Louis at the Vandeventer entrance to Route 44: “The new Chevy Impala gets 28 miles to the gallon.” Let's call 28 mpg the “reference value” for this car.

Suppose that you work for Ford Motor Co. and that you want to check this claim to see if it holds up based on some research evidence. You decide to collect some data and to use a two-side 95% confidence interval about the mean to test your results:

3.1.7 Using Excel to Find the 95% Confidence Interval for a Car's mpg Claim

Objective: To analyze the data using a two-side 95% confidence interval about the mean

You select a sample of new car owners for this car and they agree to keep track of their mileage for two tanks of gas and to record the average miles per gallon they achieve on these two tanks of gas. Your research study produces the results given in Fig. 3.1:

Chevy Impala				
Miles per gallon				
30.9				
24.5				
31.2				
28.7				
35.1				
29.0				
28.8				
23.1				
31.0				
30.2				
28.4				
29.3				
24.2				
27.0				
26.7				
31.0				
23.5				
29.4				
26.3				
27.5				
28.2				
28.4				
29.1				
21.9				
30.9				

Fig. 3.1 Worksheet Data for Chevy Impala (Practical Example)

Create a spreadsheet with these data and use Excel to find the sample size (n), the mean, the standard deviation (STDEV), and the standard error of the mean (s.e.) for these data using the following cell references.

- A3: Chevy Impala
- A5: Miles per gallon
- A6: 30.9

Enter the other mpg data in cells A7: A30

Now, highlight cells A6:A30 and format these numbers in number format (one decimal place). Center these numbers in Column A. Then, widen columns A and B by making both of them twice as wide as the original width of column A. Then, widen column C so that it is three times as wide as the original width of column A so that your table looks more professional.

- C7: n
- C10: Mean
- C13: STDEV
- C16: s.e.
- C19: 95% confidence interval
- D21: Lower limit:
- D23: Upper limit: (see Fig. 3.2)

Chevy Impala				
Miles per gallon				
30.9				
24.5	n			
31.2				
28.7				
35.1	Mean			
29.0				
28.8				
23.1	STDEV			
31.0				
30.2				
28.4	s.e			
29.3				
24.2				
27.0	95% confidence interval			
26.7				
31.0			Lower limit:	
23.5				
29.4			Upper Limit:	
26.3				
27.5				
28.2				
28.4				
29.1				
21.9				
30.9				

Fig. 3.2 Example of Chevy Impala Format for the Confidence Interval About the Mean Labels

- B26: Draw a picture below this confidence interval
- B28: 26.92
- B29: lower (then right-align this word)
- B30: limit (then right-align this word)
- C28: ‘----- 28 -----28.17 -----’ (note that you need to begin cell C28 with a *single quotation mark* (‘) to tell Excel that this is a *label*, and not a number)
- D28: ‘-----’ (notice the single quotation mark at the beginning)
- E28: ‘29.42 (note the single quotation mark)
- C29: ref. Mean
- C30: value
- E29: upper
- E30: limit
- B33: Conclusion:

Now, align the labels underneath the picture of the confidence interval so that they look like Fig. 3.3.

$$F23: = D10 + TINV(1 - .95, 24)*D16 \text{ (no spaces between symbols)}$$

The above formula gives the *upper limit of the confidence interval*, 29.42.

Now, use number format (two decimal places) in your Excel spreadsheet for the mean, standard deviation, standard error of the mean, and for both the lower limit and the upper limit of your confidence interval. If you printed this spreadsheet now, the lower limit of the confidence interval (26.92) and the upper limit of the confidence interval (29.42) would “dribble over” onto a second printed page because the information on the spreadsheet is too large to fit onto one page in its present format.

So, you need to use Excel’s “Scale to Fit” commands that we discussed in Chap. 2 (see Sect. 2.4) to reduce the size of the spreadsheet to 95% of its current size using the Page Layout/Scale to Fit function. Do that now, and notice that the dotted line to the right of 26.92 and 29.42 indicates that these numbers would now fit onto one page when the spreadsheet is printed out (see Fig. 3.4)

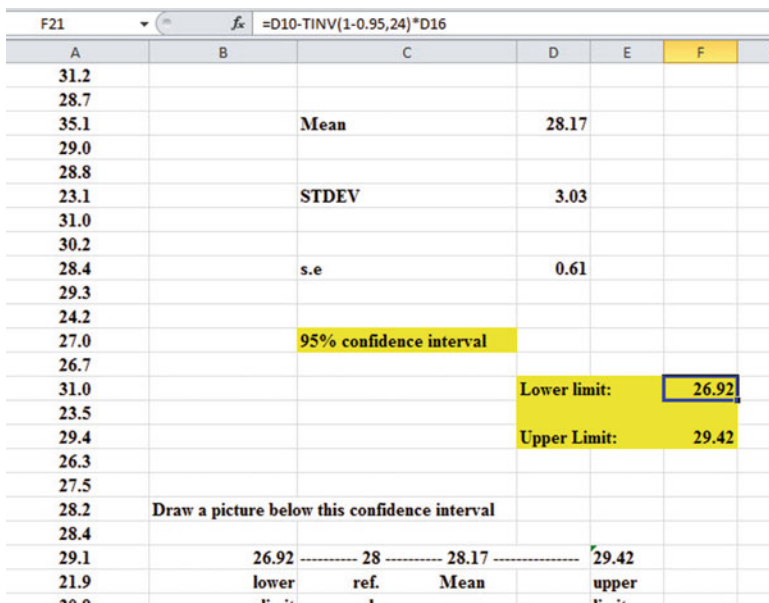


Fig. 3.4 Result of Using the TINV Function to Find the Confidence Interval About the Mean

Note that you have drawn a picture of the 95% confidence interval beneath cell B26, including the lower limit, the upper limit, the mean, and the reference value of 28 mpg given in the claim that the company wants to make about the car’s miles per gallon performance.

Now, let’s write the conclusion to your research study on your spreadsheet:

- C33: Since the reference value of 28 is inside
- C34: the confidence interval, we accept that
- C35: the Chevy Impala does get 28 mpg.

Your research study accepted the claim that the Chevy Impala did get 28 miles per gallon. The average miles per gallon in your study was 28.17 (See Fig. 3.5).

Save your resulting spreadsheet as: **CHEVY7**

Chevy Impala				
Miles per gallon				
30.9				
24.5	n		25	
31.2				
28.7				
35.1	Mean		28.17	
29.0				
28.8				
23.1	STDEV		3.03	
31.0				
30.2				
28.4	s.e		0.61	
29.3				
24.2				
27.0	95% confidence interval			
26.7				
31.0			Lower limit:	26.92
23.5				
29.4			Upper Limit:	29.42
26.3				
27.5				
28.2	Draw a picture below this confidence interval			
28.4				
29.1	26.92	----- 28 -----	28.17	----- 29.42
21.9	lower	ref.	Mean	upper
30.9	limit	value		limit
	Conclusion:	Since the reference value of 28 is inside the confidence interval, we accept that the Chevy Impala does get 28 mpg.		

Fig. 3.5 Final Spreadsheet for the Chevy Impala Confidence Interval About the Mean

3.2 Hypothesis Testing

One of the important activities of researchers, whether they are in business research, marketing research, psychological research, educational research, or in any of the social sciences is that they attempt to “check” their assumptions about the world by testing these assumptions in the form of hypotheses.

A typical hypothesis is in the form: “If *x*, then *y*.”

Some examples would be:

1. “If we raise our price by 5%, then our sales dollars for our product will decrease by 8%.”

2. “If we increase our advertising budget by \$400,000 for our product, then our market share will go up by two points.”
3. “If we use this new method of teaching mathematics to ninth graders in algebra, then our math achievement scores will go up by 10%.”
4. “If we change the raw materials for this product, then our production cost per unit will decrease by 5%.”

A hypothesis, then, to a social science researcher is a “guess” about what we think is true in the real world. We can test these guesses using statistical formulas to see if our predictions come true in the real world.

So, in order to perform these statistical tests, we must first state our hypotheses so that we can test our results against our hypotheses to see if our hypotheses match reality.

So, how do we generate hypotheses in business?

3.2.1 Hypotheses Always Refer to the Population of People or Events That You Are Studying

The first step is to understand that our hypotheses always refer to the *population* of people under study.

For example, if we are interested in studying 18–24 year-olds in St. Louis as our target market, and we select a sample of people in this age group in St. Louis, depending on how we select our sample, we are hoping that our results of this study are useful in generalizing our findings to *all* 18–24 year-olds in St. Louis, and not just to the particular people in our sample.

The entire group of 18–24 year-olds in St. Louis would be the *population* that we are interested in studying, while the particular group of people in our study are called the *sample* from this population.

Since our sample sizes typically contain only a few people, we interested in the results of our sample *only insofar as the results of our sample can be “generalized” to the population in which we are really interested.*

That is why our hypotheses always refer to the population, and never to the sample of people in our study.

You will recall from Chap. 1 that we used the symbol: \bar{X} to refer to the mean of the sample we use in our research study (See Sect. 1.1).

We will use the symbol: μ (the Greek letter “mu”) to refer to the *population mean*.

In testing our hypotheses, we are trying to decide which one of two competing hypotheses *about the population mean* we should accept given our data set.

3.2.2 *The Null Hypothesis and the Research (Alternative) Hypothesis*

These two hypotheses are called the *null hypothesis* and the *research hypothesis*.

Statistics textbooks typically refer to the *null hypothesis* with the notation: H_0 .

The *research hypothesis* is typically referred to with the notation: H_1 , and it is sometimes called the *alternative hypothesis*.

Let's explain first what is meant by the null hypothesis and the research hypothesis:

- (1) *The null hypothesis is what we accept as true unless we have compelling evidence that it is not true.*
- (2) *The research hypothesis is what we accept as true whenever we reject the null hypothesis as true.*

This is similar to our legal system in America where we assume that a supposed criminal is innocent until he or she is proven guilty in the eyes of a jury. Our null hypothesis is that this defendant is innocent, while the research hypothesis is that he or she is guilty.

In the great state of Missouri, every license plate has the state slogan: "Show me." This means that people in Missouri think of themselves as not gullible enough to accept everything that someone says as true unless that person's actions indicate the truth of his or her claim. In other words, people in Missouri believe strongly that a person's actions speak much louder than that person's words.

Since both the null hypothesis and the research hypothesis cannot both be true, the task of hypothesis testing using statistical formulas is to decide which one you will accept as true, and which one you will reject as true.

Sometimes in business research a series of rating scales is used to measure people's attitudes toward a company, toward one of its products, or toward their intention-to-buy that company's products. These rating scales are typically 5-point, 7-point, or 10-point scales, although other scale values are often used as well.

3.2.2.1 **Determining the Null Hypothesis and the Research Hypothesis When Rating Scales Are Used**

Here is a typical example of a 7-point scale in attitude research in customer satisfaction studies (see Fig. 3.6):

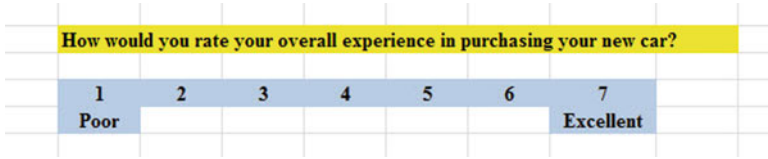


Fig. 3.6 Example of a Rating Scale Item for a New Car Purchase (Practical Example)

So, how do we decide what to use as the null hypothesis and the research hypothesis whenever rating scales are used?

Objective: To decide on the null hypothesis and the research hypothesis whenever rating scales are used.

In order to make this determination, we will use a simple rule:

Rule: Whenever rating scales are used, we will use the “middle” of the scale as the null hypothesis and the research hypothesis.

In the above example, since 4 is the number in the middle of the scale (i.e., three numbers are below it, and three numbers are above it), our hypotheses become:

Null hypothesis: $\mu = 4$

Research hypothesis: $\mu \neq 4$

In the above rating scale example, if the result of our statistical test for this one attitude scale item indicates that our population mean is “close to 4,” we say that we accept the null hypothesis that our new car purchase experience was neither positive nor negative.

In the above example, if the result of our statistical test indicates that the population mean is significantly different from 4, we reject the null hypothesis and accept the research hypothesis by stating either that:

“The new car purchase experience was significantly positive” (this is true whenever our sample mean is significantly greater than our expected population mean of 4).

or

“The new car purchase experience was significantly negative” (this is accepted as true whenever our sample mean is significantly less than our expected population mean of 4).

Both of these conclusions cannot be true. We accept one of the hypotheses as “true” based on the data set in our research study, and the other one as “not true” based on our data set.

The job of the business researcher, then, is to decide which of these two hypotheses, the null hypothesis or the research hypothesis, he or she will accept as true given the data set in the research study.

Let’s try some examples of rating scales so that you can practice figuring out what the null hypothesis and the research hypothesis are for each rating scale.

In the spaces in Fig. 3.7, write in the null hypothesis and the research hypothesis for the rating scales:

1. Webster University is an excellent university.										
	1	2	3	4	5					
	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree					
	Null hypothesis:			$\mu =$	_____					
	Research hypothesis:			$\mu \neq$	_____					
2. How would you rate the quality of teaching at Webster University?										
	poor	1	2	3	4	5	6	7	excellent	
	Null hypothesis:			$\mu =$	_____					
	Research hypothesis:			$\mu \neq$	_____					
3. How would you rate the quality of the faculty at Webster University?										
	1	2	3	4	5	6	7	8	9	10
	very poor									very good
	Null hypothesis:			$\mu =$	_____					
	Research hypothesis:			$\mu \neq$	_____					

Fig. 3.7 Examples of Rating Scales for Determining the Null Hypothesis and the Research Hypothesis

How did you do?

Here are the answers to these three questions:

1. The null hypothesis is 3, and the research hypothesis is not equal to 3 on this 5-point scale (i.e. the “middle” of the scale is 3).
2. The null hypothesis is 4, and the research hypothesis is not equal to 4 on this 7-point scale (i.e., the “middle” of the scale is 4).
3. The null hypothesis is 5.5, and the research hypothesis is not equal to 5.5 on this 10-point scale (i.e., the “middle” of the scale is 5.5 since there are five numbers below 5.5 and five numbers above 5.5).

As another example, Holiday Inn Express in its Stay Smart Experience Survey uses 4-point scales where:

- 1 = Not So Good
- 2 = Average
- 3 = Very Good
- 4 = Great

On this scale, the null hypothesis is: $\mu = 2.5$ and the research hypothesis is: $\mu \neq 2.5$, because there are two numbers below 2.5, and two numbers above 2.5 on that rating scale.

Now, let's discuss the 7 STEPS of hypothesis testing for using the confidence interval about the mean.

3.2.3 The 7 Steps for Hypothesis-Testing Using the Confidence Interval About the Mean

Objective: To learn the 7 steps of hypothesis-testing using the confidence interval about the mean

There are seven basic steps of hypothesis-testing for this statistical test.

3.2.3.1 STEP 1: State the Null Hypothesis and the Research Hypothesis

If you are using numerical scales in your survey, you need to remember that these hypotheses refer to the “middle” of the numerical scale. For example, if you are using 7-point scales with 1 = poor and 7 = excellent, these hypotheses would refer to the middle of these scales and would be:

Null hypothesis H_0 : $\mu = 4$
 Research hypothesis H_1 : $\mu \neq 4$

3.2.3.2 STEP 2: Select the Appropriate Statistical Test

In this chapter we are studying the confidence interval about the mean, and so we will select that test.

3.2.3.3 STEP 3: Calculate the Formula for the Statistical Test

You will recall (see Sect. 3.1.5) that the formula for the confidence interval about the mean is:

$$\bar{X} \pm t \text{ s.e.} \quad (3.2)$$

We discussed the procedure for computing this formula for the confidence interval about the mean using Excel earlier in this chapter, and the steps involved in using that formula are:

1. Use Excel's =COUNT function to find the sample size.
2. Use Excel's =AVERAGE function to find the sample mean, \bar{X} .
3. Use Excel's =STDEV function to find the standard deviation, STDEV.
4. Find the standard error of the mean (s.e.) by dividing the standard deviation (STDEV) by the square root of the sample size, n .
5. Use Excel's TINV function to find the lower limit of the confidence interval.
6. Use Excel's TINV function to find the upper limit of the confidence interval.

3.2.3.4 STEP 4: Draw a Picture of the Confidence Interval About the Mean, Including the Mean, the Lower Limit of the Interval, the Upper Limit of the Interval, and the Reference Value Given in the Null Hypothesis, H_0

3.2.3.5 STEP 5: Decide on a Decision Rule

- (a) *If the reference value is inside the confidence interval, accept the null hypothesis, H_0*
- (b) *If the reference value is outside the confidence interval, reject the null hypothesis, H_0 , and accept the research hypothesis, H_1*

3.2.3.6 STEP 6: State the Result of Your Statistical Test

There are two possible results when you use the confidence interval about the mean, and only one of them can be accepted as "true." So your result would be one of the following:

Either: Since the reference value is inside the confidence interval, *we accept the null hypothesis, H_0*

Or: Since the reference value is outside the confidence interval, *we reject the null hypothesis, H_0 , and accept the research hypothesis, H_1*

3.2.3.7 STEP 7: State the Conclusion of Your Statistical Test in Plain English!

In practice, this is more difficult than it sounds because you are trying to summarize the result of your statistical test in simple English that is both concise and accurate so that someone who has never had a statistics course (such as your boss, perhaps) can understand the conclusion of your test. This is a difficult task, and we will give you lots of practice doing this last and most important step throughout this book.

Objective: To write the conclusion of the confidence interval about the mean test

Let's set some basic rules for stating the conclusion of a hypothesis test.

Rule #1: *Whenever you reject H_0 and accept H_1 , you must use the word “significantly” in the conclusion to alert the reader that this test found an important result.*

Rule #2: *Create an outline in words of the “key terms” you want to include in your conclusion so that you do not forget to include some of them.*

Rule #3: *Write the conclusion in plain English so that the reader can understand it even if that reader has never taken a statistics course.*

Let's practice these rules using the Chevy Impala Excel spreadsheet that you created earlier in this chapter, but first we need to state the hypotheses for that car.

Since the billboard wants to claim that the Chevy Impala gets 28 miles per gallon, the hypotheses would be:

$$H_0: \mu = 28 \text{ mpg}$$

$$H_1: \mu \neq 28 \text{ mpg}$$

You will remember that the reference value of 28 mpg was inside the 95% confidence interval about the mean for your data, so we would accept H_0 for the Chevy Impala that the car does get 28 mpg.

Objective: To state the result when you accept H_0

Result: Since the reference value of 28 mpg is inside the confidence interval, we accept the null hypothesis, H_0

Let's try our three rules now:

Objective: To write the conclusion when you accept H_0

Rule #1: *Since the reference value was inside the confidence interval, we cannot use the word “significantly” in the conclusion. This is a basic rule we are using in this chapter for every problem.*

Rule #2: The key terms in the conclusion would be:

- Chevy Impala
- reference value of 28 mpg

Rule #3: The Chevy Impala did get 28 mpg.

The process of writing the conclusion when you accept H_0 is relatively straightforward since you put into words what you said when you wrote the null hypothesis.

However, the process of stating the conclusion when you reject H_0 and accept H_1 is more difficult, so let’s practice writing that type of conclusion with three practice case examples:

Objective: To write the result and conclusion when you reject H_0

CASE #1: Suppose that an ad in *Business Week* claimed that the Ford Escape Hybrid got 34 miles per gallon. The hypotheses would be:

$$H_0: \mu = 34 \text{ mpg}$$

$$H_1: \mu \neq 34 \text{ mpg}$$

Suppose that your research yields the following confidence interval:

30	31	32	34
lower	Mean	upper	Ref.
limit		limit	Value

Result: Since the reference value is outside the confidence interval, we reject the null hypothesis and accept the research hypothesis

The three rules for stating the conclusion would be:

Rule #1: We must include the word “significantly” since the reference value of 34 is outside the confidence interval.

Rule #2: The key terms would be:

- Ford Escape Hybrid
- significantly
- either “more than” or “less than”
- and probably closer to

Rule #3: The Ford Escape Hybrid got significantly less than 34 mpg, and it was probably closer to 31 mpg.

Note that this conclusion says that the mpg was less than 34 mpg because the sample mean was only 31 mpg. Note, also, that when you find a significant result by rejecting the null hypothesis, *it is not sufficient to say only: “significantly less than 34 mpg,”* because that does not tell the reader “how much less than 34 mpg” the

sample mean was from 34 mpg. To make the conclusion clear, you need to add: “probably closer to 31 mpg” since the sample mean was only 31 mpg.

CASE #2: Suppose that you have been hired as a consultant by the St. Louis Symphony Orchestra (SLSO) to analyze the data from an Internet survey of attendees for a concert in Powell Symphony Hall in St. Louis last month. You have decided to practice your data analysis skills on Question #7 given in Fig. 3.8:

Question #7:	"Overall, how satisfied have you been with your experience(s) at SLSO concerts?"						
	1	2	3	4	5	6	7
	Extremely dissatisfied						Extremely satisfied

Fig. 3.8 Example of a Survey Item Used by the St. Louis Symphony Orchestra (SLSO)

The hypotheses for this one item would be:

$$H_0: \mu = 4$$

$$H_1: \mu \neq 4$$

Essentially, the null hypothesis equal to four states that if the obtained mean score for this question is not significantly different from 4 on the rating scale, then attendees, overall, were neither satisfied nor dissatisfied with their SLSO concerts.

Suppose that your analysis produced the following confidence interval for this item on the survey.

1.8	2.8	3.8	4
lower limit	Mean	upper limit	Ref. Value

Result: Since the reference value is outside the confidence interval, we reject the null hypothesis and accept the research hypothesis.

Rule #1: You must include the word “significantly” since the reference value is outside the confidence interval

Rule #2: The key terms would be:

- attendees
- SLSO Internet survey
- significantly
- last month
- either satisfied or dissatisfied (since the result is significant)
- experiences at concerts
- overall

Rule #3: Attendees were significantly dissatisfied, overall, on last month’s Internet survey with their experiences at concerts of the SLSO.

Note that you need to use the word “dissatisfied” since the sample mean of 2.8 was on the dissatisfied side of the middle of the rating scale.

CASE #3: Suppose that Marriott Hotel at the St. Louis Airport location had the results of one item in its Guest Satisfaction Survey from last week’s customers that was the following (see Fig. 3.9):

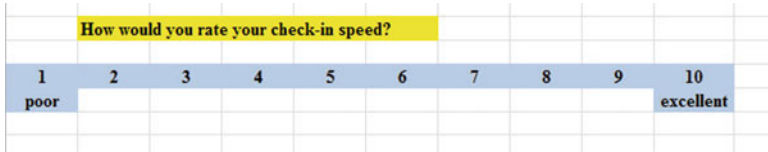


Fig. 3.9 Example of a Survey Item from Marriott Hotels

This item would have the following hypotheses:

$$H_0: \mu = 5.5$$

$$H_1: \mu \neq 5.5$$

Suppose that your research produced the following confidence interval for this item on the survey:

5.5	5.7	5.8	5.9
Ref. Value	lower limit	Mean	upper limit

Result: Since the reference value is outside the confidence interval, we reject the null hypothesis and accept the research hypothesis

The three rules for stating the conclusion would be:

Rule #1: You must include the word “significantly” since the reference value is outside the confidence interval

Rule #2: The key terms would be:

- Marriott Hotel
- St. Louis Airport
- significantly
- check-in speed
- survey
- last week
- customers
- either “positive” or “negative” (we will explain this)

Rule #3: Customers at the St. Louis Airport Marriott Hotel last week rated their check-in speed in a survey as significantly positive.

Note two important things about this conclusion above: (1) people when speaking English do not normally say “significantly excellent” since something is either excellent or is not excellent without any modifier, and (2) since the mean rating of the check-in speed (5.8) was significantly greater than 5.5 on the positive side of the scale, we would say “significantly positive” to indicate this fact.

The three practice problems at the end of this chapter will give you additional practice in stating the conclusion of your result, and this book will include many more examples that will help you to write a clear and accurate conclusion to your research findings.

3.3 Alternative Ways to Summarize the Result of a Hypothesis Test

It is important for you to understand that in this book we are summarizing an hypothesis test in one of two ways: (1) We accept the null hypothesis, or (2) We reject the null hypothesis and accept the research hypothesis. We are consistent in the use of these words so that you can understand the concept underlying hypothesis testing.

However, there are many other ways to summarize the result of an hypothesis test, and all of them are correct theoretically, even though the terminology differs. If you are taking a course with a professor who wants you to summarize the results of a statistical test of hypotheses in language which is different from the language we are using in this book, do not panic! If you understand the concept of hypothesis testing as described in this book, you can then translate your understanding to use the terms that your professor wants you to use to reach the same conclusion to the hypothesis test.

Statisticians and professors of business statistics all have their own language that they like to use to summarize the results of an hypothesis test. There is no one set of words that these statisticians and professors will ever agree on, and so we have chosen the one that we believe to be easier to understand in terms of the concept of hypothesis testing.

To convince you that there are many ways to summarize the results of an hypothesis test, we present the following quotes from prominent statistics and research books to give you an idea of the different ways that are possible.

3.3.1 Different Ways to Accept the Null Hypothesis

The following quotes are typical of the language used in statistics and research books when the null hypothesis is accepted:

- “The null hypothesis is not rejected.” (Black 2010, p. 310)
- “The null hypothesis cannot be rejected.” (McDaniel and Gates 2010, p. 545)
- “The null hypothesis . . . claims that there is no difference between groups.” (Salkind 2010, p. 193)
- “The difference is not statistically significant.” (McDaniel and Gates 2010, p. 545)
- “ . . . the obtained value is not extreme enough for us to say that the difference between Groups 1 and 2 occurred by anything other than chance.” (Salkind 2010, p. 225)
- “If we do not reject the null hypothesis, we conclude that there is not enough statistical evidence to infer that the alternative (hypothesis) is true.” (Keller 2009, p. 358)
- “The research hypothesis is not supported.” (Zikmund and Babin 2010, p. 552)

3.3.2 *Different Ways to Reject the Null Hypothesis*

The following quotes are typical of the quotes used in statistics and research books when the null hypothesis is rejected:

- “The null hypothesis is rejected.” (McDaniel and Gates 2010, p. 546)
- “If we reject the null hypothesis, we conclude that there is enough statistical evidence to infer that the alternative hypothesis is true.” (Keller 2009, p. 358)
- “If the test statistic’s value is inconsistent with the null hypothesis, we reject the null hypothesis and infer that the alternative hypothesis is true.” (Keller 2009, p. 348)
- “Because the observed value . . . is greater than the critical value . . . , the decision is to reject the null hypothesis.” (Black 2010, p. 359)
- “If the obtained value is more extreme than the critical value, the null hypothesis cannot be accepted.” (Salkind 2010, p. 243)
- “The critical t-value . . . must be surpassed by the observed t-value if the hypothesis test is to be statistically significant” (Zikmund and Babin 2010, p. 567)
- “The calculated test statistic . . . exceeds the upper boundary and falls into this rejection region. The null hypothesis is rejected.” (Weiers 2011, p. 330)

You should note that all of the above quotes are used by statisticians and professors when discussing the results of an hypothesis test, and so you should not be surprised if someone asks you to summarize the results of a statistical test using a different language than the one we are using in this book.

3.4 End-of-Chapter Practice Problems

1. Many US airlines charge a “change fee” if a passenger cancels a scheduled flight and wants to change the date/time of that flight. This fee typically ranges from US\$150 to US\$200. It is typical that if the airline cancels a flight due to inclement weather or mechanical problems with the aircraft, that this fee is waived.

Suppose that you purchased a ticket on Delta Airlines to fly from St. Louis, Missouri (US) to Boston, Massachusetts (US) and that Delta cancelled your flight because a major winter storm came up the east coast of the US and resulted in many flights of many airlines being cancelled. Suppose, further, that you called

Create an Excel spreadsheet with these data.

- (a) Use Excel to the right of the table to find the sample size, mean, standard deviation, and standard error of the mean for these data. Label your answers, and use two decimal places for the mean, standard deviation, and standard error of the mean
 - (b) Enter the null hypothesis and the research hypothesis for this item on your spreadsheet.
 - (c) Use Excel's TINV function to find the 95% confidence interval about the mean for these data. Label your answers on your spreadsheet. Use two decimal places for the lower limit and the upper limit of the confidence interval.
 - (d) Enter the *result* of the test on your spreadsheet.
 - (e) Enter the *conclusion* of the test in plain English on your spreadsheet.
 - (f) Print your final spreadsheet so that it fits onto one page (if you need help remembering how to do this, see the objectives at the end of Chap. 2 in Sect. 2.4).
 - (g) Draw a picture of the confidence interval, including the reference value, onto your spreadsheet.
 - (h) Save the final spreadsheet as: LIFETIME11
3. Suppose that you have been asked to conduct three focus groups in different cities with adult women (ages 25–44) to determine how much they liked a new design of a blouse that was created by a well-known designer. The designer is hoping to sell this blouse in department stores at a retail price of \$68.00. You conduct a one-hour focus group discussion with three groups of adult women in this age range, and the last question on the survey at the end of the discussion period produced the hypothetical results given in Fig. 3.12:

FOCUS GROUP PRICING STUDY	
Question #10: "How much would you be willing to pay for this blouse?"	
	\$ _____
Groups 1,2,3 in \$	
62	
55	
73	
53	
46	
48	
57	
59	
65	
68	
64	
72	
62	
67	
59	
71	
65	
63	
69	
71	
70	
58	
67	
65	
63	
59	
70	
67	
64	
65	

Fig. 3.12 Worksheet Data for Chap. 3: Practice Problem #3

Create an Excel spreadsheet with these data.

- (a) Use Excel to the right of the table to find the sample size, mean, standard deviation, and standard error of the mean for these data. Label your answers, and use two decimal places and currency format for the mean, standard deviation, and standard error of the mean
- (b) Enter the null hypothesis and the research hypothesis for this item onto your spreadsheet.
- (c) Use Excel's TINV function to find the 95% confidence interval about the mean for these data. Label your answers on your spreadsheet. Use two decimal places in currency format for the lower limit and the upper limit of the confidence interval.

- (d) Enter the *result* of the test on your spreadsheet.
- (e) Enter the *conclusion* of the test in plain English on your spreadsheet.
- (f) Print your final spreadsheet so that it fits onto one page (if you need help remembering how to do this, see the objectives at the end of Chap. 2 in Sect. 2.4).
- (g) Draw a picture of the confidence interval, including the reference value, onto your spreadsheet.
- (h) Save the final spreadsheet as: blouse9

References

- Black, K. Business Statistics: for Contemporary Decision Making (6th ed.). Hoboken, NJ: John Wiley & Sons, Inc., 2010.
- Keller, G. Statistics for Management and Economics (8th ed.). Mason, OH: South-Western Cengage learning, 2009.
- Levine, D.M. Statistics for Managers using Microsoft Excel (6th ed.). Boston, MA: Prentice Hall/Pearson, 2011.
- McDaniel, C. and Gates, R. Marketing Research (8th ed.). Hoboken, NJ: John Wiley & Sons, Inc., 2010.
- Salkind, N.J. Statistics for People Who (think they) Hate Statistics (2nd Excel 2007 ed.). Los Angeles, CA: Sage Publications, 2010.
- Weiers, R.M. Introduction to Business Statistics (7th ed.). Mason, OH: South-Western Cengage Learning, 2011.
- Zikmund, W.G. and Babin, B.J. Exploring Marketing Research (10th ed.). Mason, OH: South-Western Cengage learning, 2010.

Chapter 4

One-Group t-Test for the Mean



In this chapter, you will learn how to use one of the most popular and most helpful statistical tests in social science research: the one-group t-test for the mean.

The formula for the one-group t-test is as follows:

$$t = \frac{\bar{X} - \mu}{S_{\bar{X}}} \text{ where} \tag{4.1}$$

$$\text{s.e.} = S_{\bar{X}} = \frac{S}{\sqrt{n}} \tag{4.2}$$

This formula asks you to take the mean (\bar{X}) and subtract the population mean (μ) from it, and then divide the answer by the standard error of the mean (s.e.). The standard error of the mean equals the standard deviation divided by the square root of n (the sample size).

Let's discuss the seven STEPS of hypothesis testing using the one-group t-test so that you can understand how this test is used.

4.1 The 7 STEPS for Hypothesis-Testing Using the One-Group t-Test

Objective: To learn the 7 steps of hypothesis-testing using the one-group t-test

Before you can try out your Excel skills on the one-group t-test, you need to learn the basic steps of hypothesis-testing for this statistical test. There are 7 steps in this process:

4.1.1 STEP 1: State the Null Hypothesis and the Research Hypothesis

If you are using numerical scales in your survey, you need to remember that these hypotheses refer to the “middle” of the numerical scale. For example, if you are using 7-point scales with 1 = poor and 7 = excellent, these hypotheses would refer to the middle of these scales and would be:

Null hypothesis H_0 : $\mu = 4$

Research hypothesis H_1 : $\mu \neq 4$

As a second example, suppose that you worked for Honda Motor Company and that you wanted to place a magazine ad that claimed that the new Honda Fit got 35 miles per gallon (mpg). The hypotheses for testing this claim on actual data would be:

H_0 : $\mu = 35$ mpg

H_1 : $\mu \neq 35$ mpg

4.1.2 STEP 2: Select the Appropriate Statistical Test

In this chapter we will be studying the one-group t-test, and so we will select that test.

4.1.3 STEP 3: Decide on a Decision Rule for the One-Group t-Test

- (a) If the absolute value of t is less than the critical value of t , accept the null hypothesis.
- (b) If the absolute value of t is greater than the critical value of t , reject the null hypothesis and accept the research hypothesis.

You are probably saying to yourself: “That sounds fine, but how do I find the absolute value of t ?”

4.1.3.1 Finding the Absolute Value of a Number

To do that, we need another objective:

Objective: To find the absolute value of a number

If you took a basic algebra course in high school, you may remember the concept of “absolute value.” In mathematical terms, the absolute value of any number is *always* that number expressed as a positive number.

For example, the absolute value of 2.35 is +2.35.

And the absolute value of minus 2.35 (i.e. -2.35) is also +2.35.

This becomes important when you are using the t-table in Appendix E of this book. We will discuss this table later when we get to Step 5 of the one-group t-test where we explain how to find the critical value of t using Appendix E.

4.1.4 STEP 4: Calculate the Formula for the One-Group t-Test

Objective: To learn how to use the formula for the one-group t-test

The formula for the one-group t-test is as follows:

$$t = \frac{\bar{X} - \mu}{S_{\bar{X}}} \text{ where} \quad (4.1)$$

$$\text{s.e.} = S_{\bar{X}} = \frac{S}{\sqrt{n}} \quad (4.2)$$

This formula makes the following assumptions about the data (Foster et al. 1998): (1) The data are independent of each other (i.e., each person receives only one score), (2) the *population* of the data is normally distributed, and (3) the data have a constant variance (note that the standard deviation is the square root of the variance).

To use this formula, you need to follow these steps:

1. Take the sample mean in your research study and subtract the population mean μ from it (remember that the population mean for a study involving numerical rating scales is the “middle” number in the scale).
2. Then take your answer from the above step, and divide your answer by the standard error of the mean for your research study (you will remember that you learned how to find the standard error of the mean in Chap. 1; to find the standard error of the mean, just take the standard deviation of your research study and divide it by the square root of n , where n is the number of people used in your research study).
3. The number you get after you complete the above step is the value for t that results when you use the formula stated above.

4.1.5 *STEP 5: Find the Critical Value of t in the t-Table in Appendix E*

Objective: To find the critical value of t in the t-table in Appendix E

Before we get into an explanation of what is meant by “the critical value of t ,” let’s give you practice in finding the critical value of t by using the t-table in Appendix E.

Keep your finger on Appendix E as we explain how you need to “read” that table.

Since the test in this chapter is called the “one-group t-test,” you will use the first column on the left in Appendix E to find the critical value of t for your research study (note that this column is headed: “sample size n ”).

To find the critical value of t , you go down this first column until you find the sample size in your research study, and then you go to the right and read the critical value of t for that sample size in the critical t column in the table (note that *this is the column that you would use for both the one-group t-test and the 95% confidence interval about the mean*).

For example, if you have 27 people in your research study, the critical value of t is 2.056.

If you have 38 people in your research study, the critical value of t is 2.026.

If you have more than 40 people in your research study, the critical value of t is always 1.96.

Note that the “critical t column” in Appendix E represents the value of t that you need to obtain to be 95% confident of your results as “significant” results.

The critical value of t is the value that tells you whether or not you have found a “significant result” in your statistical test.

The t-table in Appendix E represents a series of “bell-shaped normal curves” (they are called bell-shaped because they look like the outline of the Liberty Bell that you can see in Philadelphia outside of Independence Hall).

The “middle” of these normal curves is treated as if it were zero point on the x -axis (the technical explanation of this fact is beyond the scope of this book, but any good statistics book (e.g. Zikmund and Babin 2010) will explain this concept to you if you are interested in learning more about it).

Thus, values of t that are to the right of this zero point are positive values that use a plus sign before them, and values of t that are to the left of this zero point are negative values that use a minus sign before them. Thus, some values of t are positive, and some are negative.

However, every statistics book that includes a t-table only reprints the *positive* side of the t-curves because the negative side is the mirror image of the positive side; this means that the negative side contains the exact same numbers as the positive side, but the negative numbers all have a minus sign in front of them.

Therefore, to use the t-table in Appendix E, you need to *take the absolute value of the t-value you found when you use the t-test formula* since the t-table in Appendix E only has the values of t that are the positive values for t.

Throughout this book, we are assuming that you want to be 95% confident in the results of your statistical tests. Therefore, the value for t in the t-table in Appendix E tells you whether or not the t-value you obtained when you used the formula for the one-group t-test is within the 95% interval of the t-curve range that that t-value would be expected to occur with 95% confidence.

If the t-value you obtained when you used the formula for the one-group t-test is *inside* of the 95% confidence range, we say that the result you found is *not significant* (note that this is equivalent to *accepting the null hypothesis!*).

If the t-value you found when you used the formula for the one-group t-test is *outside* of this 95% confidence range, we say that you have found a *significant result* that would be expected to occur less than 5% of the time (note that this is equivalent to *rejecting the null hypothesis and accepting the research hypothesis*).

4.1.6 STEP 6: State the Result of Your Statistical Test

There are two possible results when you use the one-group t-test, and only one of them can be accepted as “true.”

Either: Since the absolute value of t that you found in the t-test formula is *less than the critical value of t* in Appendix E, you accept the null hypothesis.

Or: Since the absolute value of t that you found in the t-test formula is *greater than the critical value of t* in Appendix E, you reject the null hypothesis, and accept the research hypothesis.

4.1.7 STEP 7: State the Conclusion of Your Statistical Test in Plain English!

In practice, this is more difficult than it sounds because you are trying to summarize the result of your statistical test in simple English that is both concise and accurate so that someone who has never had a statistics course (such as your boss, perhaps) can understand the result of your test. This is a difficult task, and we will give you lots of practice doing this last and most important step throughout this book.

If you have read this far, you are ready to sit down at your computer and perform the one-group t-test using Excel on some hypothetical data from the Guest Satisfaction Survey used by Marriott Hotels.

Let’s give this a try.

4.2 One-Group t-Test for the Mean

Suppose that you have been hired as a statistical consultant by Marriott Hotel in St. Louis to analyze the data from a Guest Satisfaction survey that they give to all customers to determine the degree of satisfaction of these customers for various activities of the hotel.

The survey contains a number of items, but suppose item #7 is the one in Fig. 4.1:

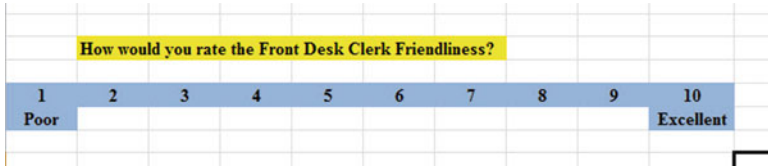


Fig. 4.1 Sample Survey Item for Marriot Hotel (Practical Example)

Suppose further, that you have decided to analyze the data from last week's customers using the one-group t-test.

Important note: You would need to use this test for each of the survey items separately.

Suppose that the hypothetical data for Item #7 from last week at the St. Louis Marriott Hotel were based on a sample size of 124 guests who had a mean score on this item of 6.58 and a standard deviation on this item of 2.44.

Objective: To analyze the data for each question separately using the one-group t-test for each survey item.

Create an Excel spreadsheet with the following information:

B11: Null hypothesis:

B14: Research hypothesis:

Note: Remember that when you are using a rating scale item, both the null hypothesis and the research hypothesis refer to the "middle of the scale." In the 10-point scale in this example, the middle of the scale is 5.5 since five numbers are below 5.5 (i.e., 1–5) and five numbers are above 5.5 (i.e. 6–10). Therefore, the hypotheses for this rating scale item are:

H₀: $\mu = 5.5$

H₁: $\mu \neq 5.5$

B17: n

B20: mean

B23: STDEV

- B26: s.e.
- B29: critical t
- B32: t-test
- B36: Result:
- B41: Conclusion:

Now, use Excel:

- D17: enter the sample size
- D20: enter the mean
- D23: enter the STDEV (see Fig. 4.2)

Fig. 4.2 Basic Data
Table for Front Desk Clerk
Friendliness

Null hypothesis:					
Research hypothesis:					
n		124			
mean		6.58			
STDEV		2.44			
s.e.					
critical t					
t-test					
Result:					
Conclusion:					

D26 compute the standard error using the formula in Chap. 1

D29: find the critical t value of t in the t-table in Appendix E

Now, enter the following formula in cell D32 to find the t-test result:

$$= (D20 - 5.5)/D26 \quad (\text{no spaces between symbols})$$

This formula takes the sample mean (D20) and subtracts the population hypothesized mean of 5.5 from the sample mean, and THEN divides the answer by the standard error of the mean (D26). Note that you need to enter D20–5.5 with an open-parenthesis *before* D20 and a closed-parenthesis *after* 5.5 so that the *answer of 1.08 is THEN divided by the standard error of 0.22* to get the t-test result of 4.93.

Now, use two decimal places for both the s.e. and the t-test result (see Fig. 4.3).

Fig. 4.3 t-test Formula
Result for Front Desk Clerk
Friendliness

Null hypothesis:		
Research hypothesis:		
n		124
mean		6.58
STDEV		2.44
s.e.		0.22
critical t		1.96
t-test		4.93
Result:		
Conclusion:		

Now, write the following sentence in D36:D39 to summarize the result of the t-test:

D36: Since the absolute value of t of 4.93 is
D37: greater than the critical t of 1.96, we
D38: reject the null hypothesis and accept
D39: the research hypothesis.

Lastly, write the following sentence in D41:D43 to summarize the conclusion of the result for Item #7 of the Marriott Guest Satisfaction Survey:

D41: St. Louis Marriott Hotel guests rated the
D42: Front Desk Clerks as significantly
D43: friendly last week.

Save your file as: MARRIOTT3

Important note: You are probably wondering why we entered both the result and the conclusion in separate cells instead of in just one cell. This is because if you enter them in one cell, you will be very disappointed when you print out your final spreadsheet, because one of two things will happen that you will not like: (1) if you print the spreadsheet to fit onto only one page, the result and the conclusion will force the entire spreadsheet to be printed in such small font size that you will be unable to read it, or (2) if you do not print the final spreadsheet to fit onto one page, both the result and the conclusion will “dribble over” onto a second page instead of fitting the entire spread-sheet onto one page. In either case, your spreadsheet will not have a “professional look.”

Print the final spreadsheet so that it fits onto one page as given in Fig. 4.4. Enter the null hypothesis and the research hypothesis by hand on your spreadsheet

Fig. 4.4 Final Spreadsheet for Front Desk Clerk Friendliness

Null hypothesis:	$\mu = 5.5$
Research hypothesis:	$\mu \neq 5.5$
n	124
mean	6.58
STDEV	2.44
s.e.	0.22
critical t	1.96
t-test	4.93
Result:	Since the absolute value of t of 4.93 is greater than the critical t of 1.96, we reject the null hypothesis and accept the research hypothesis.
Conclusion:	St. Louis Marriott Hotel guests rated the Front Desk Clerks as significantly friendly last week.

Important note: It is important for you to understand that “technically” the above conclusion in statistical terms should state:

“St. Louis Marriott Hotel Guests rated the Front Desk Clerks as friendly last week, and this result was probably not obtained by chance.”

However, throughout this book, we are using the term “significantly” in writing the conclusion of statistical tests to alert the reader that the result of the statistical test was probably not a chance finding, but instead of writing all of those words each time, we use the word “significantly” as a shorthand to the longer explanation. This makes it much easier for the reader to understand the conclusion when it is written “in plain English,” instead of technical, statistical language.

4.3 Can You Use Either the 95% Confidence Interval About the Mean OR the One-Group t-Test When Testing Hypotheses?

You are probably asking yourself:

“It sounds like you could use *either* the 95% confidence interval about the mean *or* the one-group t-test to analyze the results of the types of problems described so far in this book? Is this a correct statement?”

The answer is a resounding: “*Yes!*”

Both the confidence interval about the mean and the one-group t-test are used often in social science research on the types of problems described so far in this book. *Both of these tests produce the same result and the same conclusion from the data set!*

Both of these tests are explained in this book because some researchers prefer the confidence interval about the mean test, others prefer the one-group t-test, and still others prefer to use both tests on the same data to make their results and conclusions clearer to the reader of their research reports. Since we do not know which of these tests your researcher prefers, we have explained both of them so that you are competent in the use of both tests in the analysis of statistical data.

Now, let’s try your Excel skills on the one-group t-test on these three problems at the end of this chapter.

4.4 End-of-Chapter Practice Problems

1. Subaru of America rates the customer satisfaction of its dealers on a weekly basis on its Purchase Experience Survey, and demands that dealers achieve a 93% satisfaction score, or the dealers are required to take additional training to improve their customer satisfaction scores. Suppose that you have selected a random sample of rating forms submitted by new car purchasers (either online or through the mail) for the St. Louis Subaru dealer in the past week and that you have prepared the hypothetical table in Fig. 4.5 for Question #1d:

SUBARU Customer Satisfaction Survey						
Question #1d: The salesperson was knowledgeable about the Subaru model line.						
1	2	3	4	5	6	7
Completely Disagree						Completely Agree
	Rating					
	5					
	7					
	6					
	4					
	3					
	5					
	6					
	7					
	2					
	3					
	5					
	7					
	4					
	7					
	7					
	5					
	6					
	6					
	4					
	3					
	5					
	5					

Fig. 4.5 Worksheet Data for Chap. 4: Practice Problem #1

- (a) Write the null hypothesis and the research hypothesis on your spreadsheet
- (b) Use Excel to find the sample size, mean, standard deviation, and standard error of the mean to the right of the data set. Use number format (two decimal places) for the mean, standard deviation, and standard error of the mean.
- (c) Enter the critical t from the t-table in Appendix E onto your spreadsheet, and label it.
- (d) Use Excel to compute the t-value for these data (use two decimal places) and label it on your spreadsheet
- (e) Type the result on your spreadsheet, and then type the conclusion in plain English on your spreadsheet
- (f) Save the file as: subaru4

- (a) On your Excel spreadsheet, write the null hypothesis and the research hypothesis for these data.
 - (b) Use Excel to find the *sample size, mean, standard deviation, and standard error of the mean* for these data (two decimal places for the mean, standard deviation, and standard error of the mean).
 - (c) Use Excel to perform a *one-group t-test* on these data (two decimal places).
 - (d) On your printout, type the *critical value of t* given in your t-table in Appendix E.
 - (e) On your spreadsheet, type the *result* of the t-test.
 - (f) On your spreadsheet, type the *conclusion* of your study in plain English.
 - (g) save the file as: calculus4
3. Boston University (BU) in Boston, MA US offers a graduate program for an M.S. degree in Advertising. One of the courses in this program focuses on Advertising Management. Suppose that you have been hired as a consultant by BU to analyze the student evaluation data for this course from the previous semester, and that you have created the hypothetical data for Question #12 in Fig. 4.7.

BOSTON UNIVERSITY M.S. IN ADVERTISING PROGRAM							
Course: Advertising Management							
Item #12:	"How would you rate the instructor's ability to explain advertising concepts clearly?"						
	1	2	3	4	5	6	7
	Poor						Excellent
	RATING						
	5						
	6						
	4						
	7						
	6						
	5						
	7						
	6						
	7						
	6						
	7						
	6						
	7						
	5						
	6						
	7						
	6						
	7						

Fig. 4.7 Worksheet Data for Chap. 4: Practice problem #3

- (a) Write the null hypothesis and the research hypothesis on your spreadsheet
- (b) Use Excel to find the sample size, mean, standard deviation, and standard error of the mean to the right of the data set. Use number format (two decimal places) for the mean, standard deviation, and standard error of the mean.
- (c) Enter the critical t from the t-table in Appendix E onto your spreadsheet, and label it.
- (d) Use Excel to compute the t-value for these data (use two decimal places) and label it on your spreadsheet
- (e) Type the result on your spreadsheet, and then type the conclusion in plain English on your spreadsheet
- (f) Save the file as: COURSE3

References

- Zikmund, W.G. and Babin, B.J. Exploring Marketing Research (10th ed.) Mason, OH: South-Western Cengage Learning, 2010.
- Foster, D.P., Stine, R.A., Waterman, R.P. Basic Business Statistics: A Casebook. New York, NY: Springer-Verlag, 1998.

Chapter 5

Two-Group t-Test of the Difference of the Means for Independent Groups



Up until now in this book, you have been dealing with the situation in which you have had only one group of people in your research study and only one measurement “number” on each of these people. We will now change gears and deal with the situation in which you are measuring two groups of people instead of only one group of people.

Whenever you have two completely different groups of people (i.e., no one person is in both groups, but every person is measured on only one variable to produce one “number” for each person), we say that the two groups are “independent of one another.” This chapter deals with just that situation and that is why it is called the two-group t-test for independent groups.

The two assumptions underlying the two-group t-test are the following (Zikmund and Babin 2010): (1) both groups are sampled from a normal population, and (2) the variances of the two populations are approximately equal. Note that the standard deviation is merely the square root of the variance. (There are different formulas to use when each person is measured twice to create two groups of data, and this situation is called “dependent,” but those formulas are beyond the scope of this book.) This book only deals with two groups that are independent of one another so that no person is in both groups of data.

When you are testing for the difference between the means for two groups, it is important to remember that there are two different formulas that you need to use depending on the sample sizes of the two groups:

- (1) Use Formula #1 in this chapter when both of the groups have more than 30 people in them, and
- (2) Use Formula #2 in this chapter when either one group, or both groups, have sample sizes less than 30 people in them.

We will illustrate both of these situations in this chapter.

But, first, we need to understand the steps involved in hypothesis-testing when two groups of people are involved before we dive into the formulas for this test.

5.1 The 9 STEPS for Hypothesis-Testing Using the Two-Group t-Test

Objective: To learn the 9 steps of hypothesis-testing using two groups of people and the two-group t-test

You will see that these steps parallel the steps used in the previous chapter that dealt with the one-group t-test, but there are some important differences between the steps that you need to understand clearly before we dive into the formulas for the two-group t-test.

5.1.1 *STEP 1: Name One Group, Group 1, and the Other Group, Group 2*

The formulas used in this chapter will use the subscripts 1 and 2 to distinguish between the two groups. If you define which group is Group 1 and which group is Group 2, you can use these subscripts in your computations without having to write out the names of the groups.

For example, if you are testing teenage boys on their preference for the taste of Coke or Pepsi, you could call the groups: “Coke” and “Pepsi.” but this would require your writing out the words “Coke” or “Pepsi” whenever you wanted to refer to one of these groups. If you call the Coke group, Group 1, and the Pepsi group, Group 2, this makes it much easier to refer to the groups because it saves you writing time.

As a second example, you could be comparing the test market results for Kansas City versus Indianapolis, but if you had to write out the names of those cities whenever you wanted to refer to them, it would take you more time than it would if, instead, you named one city, Group 1, and the other city, Group 2.

Note, also, that it is completely arbitrary which group you call Group 1, and which Group you call Group 2. You will achieve the same result and the same conclusion from the formulas however you decide to define these two groups.

5.1.2 *STEP 2: Create a Table That Summarizes the Sample Size, Mean Score, and Standard Deviation of Each Group*

This step makes it easier for you to make sure that you are using the correct numbers in the formulas for the two-group t-test. If you get the numbers “mixed-up,” your entire formula work will be incorrect and you will botch the problem terribly.

For example, suppose that you tested teenage boys on their preference for the taste of Coke versus Pepsi in which the boys were randomly assigned to taste just one of these brands and then rate its taste on a 100-point scale from 0 = poor to 100 = excellent. After the research study was completed, suppose that the Coke group had 52 boys in it, their mean taste rating was 55 with a standard deviation of 7, while the Pepsi group had 57 boys in it and their average taste rating was 64 with a standard deviation of 13.

The formulas for analyzing these data to determine if there was a significant different in the taste rating for teenage boys for these two brands require you to use six numbers correctly in the formulas: the sample size, the mean, and the standard deviation of each of the two groups. All six of these numbers must be used correctly in the formulas if you are to analyze the data correctly.

If you create a table to summarize these data, a good example of the table, using both Step 1 and Step 2, would be the data presented in Fig. 5.1:

Fig. 5.1 Basic Table Format for the Two-group t-test

	A	B	C	D	E	F
1						
2						
3		Group	n	Mean	STDEV	
4		1 (name it)				
5		2 (name it)				
6						
7						

For example, if you decide to call Group 1 the Coke group and Group 2 the Pepsi group, the following table would place the six numbers from your research study into the proper cells of the table as in Fig. 5.2:

Fig. 5.2 Results of Entering the Data Needed for the Two-group t-test

	A	B	C	D	E	F
1						
2						
3		Group	n	Mean	STDEV	
4		1 (name it)	52	55	7	
5		2 (name it)	57	64	13	
6						
7						

You can now use the formulas for the two-group t-test with more confidence that the six numbers will be placed in the proper place in the formulas.

Note that you could just as easily call Group 1 the Pepsi group and Group 2 the Coke group; it makes no difference how you decide to name the two groups; this decision is up to you.

5.1.3 STEP 3: State the Null Hypothesis and the Research Hypothesis for the Two-Group t-Test

If you have completed Step 1 above, this step is very easy because the null hypothesis and the research hypothesis will always be stated in the same way for the two-group t-test. The null hypothesis states that the population means of the two groups are equal, while the research hypothesis states that the population means of the two groups are not equal. In notation format, this becomes:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

You can now see that this notation is much simpler than having to write out the names of the two groups in all of your formulas.

5.1.4 STEP 4: Select the Appropriate Statistical Test

Since this chapter deals with the situation in which you have two groups of people but only one measurement on each person in each group, we will use the two-group t-test throughout this chapter.

5.1.5 STEP 5: Decide on a Decision Rule for the Two-Group t-Test

The decision rule is exactly what it was in the previous chapter (see Sect. 4.1.3) when we dealt with the one-group t-test.

- (a) If the absolute value of t is less than the critical value of t , accept the null hypothesis.
- (b) If the absolute value of t is greater than the critical value of t , reject the null hypothesis and accept the research hypothesis.

Since you learned how to find the absolute value of t in the previous chapter (see Sect. 4.1.3.1), you can use that knowledge in this chapter.

5.1.6 STEP 6: Calculate the Formula for the Two-Group t-Test

Since we are using two different formulas in this chapter for the two-group t-test depending on the sample size of the people in the two groups, we will explain how to use those formulas later in this chapter.

5.1.7 STEP 7: Find the Critical Value of t in the t -Table in Appendix E

In the previous chapter where we were dealing with the one-group t -test, you found the critical value of t in the t -table in Appendix E by finding the sample size for the one group of people in the first column of the table, and then reading the critical value of t across from it on the right in the “critical t column” in the table (see Sect. 4.1.5). This process was fairly simple once you have had some practice in doing this step.

However, for the two-group t -test, the procedure for finding the critical value of t is more complicated because you have two different groups of people in your study, and they often have different sample sizes in each group.

To use Appendix E correctly in this chapter, you need to learn how to find the “degrees of freedom” for your study. We will discuss that process now.

5.1.7.1 Finding the Degrees of Freedom (df) for the Two-Group t -Test

Objective: To find the degrees of freedom for the two-group t -test and to use it to find the critical value of t in the t -table in Appendix E

The mathematical explanation of the concept of the “degrees of freedom” is beyond the scope of this book, but you can find out more about this concept by reading any good statistics book (e.g. Keller, 2009). For our purposes, you can easily understand how to find the degrees of freedom and to use it to find the critical value of t in Appendix E. The formula for the degrees of freedom (df) is:

$$\text{degrees of freedom} = df = n_1 + n_2 - 2 \quad (5.1)$$

In other words, you add the sample size for Group 1 to the sample size for Group 2 and then subtract 2 from this total to get the number of degrees of freedom to use in Appendix E.

Take a look at Appendix E.

Instead of using the first column as we did in the one-group t -test that is based on the sample size, n , of one group of people, we need to use the second-column of this table (df) to find the critical value of t for the two-group t -test.

For example, if you had 13 people in Group 1 and 17 people in Group 2, the degrees of freedom would be: $13 + 17 - 2 = 28$, and the critical value of t would be 2.048 *since you look down the second column which contains the degrees of freedom until you come to the number 28, and then read 2.048 in the “critical t column” in the table to find the critical value of t when $df = 28$.*

As a second example, if you had 52 people in Group 1 and 57 people in Group 2, the degrees of freedom would be: $52 + 57 - 2 = 107$ When you go down the

second column in Appendix E for the degrees of freedom, you find that *once you go beyond the degrees of freedom equal to 39, the critical value of t is always 1.96*, and that is the value you would use for the critical t with this example.

5.1.8 STEP 8: State the Result of Your Statistical Test

The result follows the exact same result format that you found for the one-group t -test in the previous chapter (see Sect. 4.1.6):

Either: Since the absolute value of t that you found in the t -test formula is *less than the critical value of t* in Appendix E, you accept the null hypothesis.

Or: Since the absolute value of t that you found in the t -test formula is *greater than the critical value of t* in Appendix E, you reject the null hypothesis and accept the research hypothesis.

5.1.9 STEP 9: State the Conclusion of Your Statistical Test in Plain English!

Writing the conclusion for the two-group t -test is more difficult than writing the conclusion for the one-group t -test because you have to decide what the difference was between the two groups.

When you accept the null hypothesis, the conclusion is simple to write: “There is no difference between the two groups in the variable that was measured.”

But when you reject the null hypothesis and accept the research hypothesis, you need to be careful about writing the conclusion so that it is both accurate and concise.

Let’s give you some practice in writing the conclusion of a two-group t -test.

5.1.9.1 Writing the Conclusion of the Two-Group t -Test When You Accept the Null Hypothesis

Objective: To write the conclusion of the two-group t -test when you have accepted the null hypothesis.

Suppose that you have been hired as a statistical consultant by Marriott Hotel in St. Louis to analyze the data from a Guest Satisfaction Survey that they give to all customers to determine the degree of satisfaction of these customers for various activities of the hotel.

The survey contains a number of items, but suppose Item #7 is the one in Fig. 5.3:

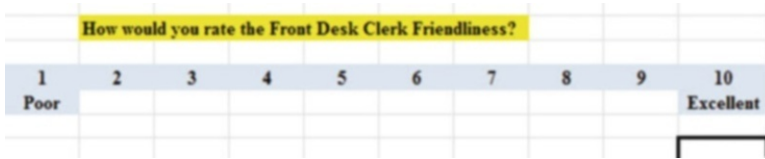


Fig. 5.3 Marriott Hotel Guest Satisfaction Survey Item #7

Suppose further, that you have decided to analyze the data from last week’s customers comparing men and women using the two-group t-test.

Important note: You would need to use this test for each of the survey items separately.

Suppose that the hypothetical data for Item #7 from last week at the St. Louis Marriott Hotel were based on a sample size of 124 men who had a mean score on this item of 6.58 and a standard deviation on this item of 2.44. Suppose that you also had data from 86 women from last week who had a mean score of 6.45 with a standard deviation of 1.86.

We will explain later in this chapter how to produce the results of the two-group t-test using its formulas, but, for now, let’s “cut to the chase” and tell you that those formulas would produce the following in Fig. 5.4:

Fig. 5.4 Worksheet Data for Males vs. Females for the St. Louis Marriott Hotel for Accepting the Null Hypothesis

	A	B	C	D	E	F
1						
2						
3		Group	n	Mean	STDEV	
4		1 Males	124	6.58	2.44	
5		2 Females	86	6.45	1.86	
6						
7						

- degrees of freedom: 208
- critical t: 1.96 (in Appendix E)
- t-test formula: 0.44 (when you use your calculator!)
- Result: Since the absolute value of 0.44 is less than the critical t of 1.96, we accept the null hypothesis.
- Conclusion: There was no difference between male and female guests last week in their rating of the friendliness of the front-desk clerk at the St. Louis Marriott Hotel.

Now, let’s see what happens when you reject the null hypothesis (H_0) and accept the research hypothesis (H_1).

5.1.9.2 Writing the Conclusion of the Two-Group t-Test When You Reject the Null Hypothesis and Accept the Research Hypothesis

Objective: To write the conclusion of the two-group t-test when you have rejected the null hypothesis and accepted the research hypothesis

Let’s continue with this same example of the Marriott Hotel, but with the result that we reject the null hypothesis and accept the research hypothesis.

Let’s assume that this time you have data on 85 males from last week and their mean score on this question was 7.26 with a standard deviation of 2.35. Let’s further suppose that you also have data on 48 females from last week and their mean score on this question was 4.37 with a standard deviation of 3.26.

Without going into the details of the formulas for the two-group t-test, these data would produce the following result and conclusion based on Fig. 5.5:

Fig. 5.5 Worksheet Data for St. Louis Marriott Hotel for Obtaining a Significant Difference between Males and Females

	A	B	C	D	E
1					
2					
3		Group	n	Mean	STDEV
4		1 Males	85	7.26	2.35
5		2 Females	48	4.37	3.26
6					
7					

Null Hypothesis: $\mu_1 = \mu_2$
 Research Hypothesis: $\mu_1 \neq \mu_2$
 degrees of freedom: 131
 critical t: 1.96 (in Appendix E)
 t-test formula: 5.40 (when you use your calculator!)
 Result: Since the absolute value of 5.40 is greater than the critical t of 1.96, we reject the null hypothesis and accept the research hypothesis.

Now, you need to compare the ratings of the men and women to find out which group had the more positive rating of the friendliness of the front-desk clerk using the following rule:

Rule: To summarize the conclusion of the two-group t-test, just compare the means of the two groups, and be sure to use the word “significantly” in your conclusion if you rejected the null hypothesis and accepted the research hypothesis.

A good way to prepare to write the conclusion of the two-group t-test when you are using a rating scale is to place the mean scores of the two groups on a drawing of the scale so that you can visualize the difference of the mean scores. For example, for

our Marriott Hotel example above, you would draw this “picture” of the scale in Fig. 5.6:

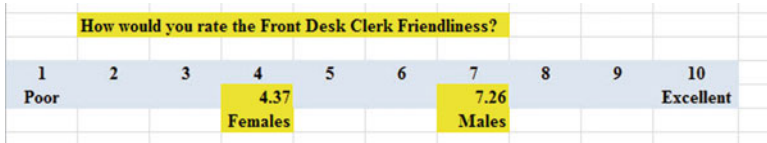


Fig. 5.6 Example of Drawing a “Picture” of the Means of the Two Groups on the Rating Scale

This drawing tells you visually that males had a higher positive rating than females on this item (7.26 vs. 4.37). *And, since you rejected the null hypothesis and accepted the research hypothesis, you know that you have found a significant difference between the two mean scores.*

So, our conclusion needs to contain the following key words:

- Male guests
- Female guests
- Marriott Hotel
- St. Louis
- last week
- significantly
- Front Desk Clerks
- more friendly *or* less friendly
- *either* (7.26 vs. 4.37) *or* (4.37 vs. 7.26)

We can use these key words to write the either of two conclusions which are *logically identical*:

Either: Male guests at the Marriott Hotel in St. Louis last week rated the Front Desk Clerks as significantly more friendly than female guests (7.26 vs. 4.37).

Or: Female guests at the Marriott Hotel in St. Louis last week rated the Front Desk Clerks as significantly less friendly than male guests (4.37 vs. 7.26).

Both of these conclusions are accurate, so you can decide which one you want to write. It is your choice.

Also, note that the mean scores in parentheses at the end of these conclusions must match the sequence of the two groups in your conclusion. For example, if you say that: “Male guests rated the Front Desk Clerks as significantly more friendly than female guests,” the end of this conclusion should be: (7.26 vs. 4.37) since you mentioned males first and females second.

Alternately, if you wrote that: “Female guests rated the Front Desk Clerks as significantly less friendly than male guests,” the end of this conclusion should be: (4.37 vs. 7.26) since you mentioned females first and males second.

Putting the two mean scores at the end of your conclusion saves the reader from having to turn back to the table in your research report to find these mean scores to see how far apart the mean scores were.

Now, let's discuss FORMULA #1 that deals with the situation in which both groups have more than 30 people in them.

Objective: To use FORMULA #1 for the two-group t-test when both groups have a sample size greater than 30 people

5.2 Formula #1: Both Groups Have More Than 30 People in Them

The first formula we will discuss will be used when you have two groups of people with more than 30 people in each group and one measurement on each person in each group. This formula for the two-group t-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} \quad (5.2)$$

$$\text{where } S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad (5.3)$$

$$\text{and where degrees of freedom} = df = n_1 + n_2 - 2. \quad (5.1)$$

This formula looks daunting when you first see it, but let's explain some of the parts of this formula:

We have explained the concept of "degrees of freedom" earlier in this chapter, and so you should be able to find the degrees of freedom needed for this formula in order to find the critical value of t in Appendix E.

In the previous chapter, *the formula for the one-group t-test was the following:*

$$t = \frac{\bar{X} - \mu}{S_{\bar{X}}} \quad (4.1)$$

$$\text{where s.e.} = S_{\bar{X}} = \frac{S}{\sqrt{n}} \quad (4.2)$$

For the one-group t-test, you found the mean score and subtracted the population mean from it, and then divided the result by the standard error of the mean (s.e.) to get the result of the t-test. You then compared the t-test result to the critical value of t to see if you either accepted the null hypothesis, or rejected the null hypothesis and accepted the research hypothesis.

The two-group t-test requires a different formula because you have two groups of people, each with a mean score on some variable. You are trying to determine

whether to accept the null hypothesis that the *population means of the two groups are equal* (in other words, there is no difference statistically between these two means), or whether the difference between the means of the two groups is “sufficiently large” that you would accept *that there is a significant difference* in the mean scores of the two groups.

The numerator of the two-group t-test asks you to find the difference of the means of the two groups:

$$\bar{X}_1 - \bar{X}_2 \quad (5.4)$$

The next step in the formula for the two-group t-test is to divide the answer you get when you subtract the two means by the standard error of the difference of the two means, and *this is a different standard error of the mean that you found for the one-group t-test because there are two means in the two-group t-test.*

The standard error of the mean when you have two groups of people is called the “standard error of the difference of the means” between the means of the two groups. This formula looks less scary when you break it down into four steps:

1. Square the standard deviation of Group 1, and divide this result by the sample size for Group 1 (n_1).
2. Square the standard deviation of Group 2, and divide this result by the sample size for Group 2 (n_2).
3. Add the results of the above two steps to get a total score.
4. *Take the square root of this total score* to find the standard error of the difference

of the means between the two groups, $S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

This last step is the one that gives students the most difficulty when they are finding this standard error using their calculator, because they are in such a hurry to get to the answer that they forget to carry the square root sign down to the last step, and thus get a larger number than they should for the standard error.

5.2.1 An example of Formula #1 for the Two-Group t-Test

Now, let’s use Formula #1 in a situation in which both groups have a sample size greater than 30 people.

Suppose that you have been hired by PepsiCo to do a taste test with teenage boys (ages 13–18) to determine if they like the taste of Pepsi the same as the taste of Coke. The boys are not told the brand name of the soft drink that they taste.

You select a group of boys in this age range, and randomly assign them to one of two groups: (1) Group 1 tastes Coke, and (2) Group 2 tastes Pepsi. Each group rates the taste of their soft drink on a 100-point scale using the following scale in Fig. 5.7:

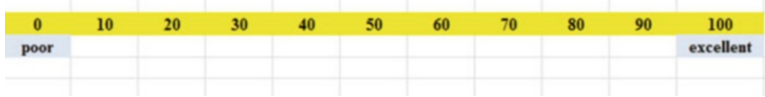


Fig. 5.7 Example of a Rating Scale for a Soft Drink Taste Test (Practical Example)

Suppose you collect these ratings and determine (using your new Excel skills) that the 52 boys in the Coke group had a mean rating of 55 with a standard deviation of 7, while the 57 boys in the Pepsi group had a mean rating of 64 with a standard deviation of 13.

Note that the two-group t-test does not require that both groups have the same sample size. This is another way of saying that the two-group t-test is “robust” (a fancy term that statisticians like to use).

Your data then produce the following table in Fig. 5.8:

Fig. 5.8 Worksheet Data for Soft Drink Taste Test

	A	B	C	D	E
1					
2					
3		Group	n	Mean	STDEV
4		1 Coke	52	55	7
5		2 Pepsi	57	64	13

Create an Excel spreadsheet, and enter the following information:

- B3: Group
- B4: 1 Coke
- B5: 2 Pepsi
- C3: n
- D3: Mean
- E3: STDEV
- C4: 52
- D4: 55
- E4: 7
- C5: 57
- D5: 64
- E5: 13

Now, widen column B so that it is twice as wide as column A, and center the six numbers and their labels in your table (see Fig. 5.9)

Fig. 5.9 Results of Widening Column B and Centering the Numbers in the Cells

	A	B	C	D	E
1					
2					
3		Group	n	Mean	STDEV
4		1 Coke	52	55	7
5		2 Pepsi	57	64	13
6					

B8: Null hypothesis:

B10: Research hypothesis:

Since both groups have a sample size greater than 30, you need to use Formula #1 for the t-test for the difference of the means of the two groups.

Let’s “break this formula down into pieces” to reduce the chance of making a mistake.

B13: $STDEV1^2/n1$ (note that you square the standard deviation of Group 1, and then divide the result by the sample size of Group 1)

B16: $STDEV2^2/n2$

B19: $D13 + D16$

B22: s.e.

B25: critical t

B28: t-test

B31: Result:

B36: Conclusion: (see Fig. 5.10)

Fig. 5.10 Formula Labels for the Two-group t-test

Group	n	Mean	STDEV
1 Coke	52	55	7
2 Pepsi	57	64	13
Null hypothesis:			
Research hypothesis:			
STDEV1 squared / n1			
STDEV2 squared / n2			
D13 + D16			
s.e.			
critical t			
t-test			
Result:			
Conclusion:			

You now need to compute the values of the above formulas in the following cells:

- D13: the result of the formula needed to compute cell B13 (use two decimals)
- D16: the result of the formula needed to compute cell B16 (use two decimals)
- D19: the result of the formula needed to compute cell B19 (use two decimals)
- D22: =SQRT(D19) (use two decimals)

This formula should give you a standard error (s.e.) of 1.98.

D25: 1.96

(Since $df = n1 + n2 - 2$, this gives $df = 109 - 2 = 107$, and the critical t is, therefore, 1.96 in Appendix E.)

D28: $= (D4 - D5) / D22$ (use two decimals) (no spaces between symbols)

This formula should give you a value for the t-test of: -4.55 .

Next, check to see if you have rounded off all figures in D13: D28 to two decimal places (see Fig. 5.11).

Fig. 5.11 Results of the t-test Formula for the Soft Drink Taste Test

Group	n	Mean	STDEV
1 Coke	52	55	7
2 Pepsi	57	64	13
Null hypothesis:			
Research hypothesis:			
STDEV1 squared / n1		0.94	
STDEV2 squared / n2		2.96	
D13 + D16		3.91	
s.e.		1.98	
critical t		1.96	
t-test		-4.55	
Result:			
Conclusion:			

Now, write the following sentence in D31 to D34 to summarize the result of the study:

- D31: Since the absolute value of -4.55
D32: is greater than the critical t of
D33: 1.96 , we reject the null hypothesis
D34: and accept the research hypothesis.

Finally, write the following sentence in D36 to D38 to summarize the conclusion of the study in plain English:

- D36: Teenage boys rated the taste of
D37: Pepsi as significantly better than
D38: the taste of Coke (64 vs. 55).

Save your file as: COKE4

Important note: You are probably wondering why we entered both the result and the conclusion in separate cells instead of in just one cell. This is because if you enter them in one cell, you will be very disappointed when you print out your final spreadsheet, because one of two things will happen that you will not like: (1) if you print the spreadsheet to fit onto only one page, the result and the conclusion will force the entire spreadsheet to be printed in such small font size that you will be unable to read it, or (2) if you do not print the final spreadsheet to fit onto one page, both the result and the conclusion will “dribble over” onto a second page instead of fitting the entire spreadsheet onto one page. In either case, your spreadsheet will not have a “professional look.”

Print this file so that it fits onto one page, and write by hand the null hypothesis and the research hypothesis on your printout.

The final spreadsheet appears in Fig. 5.12.

Group	n	Mean	STDEV
1 Coke	52	55	7
2 Pepsi	57	64	13
Null hypothesis:		$\mu_1 = \mu_2$	
Research hypothesis:		$\mu_1 \neq \mu_2$	
STDEV1 squared / n1		0.94	
STDEV2 squared / n2		2.96	
D13 + D16		3.91	
s.e.		1.98	
critical t		1.96	
t-test		-4.55	
Result:		Since the absolute value of -4.55 is greater than the critical t of 1.96, we reject the null hypothesis and accept the research hypothesis.	
Conclusion:		Teenage boys rated the taste of Pepsi as significantly better than the taste of Coke (64 vs. 55)	

Fig. 5.12 Final Worksheet for the Coke vs. Pepsi Taste Test

Now, let’s use the second formula for the two-group t-test which we use whenever either one group, or both groups, have less than 30 people or events in them.

Objective: To use Formula #2 for the two-group t-test when one or both groups have less than 30 events in them

Now, let’s look at the case when one or both groups have a sample size less than 30 people in them.

5.3 Formula #2: One or Both Groups Have Less Than 30 Events in Them

Suppose that you are an electrical engineer and you have been asked to compare the number of hours until failure of two new models of light bulbs (Model A and Model B) that have been prepared by your company’s Research & Development department. Suppose, further, that you have decided to analyze the data from this study using the two-group t-test for independent samples. You decide to try out your new Excel skills on a small sample of light bulbs of each model on the hypothetical data given in Fig. 5.13:

Fig. 5.13 Worksheet Data for Light Bulbs (Practical Example)

LIGHT BULB HOURS (hrs) UNTIL FAILURE	
Model A	Model B
910	890
940	940
980	950
1005	960
842	913
836	908
869	1030
910	1050
930	1040
897	
864	

Let’s call Model A as Group 1, and Model B as Group 2.

Null hypothesis: $\mu_1 = \mu_2$

Research hypothesis: $\mu_1 \neq \mu_2$

Note: Since both groups have a sample size less than 30, you need to use Formula #2 in the following steps:

Create an Excel spreadsheet, and enter the following information:

B2: LIGHT BULB HOURS (hrs) UNTIL FAILURE
B4: Model A
C4: Model B
B5: 910
B15: 864
C5: 890
C13: 1040

Now, enter the other figures into this table. Be sure to double-check all of your figures. If you have only one incorrect figure, you will not be able to obtain the correct answer to this problem.

Now, widen columns B and C so that all of the information fits inside the cells. To do this, click on both letters B and C at the top of these columns on your spreadsheet to highlight all of the cells in columns B and C. Then, move the mouse pointer to the right end of the B cell until you get a “cross” sign; then, click on this cross sign and drag the sign to the right until you can read all of the words on your screen. Then, stop clicking! Both Column B and Column C should now be the same width.

Then, center all information in the table except the top title by using the following steps:

Left-click your mouse and highlight cells B4:C15. Then, click on the bottom line, second from the left icon, under “Alignment” at the top-center of Home. All of the information in the table should now be in the center of each cell.

E5: Null hypothesis:
E7: Research hypothesis:
E9: Group
E10: 1 Model A
E11: 2 Model B
F9: n
G9: Mean
H9: STDEV

Your spreadsheet should now look like Fig. 5.14.

LIGHT BULB HOURS (hrs) UNTIL FAILURE				
Model A	Model B			
910	890	Null hypothesis:		
940	940			
980	950	Research hypothesis:		
1005	960			
842	913	Group	n	Mean
836	908	1 Model A		STDEV
869	1030	2 Model B		
910	1050			
930	1040			
897				
864				

Fig. 5.14 Light Bulb Hours Until Failure Worksheet Data for Hypothesis Testing

Now you need to use your Excel skills from Chap. 1 to fill in the sample sizes (n), the Means, and the Standard Deviations (STDEV) in the Table in cells F10:H11. Be sure to double-check your work or you will not be able to obtain the correct answer to this problem if you have only one incorrect figure! Round off the means and standard deviations to zero decimal places and center these six figures within their cells.

Since both groups have a sample size less than 30, you need to use Formula #2 for the t-test for the difference of the means of two independent samples.

Formula #2 for the two-group t-test is the following:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}. \quad (5.1)$$

$$\text{where } S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}. \quad (5.5)$$

$$\text{and where degrees of freedom} = df = n_1 + n_2 - 2. \quad (5.6)$$

This formula is complicated, and so it will reduce your chance of making a mistake in writing it if you “break it down into pieces” instead of trying to write the formula as one cell entry.

Now, enter these words on your spreadsheet:

E14: $(n_1 - 1) \times \text{STDEV1 squared}$

E16: $(n_2 - 1) \times \text{STDEV2 squared}$

E18: $n_1 + n_2 - 2$

E20: $1/n_1 + 1/n_2$

Note the three open-parentheses after *SQRT*, and the three closed parentheses on the right side of this formula. You need three open parentheses and three closed parentheses in this formula or the formula will not work correctly.

The above formula gives a standard error of the difference of the means equal to 25.68 (two decimals) in cell H23.

H26: Enter the critical t value from the t-table in Appendix E in this cell using $df = n_1 + n_2 - 2$ to find the critical t value

H29: $=(G10-G11)/H23$ (no spaces between symbols)

Note that you need an open-parenthesis *before* *G10* and a closed-parenthesis *after* *G11* so that this answer of -57 is *THEN* divided by the standard error of the difference of the means of 25.68, to give a t-test value of -2.22 . Use two decimal places for the t-test result (see Fig. 5.16).

LIGHT BULB HOURS (hrs) UNTIL FAILURE						
Model A	Model B					
910	890	Null hypothesis:				
940	940	Research hypothesis:				
980	950					
1005	960					
842	913	Group	n	Mean	STDEV	
836	908	1 Model A	11	908	54	
869	1030	2 Model B	9	965	61	
910	1050					
930	1040					
897		$(n1 - 1) \times STDEV1 \text{ squared}$				29224.73
864		$(n2 - 1) \times STDEV2 \text{ squared}$				29526.22
$n1 + n2 - 2$					18	
$1/n1 + 1/n2$					0.20	
s.e.					25.68	
critical t					2.101	
t-test					-2.22	
Result:						
Conclusion:						

Fig. 5.16 Light Bulb Hours Two-group t-test Formula Results

Now write the following sentence in C32 to C33 to summarize the *result* of the study:

C32: Since the absolute value of -2.22 is greater than 2.101 ,

C33: we reject the null hypothesis and accept the research hypothesis.

Finally, write the following sentence in C36 to C37 to summarize the *conclusion* of the study:

C36: Model B lasted significantly more hours until failure than Model A

C37: (965 hours vs. 908 hours).

Save your file as: bulb3

Print the final spreadsheet so that it fits onto one page.

Write the null hypothesis and the research hypothesis by hand on your printout.

The final spreadsheet appears in Fig. 5.17.

LIGHT BULB HOURS (hrs) UNTIL FAILURE			
Model A	Model B		
910	890	Null hypothesis:	$\mu_1 = \mu_2$
940	940		
980	950	Research hypothesis:	$\mu_1 \neq \mu_2$
1005	960		
842	913	Group	n
836	908	1 Model A	11
869	1030	2 Model B	9
910	1050	Mean	STDEV
930	1040		
897		$(n1 - 1) \times \text{STDEV1 squared}$	29224.73
864		$(n2 - 1) \times \text{STDEV2 squared}$	29526.22
		$n1 + n2 - 2$	18
		$1/n1 + 1/n2$	0.20
		s.e.	25.68
		critical t	2.101
		t-test	-2.22
Result:	Since the absolute value of -2.22 is greater than 2.101 , we reject the null hypothesis and accept the research hypothesis.		
Conclusion:	Model B lasted significantly more hours until failure than Model A (965 hours vs. 908 hours).		

Fig. 5.17 Light Bulb Hours Final Spreadsheet

5.4 End-of-Chapter Practice Problems

1. Suppose that you have been asked by the Director of the MS in Advertising program at the University of Illinois-Urbana to “run the data” to see if there is a gender difference in cumulative grade-point averages (GPAs) of MS in advertising students who have completed all of the required courses for this degree. The Director has obtained the cooperation of the Registrar and has promised to keep the GPA information confidential. You want to test your Excel skills on some hypothetical data to make sure that you can do this analysis. You have already determined by your Excel analysis that the 17 Males had a mean GPA of 3.15 with a standard deviation of 0.42, while the 15 Females had a mean GPA of 3.45 with a standard deviation of 0.37.
 - (a) State the null hypothesis and the research hypothesis on an Excel spreadsheet. Then, create a table summarizing the sample size, the mean, and the standard deviation of the two groups.
 - (b) Find the standard error of the difference between the means using Excel
 - (c) Find the critical t value using Appendix E, and enter it on your spreadsheet.
 - (d) Perform a t-test on these data using Excel. What is the value of t that you obtain?
 - (e) State your result on your spreadsheet.
 - (f) State your conclusion in plain English on your spreadsheet.
 - (g) Save the file as: GPA13
2. Massachusetts Mutual Financial Group (2010) placed a full-page color ad in *The Wall Street Journal* in which it used a male model hugging a two-year old daughter. The ad had the headline and sub-headline:

WHAT IS THE SIGN OF A GOOD DECISION?

It’s knowing your life insurance can help provide income for retirement. And peace of mind until you get there.

Since the majority of the subscribers to *The Wall Street Journal* are men, an interesting research question would be the following:

Research question: “Does a male model in a magazine ad affect adult men’s or adult women’s willingness to learn more about how life insurance can provide income for retirement?”

Suppose that you have shown one group of adult males (ages 25–39) and one group of adult females (ages 25–39) a mockup of an ad such that both groups saw the ad with a male model. The ads were identical in copy format. The two groups were kept separate during the experiment and could not interact with one another.

At the end of a one-hour discussion of the mockup ad, the respondents were asked the question given in Fig. 5.18:

"How interested are you in learning more about how life insurance can provide income for retirement?"						
1	2	3	4	5	6	7
Not at all interested						Very interested

Fig. 5.18 Rating Scale Item for a Magazine Ad Interest Indicator (Practical Example)

The resulting hypothetical data for this question appear in Fig. 5.19:

Fig. 5.19 Worksheet Data for Chap. 5: Practice Problem #2

Magazine ad: Male model	
Men	Women
5	3
6	4
4	6
7	5
5	2
6	3
5	1
4	3
3	2
6	4
7	3
5	5
6	6
4	3
7	4
5	2
4	5
6	3
3	4
7	5
5	4
6	3
2	2
6	4
1	3
7	5
6	1
5	3
4	2
6	3
5	2
7	5
	3
	4

- (a) On your Excel spreadsheet, write the null hypothesis and the research hypothesis.
 - (b) Create a table that summarizes these data on your spreadsheet and use Excel to find the sample sizes, the means, and the standard deviations of the two groups in this table.
 - (c) Use Excel to find the standard error of the difference of the means.
 - (d) Use Excel to perform a two-group t-test. What is the value of t that you obtain (use two decimal places)?
 - (e) On your spreadsheet, type the critical value of t using the t -table in Appendix E.
 - (f) Type your result on the test on your spreadsheet.
 - (g) Type your conclusion in plain English on your spreadsheet.
 - (h) save the file as: lifeinsur12
3. Suppose that an automobile repair parts manufacturer/supplier wants to test the crash resistance of two brands of front-bumpers for 2-door passenger sedans (BRAND X and BRAND Y). The engineer in charge of this project has decided to test these bumpers on the most recent Honda Civics that are purposely crashed into a cement wall at a speed of 15 miles per hour (mph), and then to estimate the cost of repairs to the front-bumper after this test. The engineer has used her Excel skills so far to determine that BRAND X had 11 repairs with an average cost of \$1206 and a standard deviation of \$83.45, while BRAND Y had 11 repairs with an average cost of \$1333 and a standard deviation of \$45.89.
- (a) Write the null hypothesis and the research hypothesis.
 - (b) Create an Excel table that summarizes these data.
 - (c) Use Excel to find the standard error of the difference of the means.
 - (d) Use Excel to perform a *two-group t-test*. What is the value of t that you obtain (use two decimal places)?
 - (e) On your spreadsheet, type the *critical value of t* using the t -table in Appendix E.
 - (f) Type the *result* of the test on your spreadsheet.
 - (g) Type your *conclusion in plain English* on your spreadsheet.
 - (h) Save the file as: BUMPER3
 - (i) Print the final spreadsheet so that it fits onto one page.

References

- Keller, G. Statistics for Management and Economics (8th ed.). Mason, OH: South-Western Cengage Learning, 2009.
- Zikmund, W.G. and Babin, B.J. Exploring Marketing Research (10th ed.). Mason, OH: South-Western Cengage Learning, 2010.
- Mass Mutual Financial Group. What is the Sign of a Good Decision? (Advertisement) *The Wall Street Journal*, September 29, 2010, p. A22.

Chapter 6

Correlation and Simple Linear Regression



There are many different types of “correlation coefficients,” but the one we will use in this book is the Pearson product-moment correlation which we will call: r .

6.1 What Is a “Correlation?”

Basically, a correlation is a number between -1 and $+1$ that summarizes the relationship between two variables, which we will call X and Y .

A correlation can be either positive or negative. *A positive correlation means that as X increases, Y increases. A negative correlation means that as X increases, Y decreases.* In statistics books, this part of the relationship is called the *direction* of the relationship (i.e., it is either positive or negative).

The correlation also tells us the *magnitude* of the relationship between X and Y . As the correlation approaches closer to $+1$, we say that the relationship is *strong and positive*.

As the correlation approaches closer to -1 , we say that the relationship is *strong and negative*.

A zero correlation means that there is no relationship between X and Y . This means that neither X nor Y can be used as a predictor of the other.

A good way to understand what a correlation means is to see a “picture” of the scatterplot of points produced in a chart by the data points. Let’s suppose that you want to know if variable X can be used to predict variable Y . We will place *the predictor variable X on the x -axis* (the horizontal axis of a chart) and *the criterion variable Y on the y -axis* (the vertical axis of a chart). Suppose, further, that you have collected data given in the scatterplots below (see Fig. 6.1 through Fig. 6.6).

Figure 6.1 shows the scatterplot for a perfect positive correlation of $r = +1.0$. This means that you can perfectly predict each y -value from each x -value because the data points move “upward-and-to-the-right” along a perfectly-fitting straight line (see Fig. 6.1)

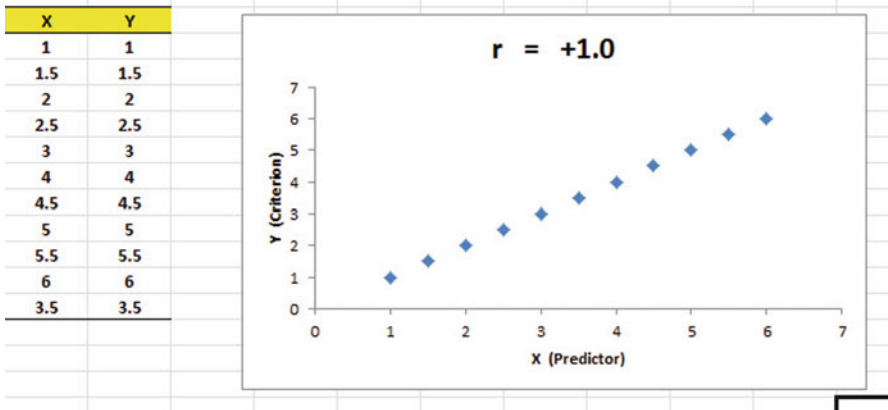


Fig. 6.1 Example of a Scatterplot for a Perfect, Positive Correlation ($r = +1.0$)

Figure 6.2 shows the scatterplot for a moderately positive correlation of $r = +.53$. This means that each x -value can predict each y -value moderately well because you can draw a picture of a “football” around the outside of the data points that move upward-and-to-the-right, but not along a straight line (see Fig. 6.2).

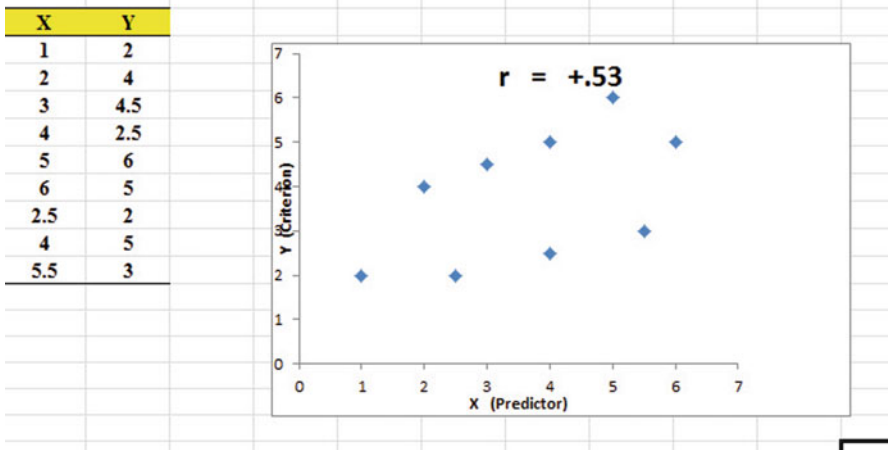


Fig. 6.2 Example of a Scatterplot for a Moderate, Positive Correlation ($r = +.53$)

Figure 6.3 shows the scatterplot for a low, positive correlation of $r = +.23$. This means that each x -value is a poor predictor of each y -value because the “picture” you could draw around the outside of the data points approaches a circle in shape (see Fig. 6.3)

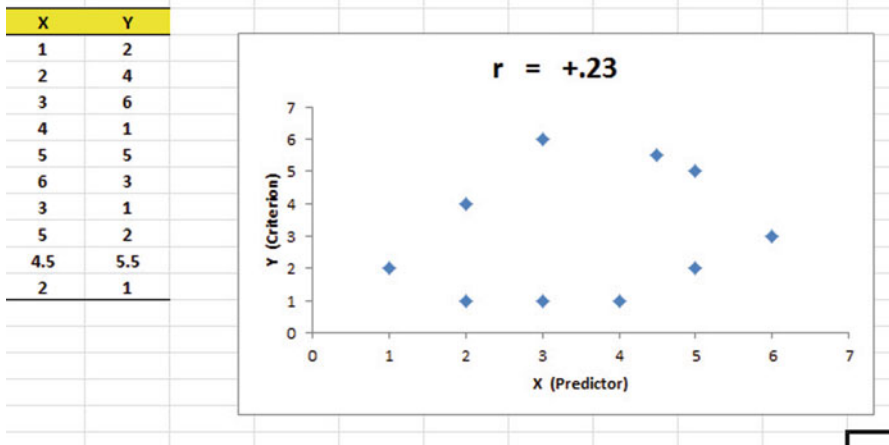


Fig. 6.3 Example of a Scatterplot for a Low, Positive Correlation ($r = +.23$)

We have not shown a Figure of a zero correlation because it is easy to imagine what it looks like as a scatterplot. A zero correlation of $r = .00$ means that there is no relationship between X and Y and the “picture” drawn around the data points would be a perfect circle in shape, indicating that you cannot use X to predict Y because these two variables are not correlated with one another.

Figure 6.4 shows the scatterplot for a low, negative correlation of $r = -.22$ which means that each X is a poor predictor of Y in an inverse relationship, meaning that as X increases, Y decreases (see Fig. 6.4). In this case, it is a negative correlation because the “football” you could draw around the data points slopes down and to the right.

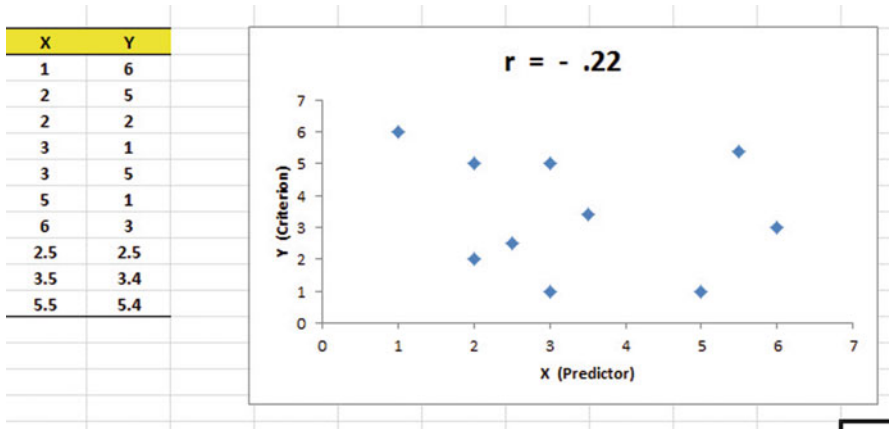


Fig. 6.4 Example of a Scatterplot for a Low, Negative Correlation ($r = -.22$)

Figure 6.5 shows the scatterplot for a moderate, negative correlation of $r = -.39$ which means that X is a moderately good predictor of Y, although there is an inverse relationship between X and Y (i.e., as X increases, Y decreases; see Fig. 6.5). In this case, it is a negative correlation because the “football” you could draw around the data points slopes down and to the right.

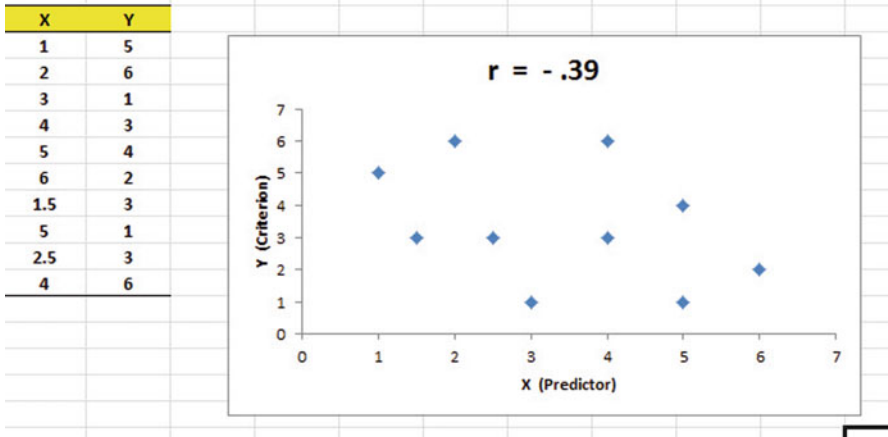


Fig. 6.5 Example of a Scatterplot for a Moderate, Negative Correlation ($r = -.39$)

Figure 6.6 shows a perfect negative correlation of $r = -1.0$ which means that X is a perfect predictor of Y, although in an inverse relationship such that as X increases, Y decreases. The data points fit perfectly along a downward-sloping straight line (see Fig. 6.6)

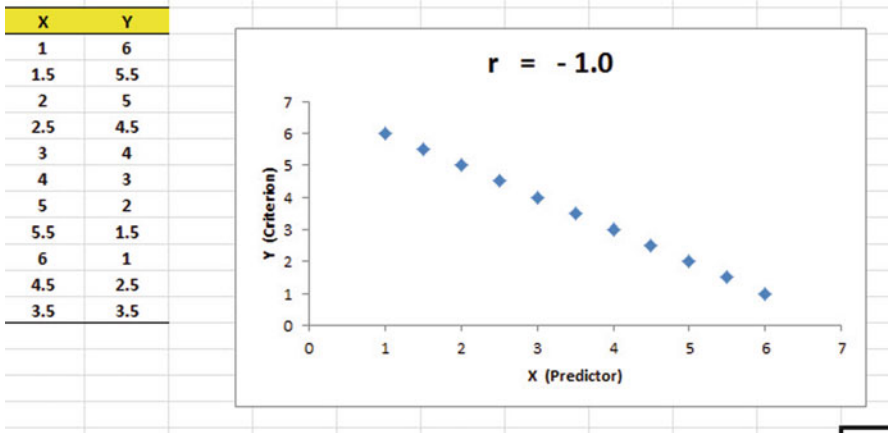


Fig. 6.6 Example of a Scatterplot for a Perfect, Negative Correlation ($r = -1.0$)

Let’s explain the formula for computing the correlation r so that you can understand where the number summarizing the correlation came from.

In order to help you to understand *where* the correlation number that ranges from -1.0 to $+1.0$ comes from, we will walk you through the steps involved to use the formula as if you were using a pocket calculator. This is the one time in this book that we will ask you to use your pocket calculator to find a correlation, but knowing how the correlation is computed step-by-step will give you the opportunity to understand *how* the formula works in practice.

To do that, let’s create a situation in which you need to find the correlation between two variables.

Suppose that you wanted to find out if there was a relationship between high school grade-point average (HSGPA) and freshman GPA (FROSH GPA) at a liberal arts college. You have decided to call HSGPA the x -variable (i.e., the predictor variable) and FROSH GPA as the y -variable (i.e., the criterion variable) in your analysis. To test your Excel skills, you take a random sample of freshmen at the end of their freshman year and record their GPA. The hypothetical data for eight students appear in Fig. 6.7. (*Note: We are using only one decimal place for these GPAs in this example to simplify the mathematical computations.*)

Fig. 6.7 Worksheet Data for High School GPA and Frosh GPA (Practical Example)

	A	B	C
1			
2		X	Y
3	Student	High School GPA	FROSH GPA
4	1	2.8	2.9
5	2	2.5	2.8
6	3	3.1	2.8
7	4	3.5	3.2
8	5	2.4	2.6
9	6	2.6	2.3
10	7	2.4	2.1
11	8	3.6	3.2
12			
13	n	8	8
14	MEAN	2.86	2.74
15	STDEV	0.48	0.39
16			

Notice also that we have used Excel to find the sample size for both variables, X and Y , and the MEAN and STDEV of both variables. (You can practice your Excel skills by seeing if you get this same results when you create an Excel spreadsheet for these data.)

Now, let’s use the above table to compute the correlation r between HSGPA and FROSH GPA using your pocket calculator.

6.1.1 Understanding the Formula for Computing a Correlation

Objective: To understand the formula for computing the correlation r

The formula for computing the correlation r is as follows:

$$r = \frac{\frac{1}{n-1} \sum (X - \bar{X})(Y - \bar{Y})}{S_x S_y} \tag{6.1}$$

This formula looks daunting at first glance, but let’s “break it down into its steps” to understand how to compute the correlation r .

6.1.2 Understanding the Nine Steps for Computing a Correlation, r

Objective: To understand the nine steps of computing a correlation r

The nine steps are as follows:

Step	Computation	Result
1	Find the sample size n by noting the number of students	8
2	Divide the number 1 by the sample size minus 1 (i.e., $1/7$)	0.14286
3	For each student, take the HSGPA and subtract the meanHSGPA for the 8 students and call this $X - \bar{X}$ (For example, for student # 6, this would be: $2.6 - 2.86$)	-0.26
	<i>Note: With your calculator, this difference is -0.26, but when Excel uses 16 decimal places for every computation, this result could be slightly different for each student</i>	
4	For each student, take the FROSH GPA and subtract the meanFROSH GPA for the 8 students and call this $Y - \bar{Y}$ (For example, for student # 6, this would be: $2.3 - 2.74$)	-0.44
5	Then, for each student, multiply $(X - \bar{X})$ times $(Y - \bar{Y})$ For example, for student # 6 this would be: $(-0.26) \times (-0.44)$	+0.1144
6	Add the results of $(X - \bar{X})$ times $(Y - \bar{Y})$ for the 8 students	+1.09

Steps 1–6 would produce the Excel table given in Fig. 6.8.

	A	B	C	D	E	F	G
1							
2		X	Y				
3	Student	High School GPA	FROSH GPA	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	
4	1	2.8	2.9	-0.06	0.16	-0.01	
5	2	2.5	2.8	-0.36	0.06	-0.02	
6	3	3.1	2.8	0.24	0.06	0.01	
7	4	3.5	3.2	0.64	0.46	0.29	
8	5	2.4	2.6	-0.46	-0.14	0.06	
9	6	2.6	2.3	-0.26	-0.44	0.11	
10	7	2.4	2.1	-0.46	-0.64	0.29	
11	8	3.6	3.2	0.74	0.46	0.34	
12						-----	
13	n	8	8		Total	1.09	
14	MEAN	2.86	2.74				
15	STDEV	0.48	0.39				
16							

Fig. 6.8 Worksheet for Computing the Correlation, r

Notice that when Excel multiplies a minus number by a minus number, the result is a plus number (for example for student #7: $(-0.46) \times (-0.64) = +0.29$. And when Excel multiplies a minus number by a plus number, the result is a negative number (for example for student #1: $(-0.06) \times (+0.16) = -0.01$).

Note: Excel computes all computation to 16 decimal places. So, when you check your work with a calculator, you frequently get a slightly different answer than Excel’s answer.

For example, when you compute above:

$$\begin{aligned}
 &(X - \bar{X}) \times (Y - \bar{Y}) \text{ for student \#2, your calculator gives:} \\
 &(-0.36) \times (+0.06) = -0.0216
 \end{aligned}
 \tag{6.2}$$

As you can see from the table, Excel’s answer is -0.02 which is really *more accurate* because Excel uses 16 decimal places for every number, even though only two decimal places are shown in Fig. 6.8.

You should also note that when you do Step 6, you have to be careful to add all of the positive numbers first to get $+1.10$ and then add all of the negative numbers second to get -0.03 , so that when you subtract these two numbers you get $+1.07$ as your answer to Step 6. When you do these computations using Excel, this total figure will be $+1.09$ because Excel carries every number and computation out to 16 decimal places which is much more accurate than your calculator.

Step		
7	Multiply the answer for step 2 above by the answer for step 6 (0.14286×1.09)	0.1557
8	Multiply the STDEV of X times the STDEV of Y (0.48×0.39)	0.1872
9	Finally, divide the answer from step 7 by the answer from step 8 (0.1557 divided by 0.1872)	+0.83

This number of 0.83 is the correlation between HSGPA (X) and FROSH GPA (Y) for these 8 students. The number $+0.83$ means that there is a strong, positive correlation between these two variables. That is, as HSGPA increases, FROSH GPA increases. For a more detailed discussion of correlation, see Larson and Farber (2015).

You could also use the results of the above table in the formula for computing the correlation r in the following way:

$$\text{correlation } r = \left[\left(\frac{1}{(n-1)} \right) \times \Sigma(X - \bar{X}) \times (Y - \bar{Y}) \right] / (\text{STDEV}_x \times \text{STDEV}_y)$$

$$\text{correlation } r = \left[\left(\frac{1}{7} \right) \times 1.09 \right] / [(.48) \times (.39)]$$

$$\text{correlation} = r = 0.83$$

When you use Excel for these computations, you obtain a slightly different correlation of $+0.82$ because Excel uses 16 decimal places for all numbers and computations and is, therefore, more accurate than your calculator.

Now, let's discuss how you can use Excel to find the correlation between two variables in a much simpler, and much faster, fashion than using your calculator.

6.2 Using Excel to Compute a Correlation Between Two Variables

Objective: To use Excel to find the correlation between two variables

Suppose that you have been asked to study the relationship between scores on the Law School Admission Test (LSAT) and the GPA of students at the end of their first-year of Law School. The LSAT is a standardized objective measure of Law School applicants and is a required exam for all Law Schools in the U.S. that are approved by the American Bar Association. About 150,000 applicants take this exam every year in the U.S. Because colleges differ in their standards for grades in courses, the LSAT provides a “level playing field” for all applicants by measuring their readiness for Law School in a single examination taken by all the applicants. There are three subtests of the LSAT (Reading Comprehension, Analytical Reasoning, and Logical Reasoning) that produce a single score that ranges between 120 and 180, with an average score about 150.

To test your Excel skills, you take a random sample of students at the end of their first-year of Law School and record their GPA. The hypothetical data appear in Fig. 6.9:

	A	B	C	D	E	F
1						
2	LAW SCHOOL ADMISSION TEST (LSAT)					
3						
4	Is there a relationship between LSAT scores and first-year GPA in law school?					
5						
6		LSAT score	First-year Law School GPA			
7		130	2.65			
8		170	3.72			
9		140	2.85			
10		160	3.25			
11		150	2.75			
12		180	3.95			
13		130	2.35			
14		160	2.74			
15		170	3.65			
16		140	2.55			
17		160	3.72			
18		140	2.35			

Fig. 6.9 Worksheet Data for LSAT Scores and GPA (Practical Example)

You want to determine if there is a *relationship* between the LSAT scores and GPA at the end of the first-year of Law School, and you decide to use a correlation to determine this relationship. Let’s call the LSAT scores the predictor, X, and first-year GPA the criterion, Y.

Create an Excel spreadsheet with the following information:

- A2: LAW SCHOOL ADMISSION TEST (LSAT)
- A4: Is there a relationship between LSAT scores and first-year GPA in law school?
- B6: LSAT score
- C6: First-year Law School GPA
- B7: 130

Next, change the width of Columns B and C so that the information fits inside the cells.

Now, complete the remaining figures in the table given above so that B18 is 140 and C18 is 2.35. (Be sure to double-check your figures to make sure that they are correct!) Then, center the information in all of these cells.

- A20: n
- A21: mean
- A22: stdev

Next, define the “name” to the range of data from B7:B18 as: LSAT
 We discussed earlier in this book (see Sect. 1.4.4) how to “name a range of data,” but here is a reminder of how to do that:

To give a “name” to a range of data:
 Click on the top number in the range of data and drag the mouse down to the bottom number of the range.

For example, to give the name: “LSAT” to the cells: B7:B18, click on B7, and drag the pointer down to B18 so that the cells B7:B18 are highlighted on your computer screen. Then, click on:

Formulas

Define name (top center of your screen)

LSAT (in the Name box; see Fig. 6.10)

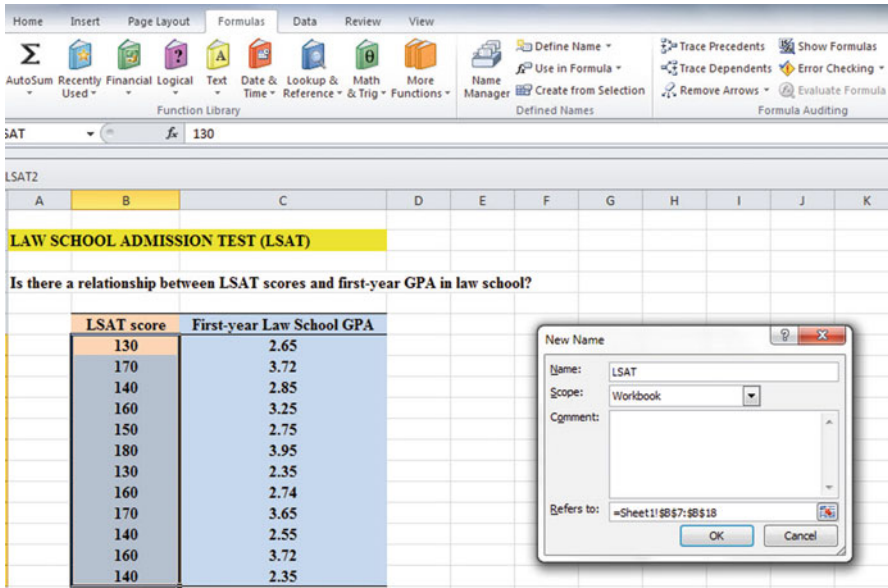


Fig. 6.10 Dialogue Box for Naming a Range of Data as: “LSAT”

OK

Now, repeat these steps to give the name: *GPA* to C7:C18

Finally, click on any blank cell on your spreadsheet to “deselect” cells C7:C18 on your computer screen.

Now, complete the data for these sample sizes, means, and standard deviations in columns B and C so that B22 is 16.58, and C22 is 0.58 (use two decimals for the means and standard deviations; see Fig. 6.11)

LAW SCHOOL ADMISSION TEST (LSAT)		
Is there a relationship between LSAT scores and first-year GPA in law school?		
	LSAT score	First-year Law School GPA
	130	2.65
	170	3.72
	140	2.85
	160	3.25
	150	2.75
	180	3.95
	130	2.35
	160	2.74
	170	3.65
	140	2.55
	160	3.72
	140	2.35
n	12	12
mean	152.50	3.04
stdev	16.58	0.58

Fig. 6.11 Example of Using Excel to Find the Sample Size, Mean, and STDEV

Objective: Find the correlation between LSAT scores and first-year GPA

B24: correlation

C24: =correl(LSAT,GPA) ; see Fig. 6.12

The screenshot shows an Excel spreadsheet with the following data and formulas:

UM					
=correl(LSAT,GPA)					
LSAT3					
A	B	C	D	E	
Is there a relationship between LSAT scores and first-year GPA in law school?					
	LSAT score	First-year Law School GPA			
	130	2.65			
	170	3.72			
	140	2.85			
	160	3.25			
	150	2.75			
	180	3.95			
	130	2.35			
	160	2.74			
	170	3.65			
	140	2.55			
	160	3.72			
	140	2.35			
n	12	12			
mean	152.50	3.04			
stdev	16.58	0.58			
	correlation	=correl(LSAT,GPA)			

Fig. 6.12 Example of Using Excel’s =correl Function to Compute the Correlation Coefficient

Hit the Enter key to compute the correlation

C24: format this cell to two decimals

Note that the equal sign in =correl(LSAT,GPA) in C24 tells Excel that you are going to use a formula in this cell.

The correlation between LSAT scores (X) and first-year GPA (Y) is +.89, a very strong positive correlation. This means that you have evidence that there is a strong relationship between these two variables. In effect, the higher the LSAT score, the higher the first-year GPA in this Law School.

Save this file as: LSAT4

The final spreadsheet appears in Fig. 6.13.

Now, let's create a chart summarizing these data.

Important note: Whenever you draw a chart, it is ESSENTIAL that you put the predictor variable (X) on the left, and the criterion variable (Y) on the right in your Excel spreadsheet, so that you know which variable is the predictor variable and which variable is the criterion variable. If you do this, you will save yourself a lot of grief whenever you do a problem involving correlation and simple linear regression using Excel!

Important note: You need to understand that in any chart that has one predictor and a criterion that there are really TWO LINES that can be drawn between the data points:

- (1) One line uses X as the predictor, and Y as the criterion
- (2) A second line uses Y as the predictor, and X as the criterion

This means that you have to be very careful to note in your input data the cells that contain X as the predictor, and Y as the criterion. If you get these cells mixed up and reverse them, you will create the wrong line for your data and you will have botched the problem terribly.

This is why we STRONGLY RECOMMEND IN THIS BOOK that you always put the X data (i.e., the predictor variable) on the LEFT of your table, and the Y data (i.e., the criterion variable) on the RIGHT of your table on your spreadsheet so that you don't get these variables mixed up.

Also note that the correlation, r , will be exactly the same correlation no matter which variable you call the predictor variable and which variable you call the criterion variable. The correlation coefficient just summarizes the relationship between two variables, and doesn't care which one is the predictor and which one is the criterion.

Let's suppose that you would like to use LSAT scores as the predictor variable, and that you would like to use it to predict first-year GPA for applicants to this Law School. Since the correlation between these two variables is $+0.89$, this shows that there is a strong, positive relationship and that LSAT scores are a good predictor of first-year GPA.

1. Open the file that you saved earlier in this chapter: LSAT4

6.3.1 Using Excel to Create a Chart and the Regression Line Through the Data Points

Objective: To create a chart and the regression line summarizing the relationship between LSAT scores and first-year GPA in Law School

2. Click and drag the mouse to highlight both columns of numbers (B7:C18), *but do not highlight the labels at the top of Column B and Column C.*

Highlight the data set: B7:C18

Insert (top left of screen)

Highlight: Scatter chart icon (immediately above the word: “Charts” at the top center of your screen)

Click on the down arrow on the right of the chart icon

Highlight the top left scatter chart icon (see Fig. 6.14)

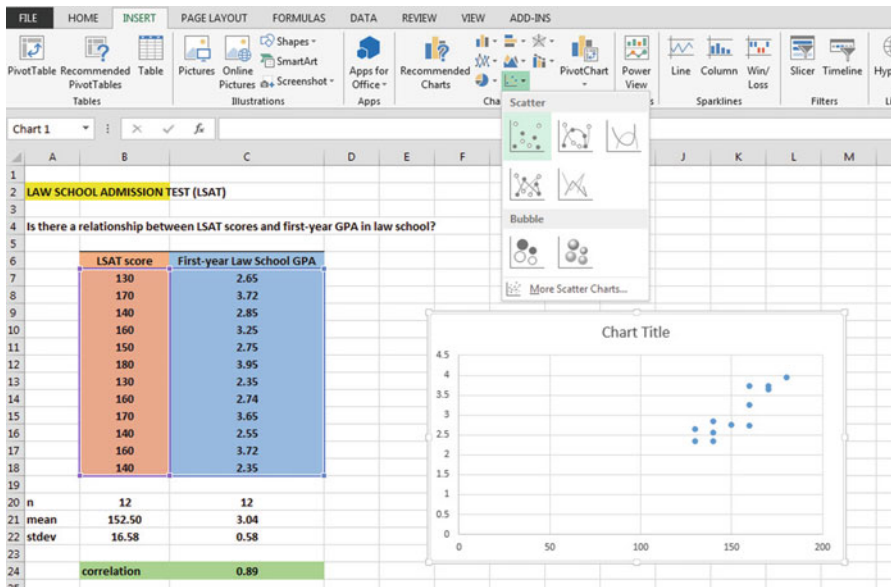


Fig. 6.14 Example of Inserting a Scatter Chart into a Worksheet

Click on the top left chart to select it

Click on the “+ icon” to the right of the chart (CHART ELEMENTS)

Click on the check mark next to “Chart Title” **and also** next to “Gridlines” to remove these check marks (see Fig. 6.15)

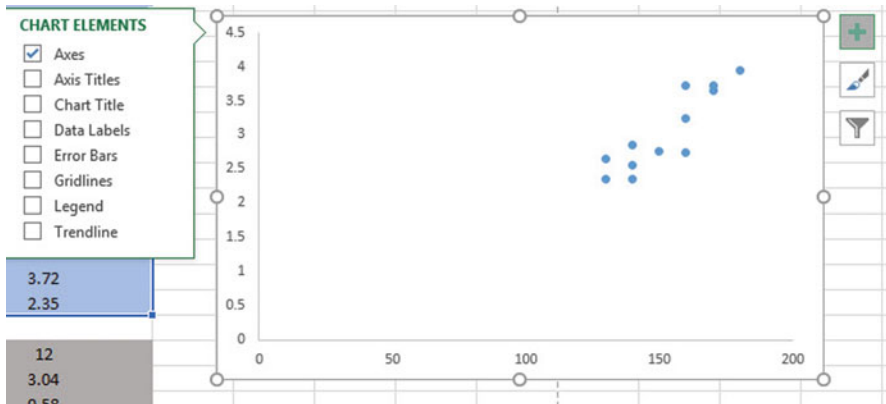


Fig. 6.15 Example of Chart Elements Selected

Click on the box next to: “Chart Title” and then click on the arrow to its right. Then, click on: “Above chart.”

Note that the words: “Chart Title” are now in a box at the top of the chart (see Fig. 6.16)

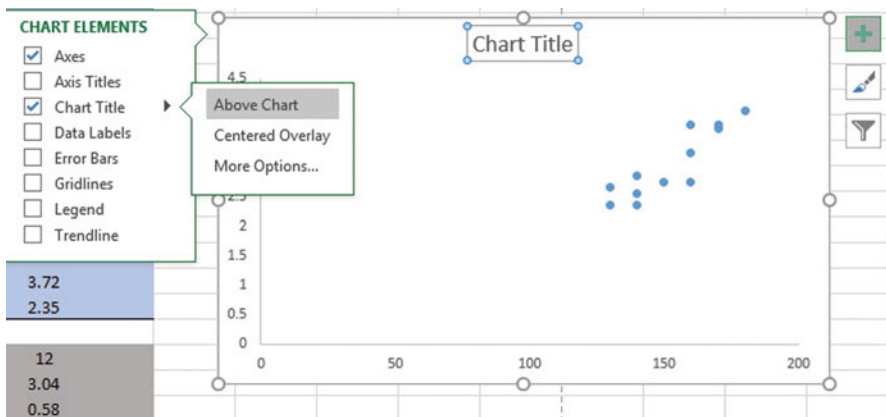


Fig. 6.16 Example of Chart Title Selected

Enter the following Chart Title to the right of f_x at the top of your screen: **RELATIONSHIP BETWEEN LSAT SCORES AND FIRST YEAR GPA IN LAW SCHOOL** (see Fig. 6.17)

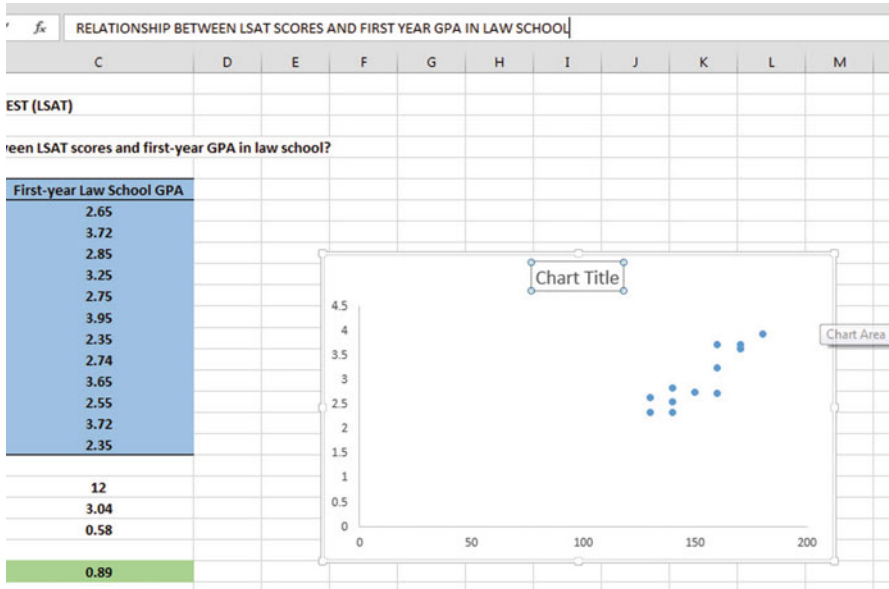


Fig. 6.17 Example of Creating a Chart Title

Hit the Enter Key to enter this chart title onto the chart. Click *inside the chart at the top right corner of the chart* to “de-select” the box around the Chart Title (see Fig. 6.18):

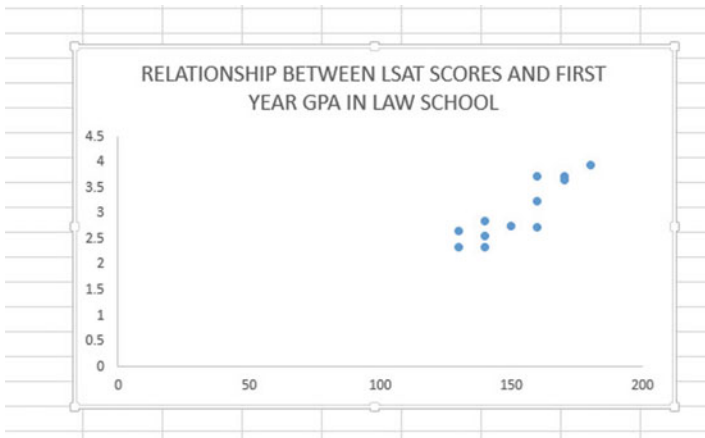


Fig. 6.18 Example of a Chart Title Inserted onto the Chart

Click on the “+ box” to the right of the chart

Add a check mark to the left of “Axis Titles” (This will create an “Axis Title” box on the y-axis of the chart.)

Click on the right arrow for: “Axis titles” and then click on: “Primary Horizontal” to remove the check mark in its box (this will create the y-axis title)

Enter the following y-axis title to the right of f_x at the top of your screen: First-year GPA in Law School

Then, hit the Enter Key to enter this y-axis title to the chart

Click *inside the chart at the top right corner of the chart* to “deselect” the box around the y-axis title (see Fig. 6.19).

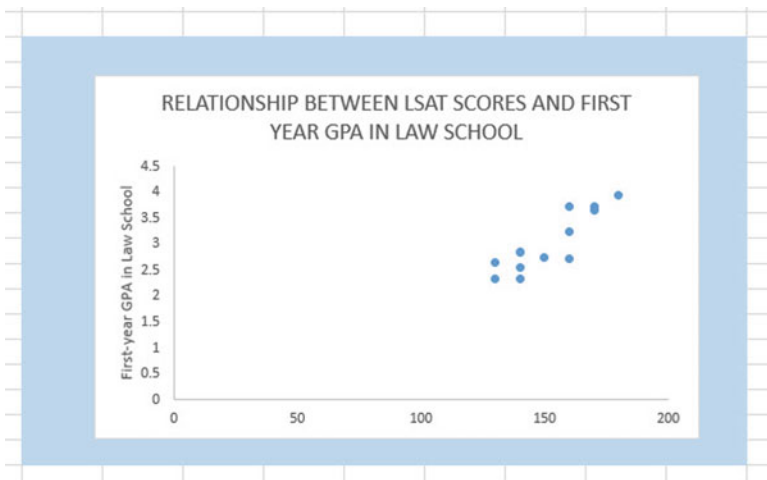


Fig. 6.19 Example of Adding a y-axis Title to the Chart

Click on the “+ box” to the right of the chart

Highlight: “Axis Titles” and click on its right arrow

Click on the words: “Primary Horizontal” to add a check mark to its box (this creates an “Axis Title” box on the x-axis of the chart)

Enter the following x-axis title to the right of f_x at the top of your screen:

LSAT Scores

Then, hit the Enter Key to add this x-axis title to the chart.

Then, click *inside the chart at the top right corner of the chart* to “deselect” the box around the x-axis title (see Fig. 6.20)

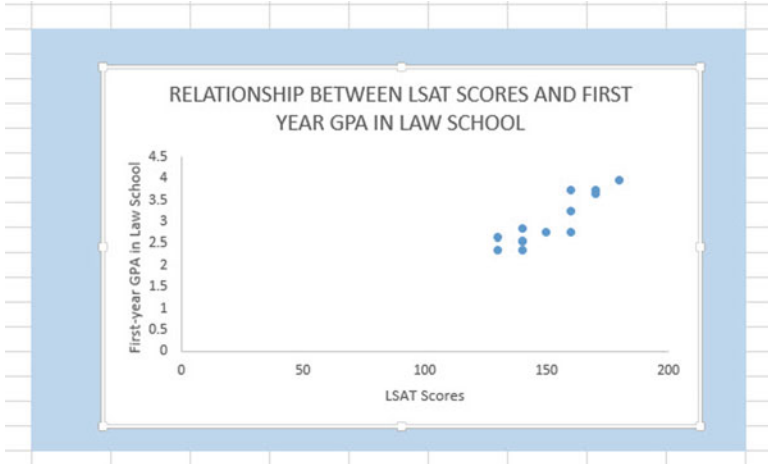


Fig. 6.20 Example of a Chart Title, an x-axis Title, and a y-axis Title

6.3.1.1 Drawing the Regression Line Through the Data Points in the Chart

Objective: To draw the regression line through the data points on the chart

Now, let's draw the regression line onto the chart. This regression line is called the "least-squares regression line" and it is the "best-fitting" straight line through the data points.

Right-click on any one of the data points inside the chart

Highlight: Add Trendline (see Fig. 6.21)

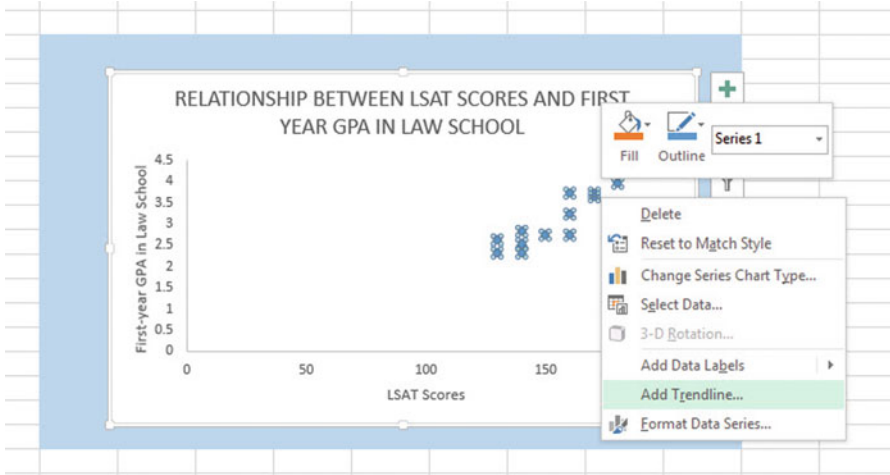


Fig. 6.21 Dialog Box for Adding a Trendline to the Chart

Click on: Add Trendline

Linear (be sure the “linear” button near the top is selected on the “Format Trendline” dialog box; see Fig. 6.22)

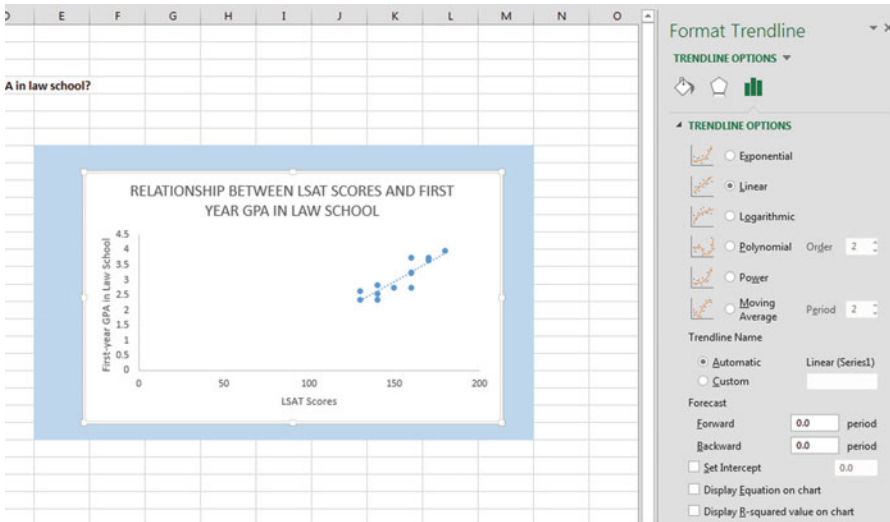


Fig. 6.22 Dialog Box for a Linear Trendline

Click on the **X** at the top right of the “Format Trendline” dialog box to close this dialog box

Click on any blank cell *outside the chart* to “deselect” the chart

Save this file as: LSAT5

Your spreadsheet should look like the spreadsheet in Fig. 6.23:

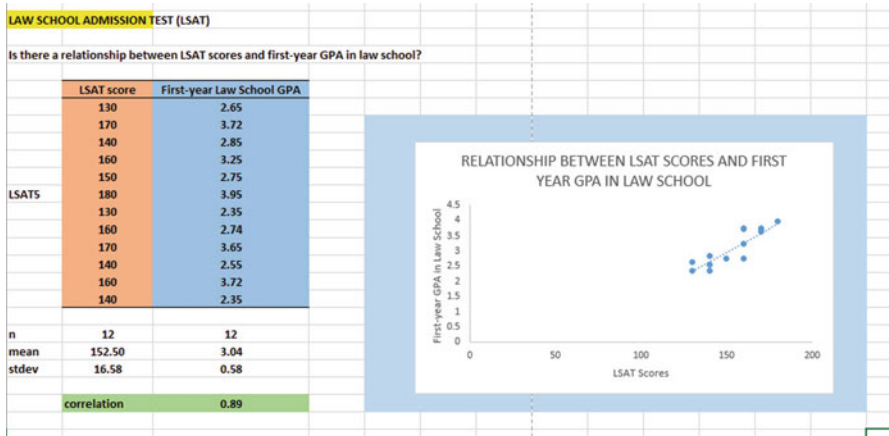


Fig. 6.23 Final Chart with the Trendline Fitted Through the Data Points of the Scatterplot

6.3.1.2 Moving the Chart Below the Table in the Spreadsheet

Objective: To move the chart below the table

Left-click your mouse on *any white space to the right of the top title inside the chart*, keep the left-click down, and drag the chart down and to the left so that the top left corner of the chart is in cell A26, then take your finger off the left-click of the mouse (see Fig. 6.24).

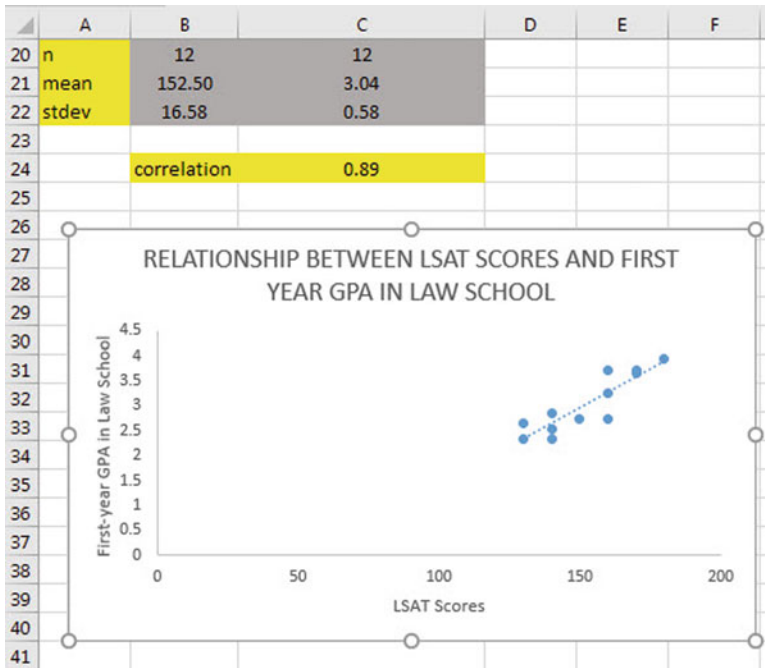


Fig. 6.24 Example of Moving the Chart Below the Table

6.3.1.3 Making the Chart “Longer” So That It Is “Taller”

Objective: To make the chart “longer” so that it is taller

Left-click your mouse on the bottom-center of the chart to create an “up-and-down-arrow” sign, hold the left-click of the mouse down and drag the bottom of the chart down to row 48 to make the chart longer, and then take your finger off the mouse.

6.3.1.4 Making the Chart “Wider”

Objective: To make the chart “wider”

Put the pointer at the middle of the right-border of the chart to create a “left-to-right arrow” sign, and then left-click your mouse and hold the left-click down while you drag the right border of the chart to the middle of Column H to make the chart wider (see Fig. 6.25).

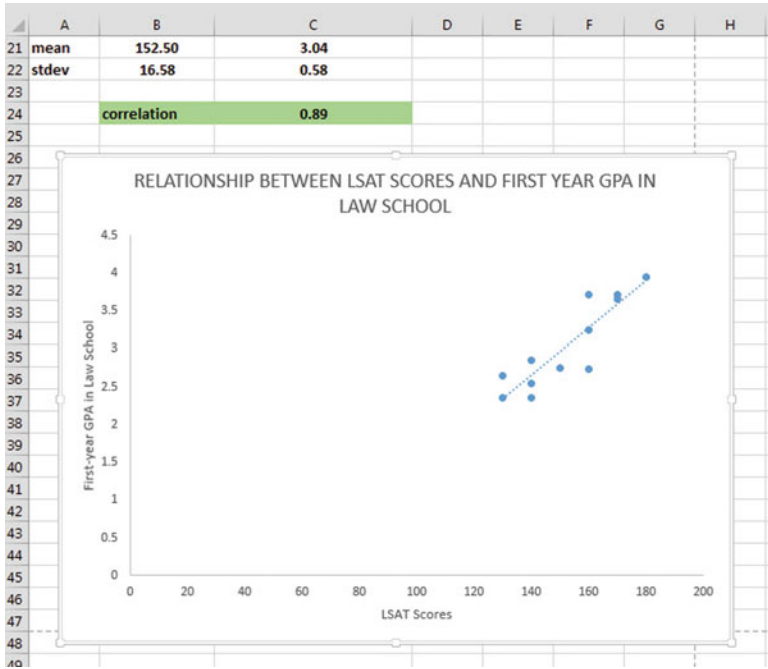


Fig. 6.25 Example of a Chart that is Enlarged to Fit the Cells: A26:H48

Save this file as: LSAT6

Note: If you printed this spreadsheet now, it is “too big” to fit onto one page, and would “dribble over” onto four pages of printout because the scale needs to be reduced below 100% in order for this worksheet to fit onto only one page. You need to complete these next steps below to print out some, or all, of this spreadsheet.

6.4 Printing a Spreadsheet So That the Table and Chart Fit onto One Page

Objective: To print the spreadsheet so that the table and the chart fit onto one page

Page Layout (top of screen)

Change the scale at the middle icon near the top of the screen “Scale to Fit” by clicking on the down-arrow until it reads “90%” so that the table and the chart will fit onto one page on your printout (see Fig. 6.26):

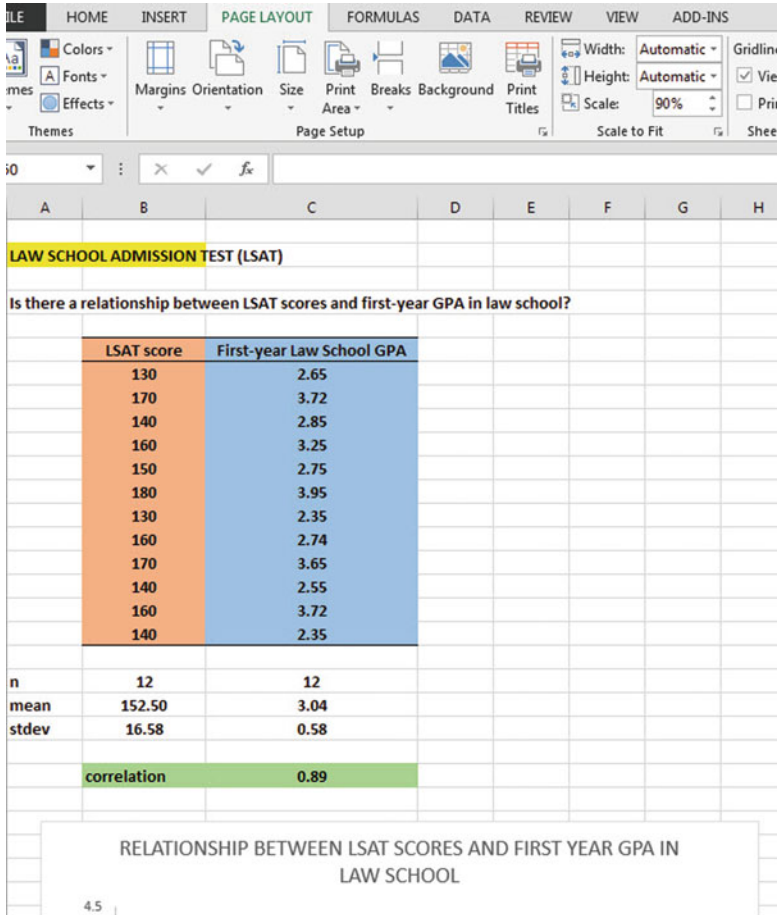


Fig. 6.26 Example of the Page Layout for Reducing the Scale of the Chart to 90% of Normal Size

File

Print

Print (see Fig. 6.27)

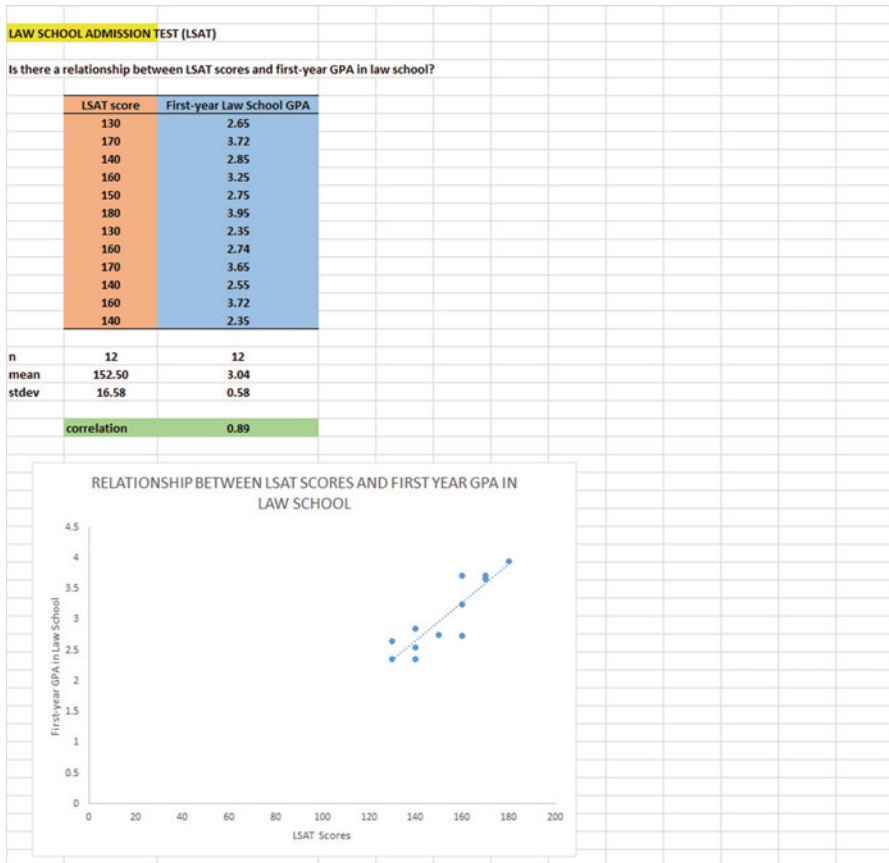


Fig. 6.27 Final Spreadsheet of a Table and a Chart (90% Scale to Fit Size)

Save your file as: LSAT7

6.5 Finding the Regression Equation

The main reason for charting the relationship between X and Y (i.e., LSAT scores as X and First-year GPA in Law School as Y in our example) is to see if there is a strong enough relationship between X and Y so that the regression equation that summarizes this relationship can be used to predict Y for a given value of X.

Since we know that the correlation between LSAT scores and GPA is $+0.89$, this tells us that it makes sense to use LSAT scores to predict first-year GPA in Law School based on past data from this Law School.

We now need to find that regression equation that is the equation of the “best-fitting straight line” through the data points.

Objective: To find the regression equation summarizing the relationship between X and Y.

In order to find this equation, we need to check to see if your version of Excel contains the “Data Analysis ToolPak” necessary to run a regression analysis.

6.5.1 Installing the Data Analysis ToolPak into Excel

Objective: To install the Data Analysis ToolPak into Excel

Since there are currently four versions of Excel in the marketplace (2007, 2010, 2013, and 2016), we will give a brief explanation of how to install the Data Analysis ToolPak into each of these versions of Excel.

6.5.1.1 Installing the Data Analysis ToolPak into Excel 2016

Open a new Excel spreadsheet

Click on: Data (at the top of your screen)

Look at the top of your monitor screen. Do you see the words: “Data Analysis” at the far right of the screen? If you do, the Data Analysis ToolPak for Excel 2016 was correctly installed when you installed Office 2016, and you should skip ahead to Sect. 6.5.2.

If the words: “Data Analysis” are not at the top right of your monitor screen, then the ToolPak component of Excel 2016 was not installed when you installed Office 2016 onto your computer. If this happens, you need to follow these steps:

File

Options (bottom left of screen)

Note: This creates a dialog box with “Excel Options” (at the top left of the box)

Add-Ins (on left of screen)

Manage: Excel Add-Ins (at the bottom of the dialog box)

Go (at bottom center of dialog box)

Highlight: Analysis ToolPak (in the Add-Ins dialog box)

Put a check mark to the left of Analysis Toolpak

OK (at the right of this dialog box)

Data

You now should have the words: “Data Analysis” at the top right of your screen to show that this feature has been installed correctly

Note: If these steps do not work, you should try these steps instead:

*File/Options (bottom left)/Add-ins/Analysis ToolPak/Go/
click to the left of Analysis ToolPak to add a check mark/OK*

If you need help doing this, ask your favorite “computer techie” for help. You are now ready to skip ahead to Sect. [6.5.2](#)

6.5.1.2 Installing the Data Analysis ToolPak into Excel 2013

Open a new Excel spreadsheet

Click on: Data (at the top of your screen)

Look at the top of your monitor screen. Do you see the words: “Data Analysis” at the far right of the screen? If you do, the Data Analysis ToolPak for Excel 2013 was correctly installed when you installed Office 2013, and you should skip ahead to Sect. [6.5.2](#).

If the words: “Data Analysis” are not at the top right of your monitor screen, then the ToolPak component of Excel 2013 was not installed when you installed Office 2013 onto your computer. If this happens, you need to follow these steps:

File

Options (bottom left of screen)

Note: This creates a dialog box with “Excel Options” (at the top left of the box)

Add-Ins (on left of screen)

Manage: Excel Add-Ins (at the bottom of the dialog box)

Go (at bottom center of dialog box)

Highlight: Analysis ToolPak (in the Add-Ins dialog box)

Put a check mark to the left of Analysis Toolpak

OK (at the right of this dialog box)

Data

You now should have the words: “Data Analysis” at the top right of your screen to show that this feature has been installed correctly

If you get a prompt asking you for the “installation CD,” put this CD in the CD drive and click on: OK

Note: If these steps do not work, you should try these steps instead:

*File/Options (bottom left)/Add-ins/Analysis ToolPak/Go/
click to the left of Analysis ToolPak to add a check mark/OK*

If you need help doing this, ask your favorite “computer techie” for help. You are now ready to skip ahead to Sect. [6.5.2](#)

6.5.1.3 Installing the Data Analysis ToolPak into Excel 2010

Open a new Excel spreadsheet

Click on: Data (at the top of your screen)

Look at the top of your monitor screen. Do you see the words: “Data Analysis” at the far right of the screen? If you do, the Data Analysis ToolPak for Excel 2010 was correctly installed when you installed Office 2010, and you should skip ahead to Sect. 6.5.2.

If the words: “Data Analysis” are not at the top right of your monitor screen, then the ToolPak component of Excel 2010 was not installed when you installed Office 2010 onto your computer. If this happens, you need to follow these steps:

File

Options

Excel options (creates a dialog box)

Add-Ins

Manage: Excel Add-Ins (at the bottom of the dialog box)

Go

Highlight: Analysis ToolPak (in the Add-Ins dialog box)

OK

Data (You now should have the words: “Data Analysis” at the top right of your screen)

If you get a prompt asking you for the “installation CD,” put this CD in the CD drive and click on: OK

Note: If these steps do not work, you should try these steps instead:

File/Options (bottom left)/Add-ins/Analysis ToolPak/Go/

click to the left of Analysis ToolPak to add a check mark/OK

If you need help doing this, ask your favorite “computer techie” for help.

You are now ready to skip ahead to Sect. 6.5.2.

6.5.1.4 Installing the Data Analysis ToolPak into Excel 2007

Open a new Excel spreadsheet

Click on: Data (at the top of your screen)

If the words “Data Analysis” do not appear at the top right of your screen, you need to install the Data Analysis ToolPak using the following steps:

Microsoft Office button (top left of your screen)

Excel options (bottom of dialog box)

Add-ins (far left of dialog box)

Go (to create a dialog box for Add-Ins)

Highlight: Analysis ToolPak

OK (If Excel asks you for permission to proceed, click on: Yes)

Data (You should now have the words: “Data Analysis” at the top right of your screen)

If you need help doing this, ask your favorite “computer techie” for help.

You are now ready to skip ahead to Sect. 6.5.2.

6.5.2 Using Excel to Find the SUMMARY OUTPUT of Regression

You have now installed *ToolPak*, and you are ready to find the regression equation for the “best-fitting straight line” through the data points by using the following steps:

Open the Excel file: *LSAT7* (if it is not already open on your screen)

Note: If this file is already open, and there is a border around the chart, you need to click on any empty cell outside of the chart to deselect the chart.

Now that you have installed *Toolpak*, you are ready to find the regression equation summarizing the relationship between LSAT scores and first-year GPA in Law School in your data set.

Remember that you gave the name: *LSAT* to the X data (the predictor), and the name: *GPA* to the Y data (the criterion) in a previous section of this chapter (see Sect. 6.2)

Data (top of screen)

Data analysis (far right at top of screen; see Fig. 6.28)

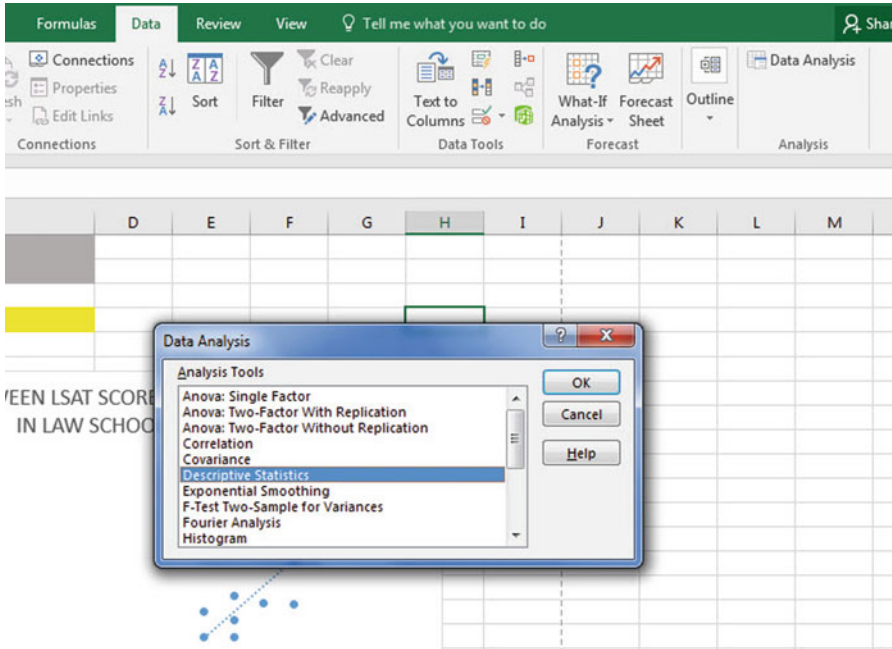
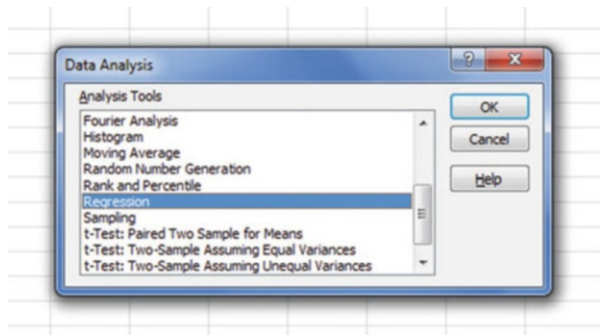


Fig. 6.28 Example of Using the Data/Data Analysis Function of Excel

Scroll down the dialog box using the down arrow and click on: Regression (see Fig. 6.29)

Fig. 6.29 Dialogue Box for Creating the Regression Function in Excel



OK

Input Y Range: GPA
Input X Range: LSAT

Click on the “button” to the left of Output Range to select this, and enter A50 in the box as the place on your spreadsheet to insert the Regression analysis in cell A50
OK

The *SUMMARY OUTPUT* should now be in cells: A50 : I67

Now, make the columns in the Regression Summary Output section of your spreadsheet *wider* so that you can read all of the column headings clearly.

Now, change the data in the following three cells to Number format (two decimal places) by first clicking on “Home” at the top left of your screen:

B53

B66

B67

Now, change the format for all other numbers that are in decimal format to number format, three decimal places, and center all numbers within their cells.

Print the file so that it fits onto one page. (*Hint: Change the scale under “Page Layout” to 65% to make it fit.*) Your file should be like the file in Fig. 6.30.

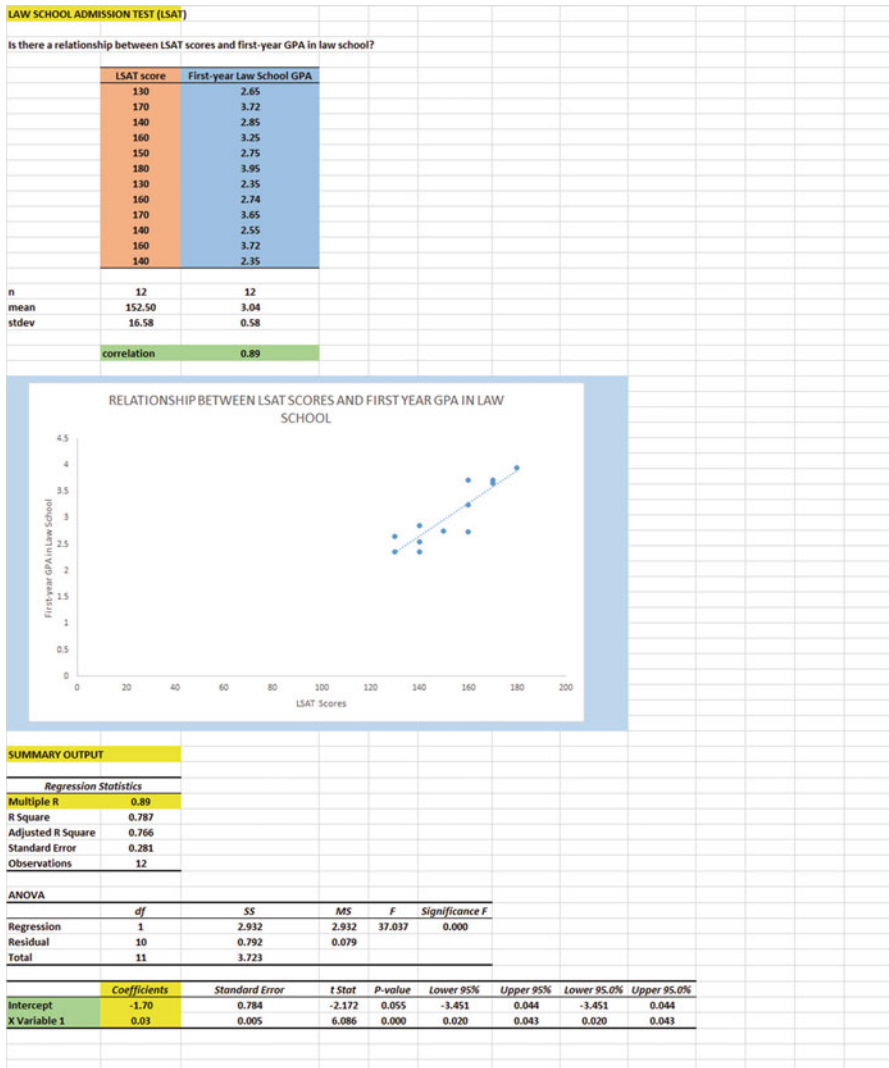


Fig. 6.30 Final Spreadsheet of Correlation and Simple Linear Regression including the SUMMARY OUTPUT for the Data

Save the resulting file as: LSAT8

Note the following problem with the summary output.

Whoever wrote the computer program for this version of Excel made a mistake and gave the name: “Multiple R” to cell A53.

This is not correct. Instead, cell A53 should say: “correlation r” since this is the notation that we are using for the correlation between X and Y.

You can now use your printout of the regression analysis to find the regression equation that is the best-fitting straight line through the data points.

But first, let's review some basic terms.

6.5.2.1 Finding the y-Intercept, a , of the Regression Line

The point on the y-axis that the regression line would intersect the y-axis if it were extended to reach the y-axis is called the “y-intercept” and *we will use the letter “a” to stand for the y-intercept of the regression line.* The y-intercept on the SUMMARY OUTPUT on the previous page is -1.70 and appears in cell B66 (note the minus sign). This means that if you were to draw an imaginary line continuing down the regression line toward the y-axis that this imaginary line would cross the y-axis at -1.70 . This is why a is called the “y-intercept.”

6.5.2.2 Finding the Slope, b , of the Regression Line

The “tilt” of the regression line is called the “slope” of the regression line. It summarizes to what degree the regression line is either above or below a horizontal line through the data points. If the correlation between X and Y were zero, the regression line would be exactly horizontal to the X-axis and would have a zero slope.

If the correlation between X and Y is positive, the regression line would “slope upward to the right” above the X-axis. Since the regression line in Fig. 6.30 slopes upward to the right, the slope of the regression line is $+0.03$ as given in cell B67. *We will use the notation “b” to stand for the slope of the regression line.* (Note that Excel calls the slope of the line: “X Variable 1” in the Excel printout.)

Since the correlation between the LSAT scores and first-year GPA was $+0.89$, you can see that the regression line for these data “slopes upward to the right” through the data. Note that the SUMMARY OUTPUT of the regression line in Fig. 6.30 gives a correlation, r , of $+0.89$ in cell B53.

If the correlation between X and Y were negative, the regression line would “slope down to the right” above the X-axis. This would happen whenever the correlation between X and Y is a negative correlation that is between zero and minus one (0 and -1).

6.5.3 Finding the Equation for the Regression Line

To find the regression equation for the straight line that can be used to predict first-year GPA in Law School from an LSAT score, we only need two numbers in the SUMMARY OUTPUT in Fig. 6.30: *B66* and *B67*.

$$\text{The format for the regression line is: } Y = a + b X \quad (6.3)$$

where a = the *y*-intercept (-1.70 in our example in cell *B66*)

and b = the *slope of the line* ($+0.03$ in our example in cell *B67*)

Therefore, the equation for the best-fitting regression line for our example is:

$$Y = a + b X$$

$$Y = -1.70 + 0.03 X$$

Remember that Y is the first-year GPA that we are trying to predict, using the LSAT scores as the predictor, X .

Let's try an example using this formula to predict first-year GPA for a hypothetical student.

6.5.4 Using the Regression Line to Predict the *y*-Value for a Given *x*-Value

Objective: To find the first-year GPA predicted from an LSAT score of 150

Since the LSAT score is 150 (i.e., $X = 150$), substituting this number into our regression equation gives:

$$Y = -1.70 + 0.03 (150)$$

$$Y = -1.70 + 4.5$$

$$Y = 2.80$$

Important note: If you look at your chart, if you go directly upwards for an LSAT score of 150 until you hit the regression line, you see that you hit this line just under the number 3 on the *y*-axis to the left when you draw a line horizontal to the *x*-axis (actually, it is 2.80), the result above for predicting first-year GPA from an LSAT score of 150.

Now, let's do a second example and predict what the first-year GPA would be if we used an LSAT score of 170.

$$Y = -1.70 + 0.03 X$$

$$Y = -1.70 + 0.03 (170)$$

$$Y = -1.70 + 5.1$$

$$Y = 3.40$$

Important note: If you look at your chart, if you go directly upwards from an LSAT score of 170 until you hit the regression line, you see that you hit this line just under the number 3.5 on the y-axis to the left (actually it is 3.40), the result above for predicting first-year GPA from this LSAT score of 170.

For a more detailed discussion of regression, see Starnes et al. (2015).

6.6 Adding the Regression Equation to the Chart

Objective: To Add the Regression Equation to the Chart

If you want to include the regression equation within the chart next to the regression line, you can do that, but a word of caution first.

Throughout this book, we are using the regression equation for one predictor and one criterion to be the following:

$$Y = a + b X \tag{6.3}$$

where a = y-intercept and
 b = slope of the line

See, for example, the regression equation in Sect. 6.5.3 where the y-intercept was $a = -1.70$ and the slope of the line was $b = +0.03$ to generate the following regression equation:

$$Y = -1.70 + 0.03 X$$

However, Excel 2016 uses a slightly different regression equation (which is logically identical to the one used in this book) when you add a regression equation to a chart:

$$Y = bX + a \tag{6.4}$$

where a = y-intercept and b = slope of the line

Note that this equation is identical to the one we are using in this book with the terms arranged in a different sequence.

For the example we used in Sect. 6.5.3, Excel 2016 would write the regression equation on the chart as:

$$Y = 0.03 X - 1.70$$

This is the format that will result when you add the regression equation to the chart using Excel 2016 using the following steps:

Open the file: LSAT8 (that you saved in Sect. 6.5.2)

Click just *inside* the outer border of the chart in the top right corner to add the “border” around the chart in order to “select the chart” for changes you are about to make

Right-click on any of the data-points in the chart

Highlight: Add Trendline, and click on this to select this command

The “Linear button” near the top of the dialog box will be selected (on its left)

Scroll down this dialog box and click on: Display Equation on chart (near the bottom of the dialog box; see Fig. 6.31)

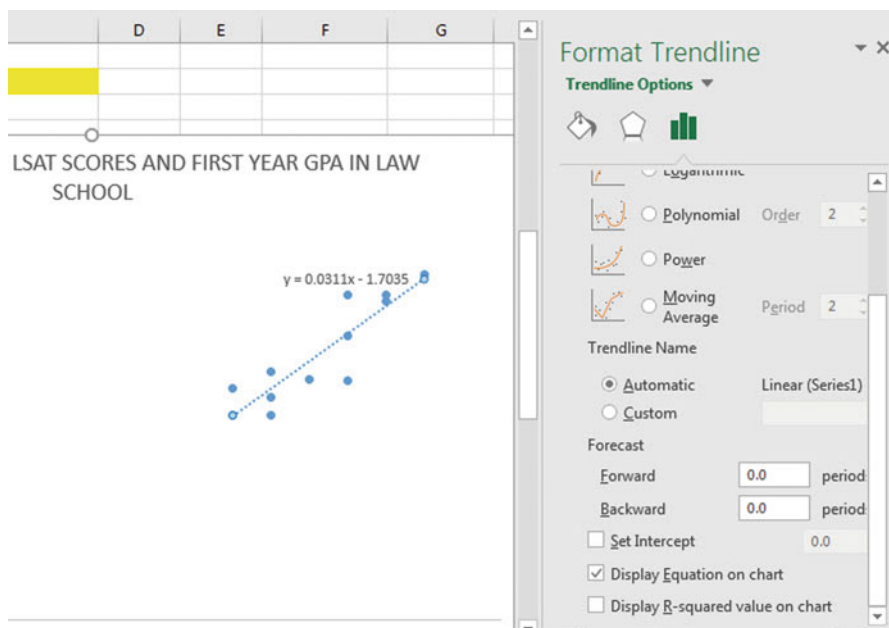


Fig. 6.31 Dialog Box for Adding the Regression Equation to the Chart Next to the Regression Line on the Chart

Click on **X** at the top right of the Format Trendline dialog box to remove this box
Click on any empty cell outside of the chart to deselect the chart.

Note that the regression equation on the chart is in the following form next to the regression line on the chart (see Fig. 6.32).

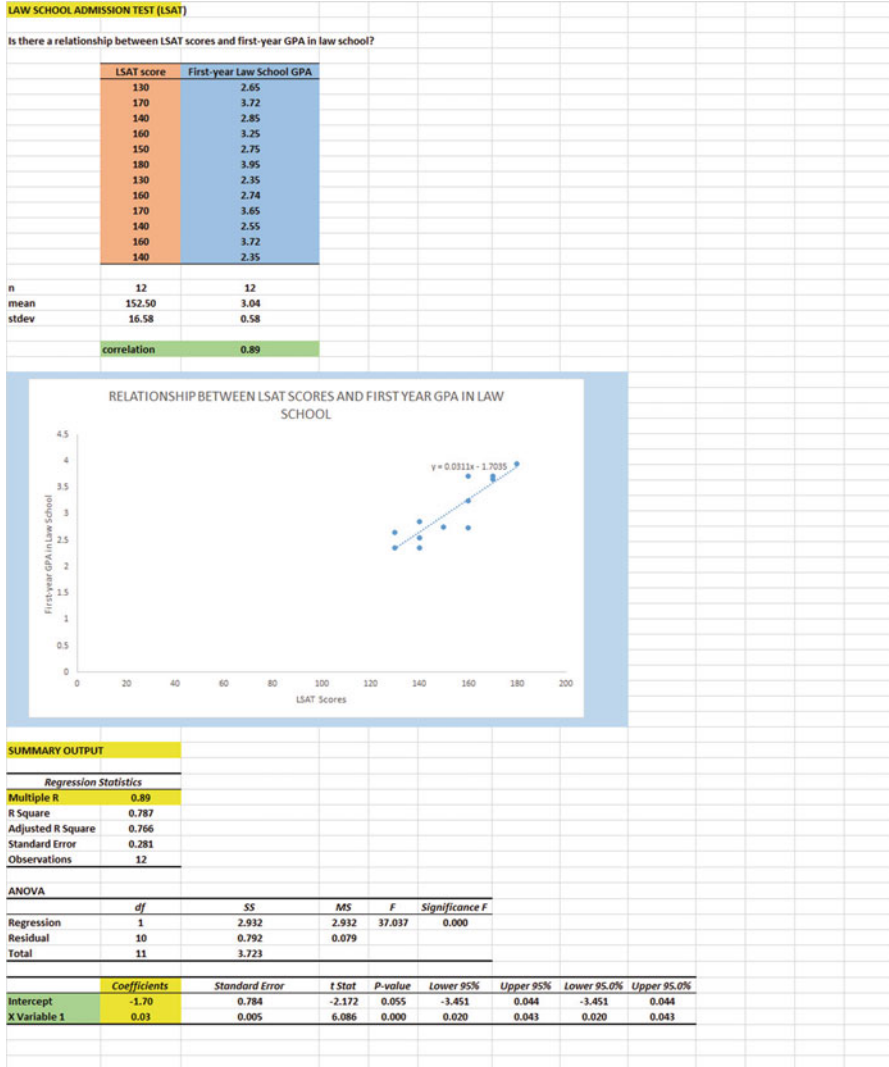


Fig. 6.32 Example of a Chart with the Regression Equation Displayed Next to the Regression Line

$$Y = 0.03 X - 1.70$$

Now, save this file as: LSAT9, and print it out so that it fits onto one page

6.7 How to Recognize Negative Correlations in the SUMMARY OUTPUT Table

Important note: Since Excel does not recognize negative correlations in the SUMMARY OUTPUT results, but treats all correlations as if they were positive correlations (this was a mistake made by the programmer), you need to be careful to note that there may be a negative correlation between X and Y even if the printout says that the correlation is a positive correlation.

You will know that the correlation between X and Y is a negative correlation when these two things occur:

- (1) THE SLOPE, b, IS A NEGATIVE NUMBER. This can only occur when there is a negative correlation.*
- (2) THE CHART CLEARLY SHOWS A DOWNWARD SLOPE IN THE REGRESSION LINE, which can only occur when the correlation between X and Y is negative.*

6.8 Printing Only Part of a Spreadsheet Instead of the Entire Spreadsheet

Objective: To print part of a spreadsheet separately instead of printing the entire spreadsheet

There will be many occasions when your spreadsheet is so large in the number of cells used for your data and charts that you only want to print part of the spreadsheet separately so that the print will not be so small that you cannot read it easily.

We will now explain how to print only part of a spreadsheet onto a separate page by using three examples of how to do that using the file, LSAT9, that you created in Sect. 6.6: (1) printing only the table and the chart on a separate page, (2) printing only the chart on a separate page, and (3) printing only the SUMMARY OUTPUT of the regression analysis on a separate page.

Note: If the file: LSAT9 is not open on your screen, you need to open it now.

If the border is around the outside of the chart, click on any white space outside the chart to deselect the chart.

Let's describe how to do these three goals with three separate objectives:

6.8.1 *Printing Only the Table and the Chart on a Separate Page*

Objective: To print only the table and the chart on a separate page

1. Left-click your mouse starting at the top left of the table in cell A2 and drag the mouse *down and to the right so that all of the table and all of the chart are highlighted in light blue on your computer screen from cell A2 to cell H48* (the light blue cells are called the “selection” cells).
2. File
Print
Print Active Sheets (hit the down arrow on the right)
Print selection
Print

The resulting printout should contain only the table of the data and the chart resulting from the data.

Then, click on any empty cell in your spreadsheet to deselect the table and chart.

6.8.2 *Printing Only the Chart on a Separate Page*

Objective: To print only the chart on a separate page

1. Click on any “white space” *just inside the outside border of the chart in the top right corner of the chart* to create the border around all of the borders of the chart in order to “select” the chart.
2. File
Print
Print selected chart
Print selected chart (again)
Print

The resulting printout should contain only the chart resulting from the data.

Important note: After each time you print a chart by itself on a separate page, you should immediately click on any white space OUTSIDE the chart to remove the border from the chart. When the border is on the borders of the chart, this tells Excel that you want to print only the chart by itself. Do this now!

6.8.3 *Printing Only the SUMMARY OUTPUT of the Regression Analysis on a Separate Page*

Objective: To print only the SUMMARY OUTPUT of the regression analysis on a separate page

1. Left-click your mouse at the cell just above SUMMARY OUTPUT in *cell A50* on the left of your spreadsheet and drag the mouse *down and to the right* until all of the regression output is highlighted in dark blue on your screen from A50 to I67. (Change the “Scale to Fit” to 65% so that the SUMMARY OUTPUT will fit onto one page when you print it out.)
2. File
Print
Print active sheets (hit the down arrow on the right)
Print selection
Print

The resulting printout should contain only the summary output of the regression analysis on a separate page.

Finally, click on any empty cell on the spreadsheet to “deselect” the regression table.

6.9 End-of-Chapter Practice Problems

1. Suppose that you worked for Snapchat and that you have been asked to create and analyze the data from an Internet survey of adults ages 18–24 in terms of their satisfaction with the Snapchat app. You are in the early stages of developing the survey, but you are sure that you want to ask one question about satisfaction with the app and another question about the likelihood of users to recommend the Snapchat app to a friend.

You have decided to use a correlation and simple linear regression analysis, and to test your Excel skills, you have created some hypothetical data for this comparison. These hypothetical data appear in Fig. 6.33:

- (d) Use Excel's *regression* function to find the equation for the least-squares regression line for these data and display the results below the chart on your spreadsheet.
- (e) Use number format (two decimal places) for the correlation, the y-intercept, and the slope of the line on the SUMMARY OUTPUT, and use number format (three decimal places) for all of the other decimal figures in the SUMMARY OUTPUT.
- (f) Print the *input data and the chart* so that this information fits onto one page.
- (g) Then, print the *regression output table* so that this information fits onto a separate page.
- (h) Save the file as: Snap4

Answer the following questions using your Excel printout:

By hand:

1. Circle and label the value of the *y-intercept* and the *slope* of the regression line on your printout.
 2. Write the *regression equation* by hand on your printout for these data
 3. Circle and label the *correlation* between the two sets of scores in the regression analysis SUMMARY OUTPUT table on your printout
 4. Underneath the regression equation you wrote by hand on your printout, use the regression equation to predict the likelihood to recommend to a friend you would expect for a satisfaction rating of 5.
 5. Read from the graph, the likelihood to recommend to a friend you would expect for a satisfaction rating of 6, and write your answer on the separate page
2. Suppose that you wanted to study the relationship between DIET (measured in calories allowed per day) and WEIGHT LOSS (measured in kilograms, kg) for adult women between the ages of 30 and 40 who are overweight for their height and body structure, and who all weigh roughly the same number of kilograms before undertaking the weight loss program. You want to test your Excel skills on a random sample of these women based on their weight change over the past four months to make sure that you can do this type of research. The hypothetical data appear in Fig. 6.34:

Fig. 6.34 Worksheet Data for Chap. 6: Practice Problem #2

RELATIONSHIP BETWEEN DIET AND WEIGHT LOSS	
ADULT WOMEN AGES 30-40	
DIET (calories allowed per day)	WEIGHT LOSS (kg)
900	16.0
1050	12.0
1150	8.0
1275	6.0
1420	3.0
1530	5.5
1610	9.5
1710	2.5
1820	6.0
1875	9.0
1930	6.0
2100	3.0

Create an Excel spreadsheet and enter the data using DIET (calories allowed per day) as the independent variable (predictor) and WEIGHT LOSS (kg) as the dependent variable (criterion). Underneath the table, use Excel's `=correl` function to find the correlation between these two variables. Label the correlation and place it underneath the table; then round off the correlation to two decimal places.

- (a) create an *XY scatterplot* of these two sets of data such that:
 - top title: RELATIONSHIP BETWEEN DIET AND WEIGHT LOSS
 - x-axis title: DIET (calories allowed per day)
 - y-axis title: WEIGHT LOSS (kg)
 - move the chart below the table and the correlation
 - re-size the chart so that it is 8 columns wide and 25 rows long
 - delete the legend
 - delete the gridlines
- (b) Create the *least-squares regression line* for these data on the scatterplot, and add the regression equation to the chart.
- (c) Use Excel to run the regression statistics to find the *equation for the least-squares regression line* for these data and display the results below the chart on your spreadsheet. Use number format (two decimal places) for the correlation and three decimal places for all other decimal figures, including the coefficients.

- (d) Print just the input data and the chart so that this information fits onto one page. Then, print the regression output table on a separate page so that it fits onto that separate page.
- (e) save the file as: DIET3

Answer the following questions using your Excel printout:

1. What is the correlation between DIET and WEIGHT LOSS?
2. What is the y-intercept?
3. What is the slope of the line?
4. What is the regression equation?
5. Use the regression equation to predict the WEIGHT LOSS you would expect for a woman who was practicing a DIET of 1500 calories allowed a day. Show your work on a separate sheet of paper.

(Note that this correlation is not the multiple correlation as the Excel table indicates, but is merely the correlation r instead.)

Important note: Since Excel does not recognize negative correlations in the SUMMARY OUTPUT but treats all correlations as if they were positive correlations, you need to be careful to note that there is a negative correlation between DIET and WEIGHT LOSS.

You know this for two reasons:

- (1) *The slope, b , is a negative -0.007 which can only occur when there is a negative correlation.*
- (2) *The chart clearly shows a downward slope in the regression line, which can only happen when the correlation is negative.*

Therefore, the correlation between DIET and WEIGHT LOSS is not $+0.64$, but -0.64 for this problem. This is a negative correlation!

3. The Advanced Placement (AP) Tests are standardized tests that allow high school students to “test out” of college courses, either by receiving college credit for these courses while in high school or by allowing the students to waive courses in which their AP scores are very high. The AP Exams are scored on a 5-point scale in which a score of “1” means “No recommendation” and a score of “5” means “Extremely well qualified.” There are 34 AP courses for which these exams are available. The AP Calculus BC Test is intended to measure the equivalent of a college introductory calculus course taken by students during their first year of college. Each spring, about 90,000 students take this test.

Suppose that you have been asked by the Chair of the Mechanical Engineering Department at a selective university to see how well the AP Calculus BC Test predicts GPA at the end of the first year of study for Mechanical Engineering

majors. This Chair has asked you to use the AP Calculus BC Test as the predictor of this GPA. The Chair would like your recommendation as to whether or not the AP Calculus BC Test should become an admissions requirement in addition to the SAT for admission to the undergraduate major in Mechanical Engineering.

You have decided to use a correlation and simple linear regression analysis, and to test your Excel skills, you have collected the data of a random sample of 19 Mechanical Engineering students who have just finished their first year of study at this university. These hypothetical data appear in Fig. 6.35.

What is the relationship between AP Calculus BC scores and freshman GPA?	
AP score	FROSH GPA
3	3.25
4	3.56
5	3.84
4	3.55
5	3.52
3	3.23
2	3.15
5	3.46
4	3.56
3	3.16
4	3.24
2	2.96
3	3.21
4	3.16
5	3.66
4	3.54
3	3.15
4	3.56
5	3.92

Fig. 6.35 Worksheet Data for Chap. 6: Practice Problem #3

- (a) create an Excel spreadsheet using FROSH GPA as the criterion and the AP Calculus BC Test as the predictor using the following format:
 - Top title: RELATIONSHIP BETWEEN AP CALCULUS BC SCORES AND FROSH GPA
 - x-axis title: AP CALCULUS BC SCORE
 - y-axis title: FROSH GPA
 - Re-size the chart so that it is 7 columns wide and 25 rows long
 - Delete the legend
 - Delete the gridlines
 - Move the chart below the table
- (b) Create the *least-squares regression line* for these data on the scatterplot.

- (c) Use Excel's *regression* function to find the equation for the least-squares regression line for these data and display the results below the chart on your spreadsheet. Add the regression line and the regression equation to the chart.
- (d) Use number format (two decimal places) for the correlation on the SUMMARY OUTPUT, and use number format (three decimal places) for all of the other decimal figures in the SUMMARY OUTPUT.
- (e) Print the input data and the chart so that this information fits onto one page.
- (f) Then, print the regression output table so that this information fits onto a separate page.
- (g) Save the file as: FROSH16

Answer the following questions using your Excel printout:

1. What is correlation r ?
2. What is the y -intercept a ?
3. What is the slope b ?
4. What is the regression equation (use three decimal places for the y -intercept and the slope)?
5. Use the regression equation to predict the FROSH GPA you would expect for an AP Calculus BC score of 4.

References

- Larson, R. and Farber, B. Elementary Statistics: Picturing the World (6th ed.) Boston, MA: Pearson Education, Inc. 2015.
- Levine, D.M., Stephan, D.F., Krehbiel, T.C., and Berenson, M.L. Statistics for Managers Using Microsoft Excel (6th ed.). Boston, MA: Prentice Hall/Pearson, 2011.
- Starnes, D.S., Tabor, J., Yates, D.S., and Moore, D.S. The Practice of Statistics For the AP Exam (5th ed.) New York: W.H. Freeman and Company 2015.

Chapter 7

Multiple Correlation and Multiple Regression



There are many times in the engineering sciences when you want to predict a criterion, Y , but you want to find out if you can develop a better prediction model by using *several predictors* in combination (e.g. X_1, X_2, X_3 , etc.) instead of a single predictor, X .

The resulting statistical procedure is called “multiple correlation” because it uses two or more predictors in combination to predict Y , instead of a single predictor, X . Each predictor is “weighted” differently based on its separate correlation with Y and its correlation with the other predictors. The job of multiple correlation is to produce a regression equation that will weight each predictor differently and in such a way that the combination of predictors does a better job of predicting Y than any single predictor by itself. We will call the multiple correlation: R_{xy} .

You will recall (see Sect. 6.5.3) that the regression equation that predicts Y when only one predictor, X , is used is:

$$Y = a + bX \tag{7.1}$$

7.1 Multiple Regression Equation

The multiple regression equation follows a similar format and is:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \text{etc. depending on the number of predictors used} \tag{7.2}$$

The “weight” given to each predictor in the equation is represented by the letter “ b ” with a subscript to correspond to the same subscript on the predictors.

You will remember from Chap. 6 that the correlation, r , ranges from $+1$ to -1 . However, the multiple correlation, R_{xy} , only ranges from zero to $+1$. R_{xy} is never a negative number!

Important note: In order to do multiple regression, you need to have installed the “Data Analysis ToolPak” that was described in Chap. 6 (see Sect. 6.5.1). If you did not install this, you need to do so now.

Let’s try a practice problem.

Suppose that you have been asked to analyze some data from the SAT Reasoning Test (formerly called the Scholastic Aptitude Test) which is a standardized test for college admissions in the U.S. This test is intended to measure a student’s readiness for academic work in college, and about 1.4 million high school students take this test every year. There are three subtest scores generated from this test: Critical Reading, Writing, and Mathematics, and each of these subtests has a score range between 200 and 800 with an average score of about 500.

Suppose that a nearby selective college in the northeast of the U.S. that is near to you wants to determine the relationship between SAT Reading scores, SAT Writing scores, and SAT Math scores in their ability to predict freshman grade-point average (FROSH GPA) for Engineering majors at the end of freshman year at this college, and that this college has asked you to determine this relationship.

You have decided to use the three subtest scores as the predictors, X_1 , X_2 , and X_3 and the freshman grade-point average (FROSH GPA) as the criterion, Y . To test your Excel skills, you have selected 11 Engineering majors randomly from last year’s freshmen class, and have recorded their scores on these variables.

Let’s use the following notation:

Y	FROSH GPA
X_1	READING SCORE
X_2	WRITING SCORE
X_3	MATH SCORE

Suppose, further, that you have collected the following hypothetical data summarizing these scores (see Fig. 7.1):

	A	B	C	D	E
1					
2	SAT REASONING TEST				
3					
4	Is there a relationship between SAT scores and Freshman GPA at a local college?				
5					
6	FROSH GPA	READING SCORE	WRITING SCORE	MATH SCORE	
7	2.55	250	230	220	
8	3.05	610	240	440	
9	3.55	620	540	530	
10	2.05	420	420	260	
11	2.45	320	520	320	
12	2.95	630	620	620	
13	3.15	650	540	530	
14	3.45	520	580	560	
15	3.30	420	490	630	
16	2.75	330	220	610	
17	3.65	440	570	660	
18					

Fig. 7.1 Worksheet Data for SAT versus FROSH GPA (Practical Example)

Create an Excel spreadsheet for these data using the following cell reference:

- A2: SAT REASONING TEST
- A4: Is there a relationship between SAT scores and Freshman GPA at a local college?
- A6: FROSH GPA
- A7: 2.55
- B6: READING SCORE
- C6: WRITING SCORE
- D6: MATH SCORE
- D17: 660

Next, change the column width to match the above table, and change all GPA figures to number format (two decimal places).

Now, fill in the additional data in the chart such that:

- A17: 3.65
- B17: 440
- C17: 570

Then, center all numbers in your table

Important note: Be sure to double-check all of your numbers in your table to be sure that they are correct, or your spreadsheets will be incorrect.

Save this file as: GPA25

Before we do the multiple regression analysis, we need to try to make one important point very clear:

Important: When we used one predictor, X, to predict one criterion, Y, we said that you need to make sure that the X variable is ON THE LEFT in your table, and the Y variable is ON THE RIGHT in your table so that you don't get these variables mixed up (see Sect. 6.3).

However, in multiple regression, you need to follow this rule which is exactly the opposite:

When you use several predictors in multiple regression, it is essential that the criterion you are trying to predict, Y, be ON THE FAR LEFT, and all of the predictors are TO THE RIGHT of the criterion, Y, in your table so that you know which variable is the criterion, Y, and which variables are the predictors. If you make this a habit, you will save yourself a lot of grief.

Notice in the table above, that the criterion Y (FROSH GPA) is on the far left of the table, and the three predictors (READING SCORE, WRITING SCORE, and MATH SCORE) are to the right of the criterion variable. If you follow this rule, you will be less likely to make a mistake in this type of analysis.

7.2 Finding the Multiple Correlation and the Multiple Regression Equation

Objective: To find the multiple correlation and multiple regression equation using Excel.

You do this by the following commands:

Data

Click on: Data Analysis (far right top of screen)

Regression (scroll down to this in the box; see Fig. 7.2)

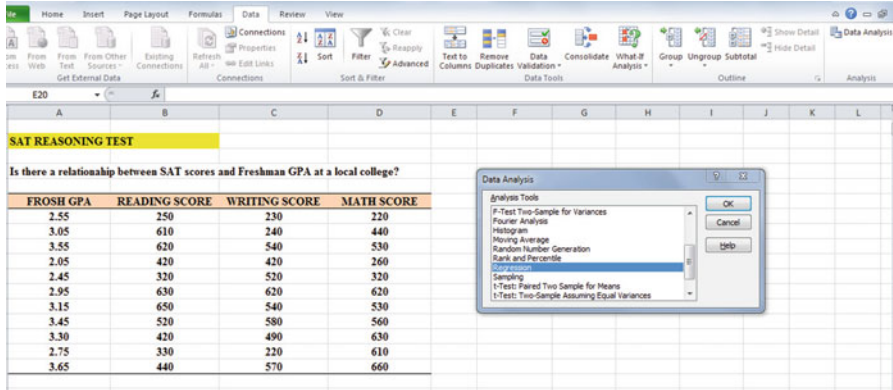


Fig. 7.2 Dialogue Box for Regression Function

OK

Input Y Range: A6:A17

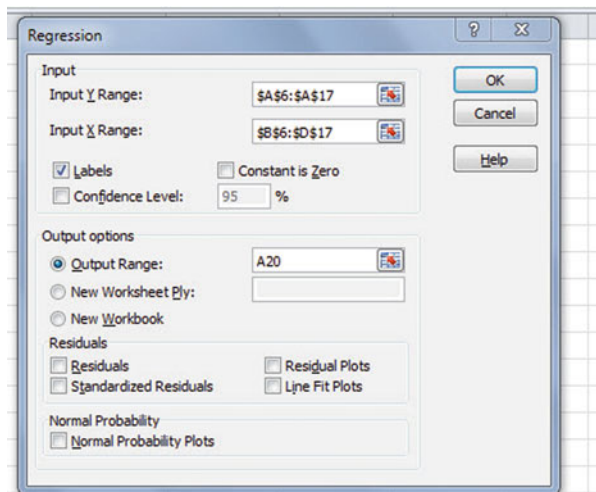
Input X Range: B6:D17

Note that both the input Y Range and the Input X Range above both include the label at the top of the columns.

Click on the Labels box to *add a check mark* to it (because you have included the column labels in row 6)

Output Range (click on the button to its left, and enter): A20 (see Fig. 7.3)

Fig. 7.3 Dialogue Box for SAT vs. FROSH GPA Data



Important note: Excel automatically assigns a dollar sign \$ in front of each column letter and each row number so that you can keep these ranges of data constant for the regression analysis.

OK (see Fig. 7.4 to see the resulting SUMMARY OUTPUT)

	A	B	C	D	E	F	G	H	I
17	3.65	440	570	660					
18									
19									
20	SUMMARY OUTPUT								
21									
22	Regression Statistics								
23	Multiple R	0.797651156							
24	R Square	0.636247266							
25	Adjusted R Square	0.48035338							
26	Standard Error	0.361446932							
27	Observations	11							
28									
29	ANOVA								
30		df	SS	MS	F	Significance F			
31	Regression	3	1.599583719	0.533194573	4.081282	0.057174747			
32	Residual	7	0.91450719	0.130643884					
33	Total	10	2.514090909						
34									
35		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
36	Intercept	1.53627108	0.468442063	3.279532734	0.013496	0.428581617	2.643960543	0.428581617	2.643960543
37	READING SCORE	0.000642945	0.000963026	0.667629207	0.525762	-0.001634251	0.00292014	-0.001634251	0.00292014
38	WRITING SCORE	0.000264354	0.000889915	0.297055329	0.775046	-0.00183996	0.002368667	-0.00183996	0.002368667
39	MATH SCORE	0.00210733	0.000848684	2.4830572	0.042022	0.000100512	0.004114149	0.000100512	0.004114149

Fig. 7.4 Regression SUMMARY OUTPUT of SAT vs. FROSH GPA Data

Next, format cell B23 in number format (two decimal places)

Next, format the following four cells in Number format (four decimal places):

- B36
- B37
- B38
- B39

Change all other decimal figures to two decimal places, and center all figures within their cells.

Save the file as: GPA26

Now, print the file so that it fits onto one page by changing the scale to 60% size. The resulting regression analysis is given in Fig. 7.5.

SAT REASONING TEST								
Is there a relationship between SAT scores and Freshman GPA at a local college?								
FROSH GPA	READING SCORE	WRITING SCORE	MATH SCORE					
2.55	250	230	220					
3.05	610	240	440					
3.55	620	540	530					
2.05	420	420	260					
2.45	320	520	320					
2.95	630	620	620					
3.15	650	540	530					
3.45	520	580	560					
3.30	420	490	630					
2.75	330	220	610					
3.65	440	570	660					
SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.80							
R Square	0.64							
Adjusted R Square	0.48							
Standard Error	0.36							
Observations	11							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	3	1.60	0.53	4.08	0.06			
Residual	7	0.91	0.13					
Total	10	2.51						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	1.5363	0.47	3.28	0.01	0.43	2.64	0.43	2.64
READING SCORE	0.0006	0.00	0.67	0.53	0.00	0.00	0.00	0.00
WRITING SCORE	0.0003	0.00	0.30	0.78	0.00	0.00	0.00	0.00
MATH SCORE	0.0021	0.00	2.48	0.04	0.00	0.00	0.00	0.00

Fig. 7.5 Final Spreadsheet for SAT vs. FROSH GPA Regression Analysis

Once you have the SUMMARY OUTPUT, you can determine the multiple correlation and the regression equation that is the best-fit line through the data points using READING SCORE, WRITING SCORE, AND MATH SCORE as the three predictors, and FROSH GPA as the criterion.

Note on the SUMMARY OUTPUT where it says: “Multiple R.” This term is correct since this is the term Excel uses for the multiple correlation, which is +0.80. This means, that from these data, that the combination of READING SCORES, WRITING SCORES, AND MATH SCORES together form a very strong positive relationship in predicting FROSH GPA.

To find the regression equation, notice the coefficients at the bottom of the SUMMARY OUTPUT:

Intercept: a (this is the y -intercept)	1.5363
READING SCORE: b_1	0.0006
WRITING SCORE: b_2	0.0003
MATH SCORE: b_3	0.0021

Since the general form of the multiple regression equation is:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 \quad (7.2)$$

we can now write the multiple regression equation for these data:

$$Y = 1.5363 + 0.0006 X_1 + 0.0003 X_2 + 0.0021X_3$$

7.3 Using the Regression Equation to Predict FROSH GPA

Objective: To find the predicted FROSH GPA using an SAT Reading Score of 600, an SAT Writing Score of 500, and an SAT Math Score of 550

Plugging these three numbers into our regression equation gives us:

$$Y = 1.5363 + 0.0006(600) + 0.0003(500) + 0.0021(550)$$

$$Y = 1.5363 + 0.36 + 0.15 + 1.155$$

$$Y = 3.20 \text{ (since GPA scores are typically measured in two decimals)}$$

If you want to learn more about the theory behind multiple regression, see Larson and Farber (2015) and Ledolter and Hogg (2010).

7.4 Using Excel to Create a Correlation Matrix in Multiple Regression

The final step in multiple regression is to find the correlation between all of the variables that appear in the regression equation.

In our example, this means that we need to find the correlation between each of the six pairs of variables:

To do this, we need to use Excel to create a “correlation matrix.” This matrix summarizes the correlations between all of the variables in the problem.

Objective: To use Excel to create a correlation matrix between the four variables in this example.

To use Excel to do this, use these steps:

Data (top of screen under “Home” at the top left of screen)

Data Analysis

Correlation (scroll *up* to highlight this formula; see Fig. 7.6)

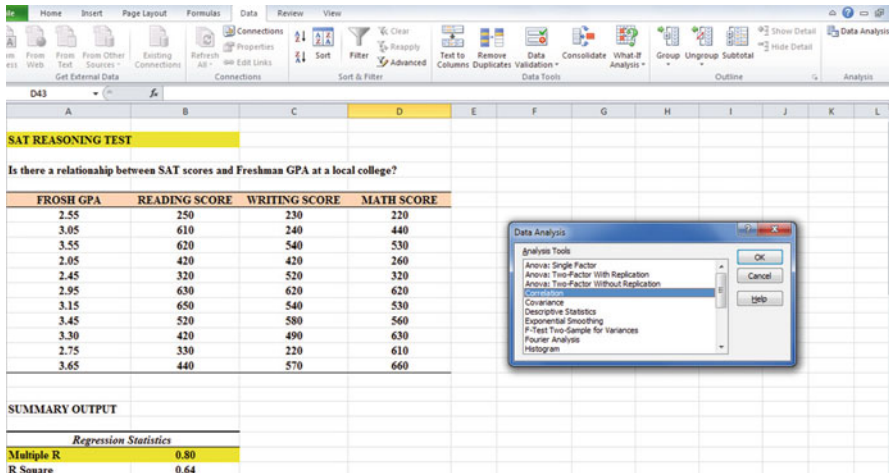


Fig. 7.6 Dialogue Box for SAT vs. FROSH GPA Correlations

OK

Input range: A6:D17

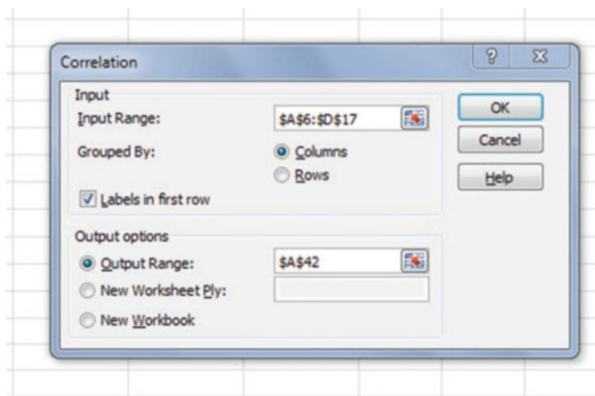
(Note that this input range includes the labels at the top of the FOUR variables (FROSH GPA, READING SCORE, WRITING SCORE, MATH SCORE) as well as all of the figures in the original data set.)

Grouped by: Columns

Put a check in the box for: Labels in the First Row (since you included the labels at the top of the columns in your input range of data above)

Output range (click on the button to its left, and enter): A42 (see Fig. 7.7)

Fig. 7.7 Dialogue Box for Input/Output Range for Correlation Matrix



OK

The resulting correlation matrix appears in A42:E46 (See Fig. 7.8).

40				
41				
42		FROSH GPA	READING SCORE	WRITING SCORE
43	FROSH GPA	1		
44	READING SCORE	0.510369686	1	
45	WRITING SCORE	0.446857676	0.468105152	1
46	MATH SCORE	0.772523347	0.444074496	0.429202393
47				

Fig. 7.8 Resulting Correlation Matrix for SAT Scores vs. FROSH GPA Data

Next, format all of the numbers in the correlation matrix that are in decimals to two decimals places. And, also, make column E wider so that the MATH SCORE label fits inside cell E42.

Save this Excel file as: GPA27

The final spreadsheet for these scores appears in Fig. 7.9.

SAT REASONING TEST								
Is there a relationship between SAT scores and Freshman GPA at a local college?								
FROSH GPA	READING SCORE	WRITING SCORE	MATH SCORE					
2.55	250	230	220					
3.05	610	240	440					
3.55	620	540	530					
2.05	420	420	260					
2.45	320	520	320					
2.95	630	620	620					
3.15	650	540	530					
3.45	520	580	560					
3.30	420	490	630					
2.75	330	220	610					
3.65	440	570	660					
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.80							
R Square	0.64							
Adjusted R Square	0.48							
Standard Error	0.36							
Observations	11							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	3	1.60	0.53	4.08	0.06			
Residual	7	0.91	0.13					
Total	10	2.51						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	1.5363	0.47	3.28	0.01	0.43	2.64	0.43	2.64
READING SCORE	0.0006	0.00	0.67	0.53	0.00	0.00	0.00	0.00
WRITING SCORE	0.0003	0.00	0.30	0.78	0.00	0.00	0.00	0.00
MATH SCORE	0.0021	0.00	2.48	0.04	0.00	0.00	0.00	0.00
	FROSH GPA	READING SCORE	WRITING SCORE	MATH SCORE				
FROSH GPA	1							
READING SCORE	0.51	1						
WRITING SCORE	0.45	0.47	1					
MATH SCORE	0.77	0.44	0.43	1				

Fig. 7.9 Final Spreadsheet for SAT Scores vs. FROSH GPA Regression and the Correlation Matrix

Note that the number “1” along the diagonal of the correlation matrix means that the correlation of each variable with itself is a perfect, positive correlation of 1.0.

Correlation coefficients are always expressed in just two decimal places.

You are now ready to read the correlation between the six pairs of variables:

The correlation between READING SCORE and FROSH GPA is: +.51

The correlation between WRITING SCORE and FROSH GPA is: +.45

The correlation between MATH SCORE and FROSH GPA is: +.77

The correlation between WRITING SCORE and READING SCORE is: +.47

The correlation between MATH SCORE and READING SCORE is: +.44

The correlation between MATH SCORE and WRITING SCORE is: +.43

This means that the best predictor of FROSH GPA is the MATH SCORE with a correlation of +.77. Adding the other two predictor variables, READING SCORE and WRITING SCORE, improved the prediction by only 0.03 to 0.80, and was, therefore, only slightly better in prediction. MATH SCORES are an excellent predictor of FROSH GPA all by themselves.

If you want to learn more about the correlation matrix, see Larson and Farber (2015).

7.5 End-of-Chapter Practice Problems

1. The Graduate Record Examinations (GRE) are a standardized test that is an admissions requirement for many U.S. graduate schools. The test is intended to measure general academic preparedness, regardless of specialization field. The General GRE test produces three subtest scores: (1) GRE Verbal Reasoning (scale 130–170), (2) GRE Quantitative Reasoning (scale 130–170), and (3) Analytical Writing (scale 0–6).

Suppose that you have been asked by the Chair of the Engineering Department at a selective graduate university to see how well the GRE predicts GPA at the end of the first year of graduate study for Engineering majors. The Chair has asked you to use the three subtest scores of the GRE as predictors, and, in addition, to use the GRE Math Test score (score range 200–990) as an additional predictor of this GPA. The Chair would like your recommendation as to whether or not the Math Test should become an admissions requirement in addition to the GRE for admission to the program in Engineering. The GRE Math Test includes problems in algebra, linear algebra, abstract algebra, number theory, and differential and integral calculus.

You have decided to use a multiple correlation and multiple regression analysis, and to test your Excel skills, you have collected the data of a random sample of 12 Engineering students who have just finished their first year of graduate study at this university. These hypothetical data appear in Fig. 7.10:

GRADUATE RECORD EXAMINATIONS				
How well does the GRE and the GRE subject area test in Math predict GPA at the end of the first year of a Masters' program in Engineering?				
FIRST-YEAR GPA	GRE VERBAL	GRE QUANTITATIVE	GRE WRITING	GRE MATH
3.25	160	161	5	650
3.42	156	158	4	600
2.85	156	157	2	500
2.65	154	153	1	510
3.65	166	166	6	630
3.16	159	160	3	550
3.56	166	163	4	610
2.35	155	154	2	430
2.86	153	154	3	450
2.95	158	157	4	550
3.15	158	159	4	580
3.45	160	160	5	620

Fig. 7.10 Worksheet Data for Chap. 7: Practice Problem #1

- Create an Excel spreadsheet using FIRST-YEAR GPA as the criterion (Y), and GRE VERBAL (X_1), GRE QUANTITATIVE (X_2), GRE WRITING (X_3), and GRE MATH (X_4) as the predictors.
- Use Excel's *multiple regression* function to find the relationship between these five variables and place it below the table.
- Use number format (two decimal places) for the multiple correlation on the SUMMARY OUTPUT, and use four decimal places for the coefficients in the SUMMARY OUTPUT
- Print the table and regression results below the table so that they fit onto one page.
- Save this file as: GRE36

Answer the following questions using your Excel printout:

- What is the multiple correlation R_{xy} ?
- What is the y-intercept a ?
- What is the coefficient for GRE VERBAL b_1 ?
- What is the coefficient for GRE QUANTITATIVE b_2 ?
- What is the coefficient for GRE WRITING b_3 ?
- What is the coefficient for GRE MATH b_4 ?
- What is the multiple regression equation?
- Predict the FIRST-YEAR GPA you would expect for a GRE VERBAL score of 150, a GRE QUANTITATIVE score of 160, a GRE WRITING score of 3, and a GRE MATH score of 610.

- (f) Now, go back to your Excel file and create a *correlation matrix* for these five variables, and place it underneath the SUMMARY OUTPUT.
- (g) Re-save this file as: GRE36
- (h) Now, print out *just this correlation matrix* on a separate sheet of paper.

Answer the following questions using your Excel printout. Be sure to include the plus or minus sign for each correlation:

9. What is the correlation between GRE VERBAL and FIRST-YEAR GPA?
 10. What is the correlation between GRE QUANTITATIVE and FIRST-YEAR GPA?
 11. What is the correlation between GRE WRITING and FIRST-YEAR GPA?
 12. What is the correlation between GRE MATH and FIRST-YEAR GPA?
 13. What is the correlation between GRE WRITING and GRE VERBAL?
 14. What is the correlation between GRE VERBAL and GRE MATH?
 15. Discuss which of the four predictors is the best predictor of FIRST-YEAR GPA.
 16. Explain in words how much better the four predictor variables together predict FIRST-YEAR GPA than the best single predictor by itself.
2. Suppose that you are the Marketing Manager of a cosmetics company, and that you have been asked to determine the effectiveness of three different types of your advertising dollars on the dollar sales of your cosmetics product line (\$000) based on three predictors: (1) Radio Ads (\$000), (2) local newspaper advertising (\$000), and (3) local television advertising (\$000). Suppose that you have completed a test market in nine cities and recorded the hypothetical data given in Fig. 7.11.

Sales (\$ 000)	Radio Ads (\$ 000)	Newspaper Ads (\$ 000)	TV Ads (\$ 000)
943	22	25	40
1212	24	26	42
937	26	27	44
651	28	29	48
890	30	28	49
923	32	24	52
1243	34	29	51
1652	36	32	53
1044	38	31	59

Fig. 7.11 Worksheet Data for Chap. 7: Practice Problem #2

- (a) create an Excel spreadsheet using Sales (\$000) as the criterion, and the other variables as the three predictors of this criterion.
 - (b) Use Excel's **multiple regression** function to find the relationship between these variables and place it below the table.
 - (c) Use number format (two decimal places) for the multiple correlation on the Summary Output, and use two decimal places for the coefficients in the SUMMARY OUTPUT
 - (d) Print the table and regression results below the table so that they fit onto one page.
 - (e) By hand on this printout, **circle and label:**
 - (1a) multiple correlation R
 - (2b) coefficients for the y-intercept, Radio Ads, Newspaper Ads, and TV Ads.
 - (f) On a separate sheet of paper by hand, write the multiple regression equation
 - (g) Underneath this regression equation by hand, predict the sales you would expect for Radio Ads of \$28,000, Newspaper Ads of \$24,000, and TV ads of \$49,000.
 - (h) Save this file as: **PRODUCT4**
 - (i) Now, go back to your Excel file and create a correlation matrix for these four variables, and place it underneath the SUMMARY OUTPUT. Use two decimal places for all correlations.
 - (j) Now, print out **just this correlation matrix** on a separate sheet of paper.
 - (k) By hand underneath the correlation matrix printed out in part j, write the answer to the following questions (label your answer as 1a, 2b, 3c, etc.). Be sure to include the plus or minus sign for each correlation:
 - (1a) What is the correlation between Radio Ads and Sales?
 - (2b) What is the correlation between Newspaper Ads and sales?
 - (3c) What is the correlation between TV Ads and sales?
 - (4d) What is the correlation between TV Ads and Radio Ads?
 - (5e) What is the correlation between Newspaper Ads and Radio Ads?
 - (6f) Discuss which predictor is the best single predictor of sales.
 - (l) Explain, in words, how much better the three variables together predict sales than the best single predictor among the predictor variables.
 - (9i) Re-save the file as: **PRODUCT4**
3. The Advanced Placement (AP) Tests are standardized tests that allow high school students to “test out” of college courses, either by receiving credit for these courses or by allowing the students to waive courses in which their AP scores are very high. The AP Exams are scored on a 5-point scale in which a score of “1” means “No recommendation” and a score of “5” means “Extremely well qualified.” There are 34 AP courses for which these exams are available. The AP Calculus BC Test is intended to measure the equivalent of a one-year course in the calculus of functions of a single variable, and before studying calculus,

students are encouraged to complete four years of secondary school mathematics that are designed for college-bound students.

Suppose that you have been asked by the Chair of the Engineering Department at a selective university to see how well the AP Calculus BC Test, along with other predictors, predicts GPA at the end of the first year of study for Engineering majors. This Chair has asked you to use as predictors: (1) SAT-Reading scores, (2) High School GPA, and (3) AP CALCULUS BC Test scores.

You have decided to use a multiple correlation and multiple regression analysis, and to test your Excel skills, you have collected the data of a random sample of 10 Engineering students who have just finished their first year of study at this university.

These hypothetical data appear in Fig. 7.12.

SAT EXAM			
FROSH GPA	SAT READING	HS GPA	AP CALCULUS BC
3.23	650	3.55	2
2.90	490	2.96	2
2.80	630	3.27	5
3.42	520	3.45	4
2.80	560	2.90	3
2.90	410	2.80	2
2.35	450	2.63	2
2.58	420	2.71	1
3.12	560	3.26	3
3.47	650	3.45	4

Fig. 7.12 Worksheet Data for Chap. 7: Practice Problem #3

- (a) create an Excel spreadsheet using FROSH GPA as the criterion and the other three variables as the predictors.
- (b) Use Excel’s *multiple regression* function to find the relationship between these five variables and place the SUMMARY OUTPUT below the table.
- (c) Use number format (two decimal places) for the multiple correlation on the Summary Output, and use number format (three decimal places) for the coefficients in the summary output, and three decimal places for all other decimal figures in the SUMMARY OUTPUT.
- (d) Save the file as: SAT11
- (e) Print the table and regression results below the table so that they fit onto one page.

Answer the following questions using your Excel printout:

1. What is multiple correlation R_{xy} ?
 2. What is the y-intercept a ?
 3. What is the coefficient for SAT-READING b_1 ?
 4. What is the coefficient for HS GPA b_2 ?
 5. What is the coefficient for AP CALCULUS BC b_3 ?
 6. What is the multiple regression equation?
 7. Predict the FROSH GPA you would expect for an SAT-READING score of 650, a HS GPA of 3.47, and an AP Calculus BC Test score of 4.
- (f) Now, go back to your Excel file and create a correlation matrix for these four variables, and place it underneath the SUMMARY OUTPUT on your spreadsheet.
- (g) Save this file as: SAT12
- (h) Now, print out *just this correlation matrix* on a separate sheet of paper.

Answer the following questions using your Excel printout. Be sure to include the plus or minus sign for each correlation:

8. What is the correlation between SAT-READING and FROSH GPA?
9. What is the correlation between HS GPA and FROSH GPA?
10. What is the correlation between AP CALCULUS BC and FROSH GPA?
11. What is the correlation between HS GPA and SAT-READING?
12. What is the correlation between SAT READING and AP Calculus BC?
13. What is the correlation between HS GPA and AP Calculus BC?
14. Discuss which of the three predictors is the best predictor of FROSH GPA.
15. Explain in words how much better the three predictor variables combined predict FROSH GPA than the best single predictor by itself.

References

- Larson, R. and Farber, B. Elementary Statistics: Picturing the World (6th ed.) Boston, MA: Pearson Education, Inc. 2015.
- Ledolter J, Hogg R. Applied Statistics for Engineers and Physical Scientists. 3rd ed. Upper Saddle River: Pearson Prentice Hall; 2010.

Chapter 8

One-Way Analysis of Variance (ANOVA)



So far in this 2016 Excel Guide, you have learned how to use a one-group t-test to compare the sample mean to the population mean, and a two-group t-test to test for the difference between two sample means. *But what should you do when you have more than two groups and you want to determine if there is a significant difference between the means of these groups?*

The answer to this question is: *Analysis of Variance (ANOVA)*.

The ANOVA test allows you to test for the difference between the means when you have *three or more groups* in your research study.

Important note: In order to do One-way Analysis of Variance, you need to have installed the “Data Analysis Toolpak” that was described in Chap. 6 (see Sect. 6.5.1). If you did not install this, you need to do that now.

Let’s suppose that you were working as a research scientist for a company and that you wanted to compare your company’s premium brand of tire (Brand A) against two major competitors’ brands (B and C). You have set up a laboratory test of the three types of tires, and you have measured the number of simulated miles driven before the tread length reached a predetermined amount. The hypothetical results are given in Fig. 8.1. Note that the data are in thousands of miles driven (1000), so, for example, 63 is really 63,000 miles.

You have been asked to analyze the data to determine if there was any significant difference in miles driven between the three brands. To test your Excel skills, you have selected a random sample of tires from each of these brands (see Fig. 8.1). Note that each brand can have a different number of tires in order for ANOVA to be used on the data. Statisticians delight in this fact by referring to this characteristic by stating that: “ANOVA is a very robust test.” (Statisticians love that term!)

TIRE MILEAGE TEST			
(Data are in thousands of miles)			
	Brand A (1000 miles)	Brand B (1000 miles)	Brand C (1000 miles)
	62	61	65
	61	62	67
	62	63	71
	64	60	66
	61	64	65
		59	64
		62	
		63	
		62	
		63	

Fig. 8.1 Worksheet Data for Tire Mileage Test (Practical Example)

Create an Excel spreadsheet for these data in this way:

- A4: TIRE MILEAGE TEST
- A6: (Data are in thousands of miles)
- B8: Brand A (1000 miles)
- C8: Brand B (1000 miles)
- D8: Brand C (1000 miles)
- B9: 62

Enter the other information into your spreadsheet table. When you have finished entering these data, the last cell on the left should have 61 in cell B13, and the last cell on the right should have 64 in cell D14. Center the numbers in each of the columns. Use number format (zero decimals) for all numbers.

Important note: Be sure to double-check all of your figures in the table to make sure that they are exactly correct or you will not be able to obtain the correct answer for this problem!

Save this file as: TIRE6

8.1 Using Excel to Perform a One-Way Analysis of Variance (ANOVA)

Objective: To use Excel to perform a one-way ANOVA test.

You are now ready to perform an ANOVA test on these data using the following steps:

Data (at top of screen)

Data Analysis (far right at top of screen)

Anova: Single Factor (*scroll up to this formula and highlight it; see Fig. 8.2*)

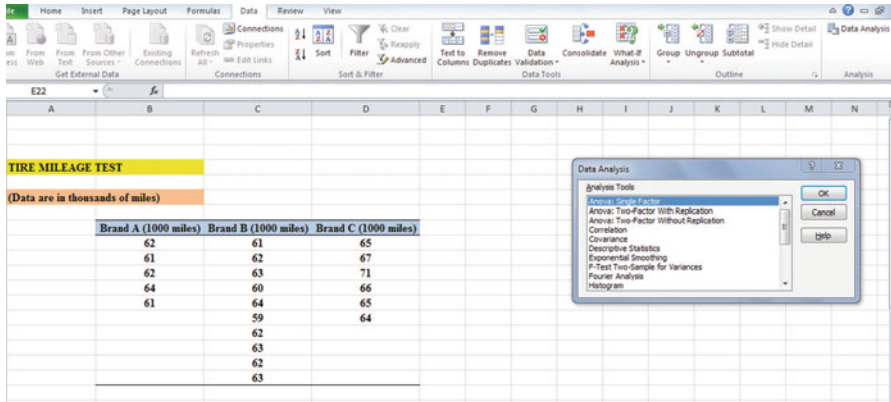


Fig. 8.2 Dialog Box for Data Analysis: Anova Single Factor

OK

Input range: B8:D18 (note that you have included in this range the column titles that are in row 8)

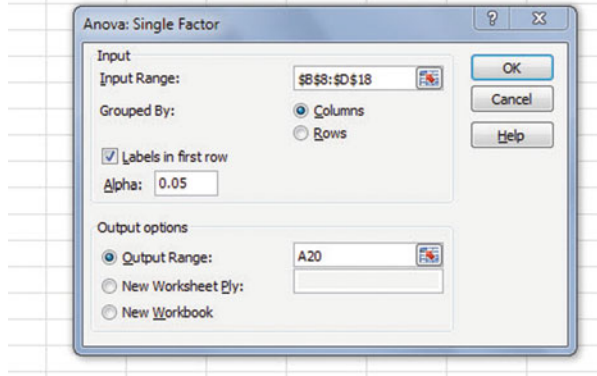
Important note: Whenever the data set has a different sample size in the groups being compared, the INPUT RANGE that you define must start at the column title of the first group on the left and go to the last column on the right to the lowest row that has a figure in it in the entire data matrix so that the INPUT RANGE has the “shape” of a rectangle when you highlight it. Since Brand B has 63 in cell C18, your “rectangle” must include row 18!

Grouped by: Columns

Put a check mark in: Labels in First Row

Output range (click on the button to its left): A20 (see Fig. 8.3)

Fig. 8.3 Dialog Box for Anova: Single Factor Input/ Output Range



OK

Center all of the numbers in the ANOVA table, and round off all numbers that are decimals to two decimal places.

Save this file as: TIRE6A

You should have generated the table given in Fig. 8.4.

TIRE MILEAGE TEST						
(Data are in thousands of miles)						
	Brand A (1000 miles)	Brand B (1000 miles)	Brand C (1000 miles)			
	62	61	65			
	61	62	67			
	62	63	71			
	64	60	66			
	61	64	65			
		59	64			
		62				
		63				
		62				
		63				
		62				
		63				
Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Brand A	5	310	62.00	1.50		
Brand B	10	619	61.90	2.32		
Brand C	6	398	66.33	6.27		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	83.00	2	41.50	12.83	0.00	3.55
Within Groups	58.23	18	3.24			
Total	141.24	20				

Fig. 8.4 ANOVA Results for Tire Mileage Test

Print out both the data table and the ANOVA summary table so that all of this information fits onto one page. (Hint: Set the Page Layout/Fit to Scale to 85% size).

As a check on your analysis, you should have the following in these cells:

A20: ANOVA: Single Factor

D24: 62.00

D31: 41.50

E31: 12.83

G31: 3.55

Now, let's discuss how you should interpret this table:

8.2 How to Interpret the ANOVA Table Correctly

Objective: To interpret the ANOVA table correctly

ANOVA allows you to test for the differences between means when you have three or more groups of data. This ANOVA test is called the F-test statistic, and is typically identified with the letter: F.

The formula for the F-test is this:

$$F = \frac{\text{Mean Square between groups (MS}_b\text{)}}{\text{Mean Square within groups (MS}_w\text{)}} \quad (8.1)$$

The derivation and explanation of this formula is beyond the scope of this *Excel Guide*. In this *Excel Guide*, we are attempting to teach you *how to use Excel*, and we are not attempting to teach you the statistical theory that is behind the ANOVA formulas. For a detailed explanation of ANOVA, see Larson and Farber (2015).

Note that cell D31 contains $MS_b = 41.50$, while cell D32 contains $MS_w = 3.24$.

When you divide these two figures using their cell references in Excel, you get the answer for the F-test of 12.83 which is in cell E31. (Remember, Excel is more accurate than your calculator!) Let's discuss now the meaning of the figure: $F = 12.83$.

In order to determine whether this figure for F of 12.83 indicates a significant difference between the means of the three groups, the first step is to write the null hypothesis and the research hypothesis for the three brands of tires.

In our statistics mileage comparisons, the null hypothesis states that the population means of the three groups are equal, while the research hypothesis states that the population means of the three groups are not equal and that there is, therefore, a significant difference between the population means of the three groups. Which of these two hypotheses should you accept based on the ANOVA results?

8.3 Using the Decision Rule for the ANOVA F-Test

To state the hypotheses, let's call Brand A as Group 1, Brand B as Group 2, and Brand C as Group 3. The hypotheses would then be:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3$$

The answer to this question is analogous to the decision rule used in this book for both the one-group t-test and the two-group t-test. You will recall that this rule (See Sects. 4.1.6 and 5.1.8) was:

If the absolute value of t is less than the critical t , you accept the null hypothesis.

or

If the absolute value of t is greater than the critical t , you reject the null hypothesis, and accept the research hypothesis.

Now, here is the decision rule for ANOVA:

Objective: To learn the decision rule for the ANOVA F-test

The decision rule for the ANOVA F-test is the following:

If the value for F is less than the critical F -value, accept the null hypothesis.

or

If the value of F is greater than the critical F -value, reject the null hypothesis, and accept the research hypothesis.

Note that Excel tells you the critical F -value in cell G31: 3.55

Therefore, our decision rule for the tire mileage ANOVA test is this:

Since the value of F of 12.83 is greater than the critical F -value of 3.55, we reject the null hypothesis and accept the research hypothesis.

Therefore, our conclusion, in plain English, is:

There was a significant difference between the number of miles driven between the three brands of tires.

Note that it is not necessary to take the absolute value of F of 12.83. The F -value can never be less than one, and so it can never be a negative value which requires us to take its absolute value in order to treat it as a positive value.

It is important to note that ANOVA tells us that there was a significant difference between the population means of the three groups, *but it does not tell us which pairs of groups were significantly different from each other.*

8.4 Testing the Difference Between Two Groups Using the ANOVA t-Test

To answer that question, we need to do a different test called the ANOVA t-test.

Objective: To test the difference between the means of two groups using an ANOVA t-test when the ANOVA F-test results indicate a significant difference between the population means.

Since we have three groups of data (one group for each of the three brands of tires), we would have to perform three separate ANOVA t-tests to determine which pairs of groups were significantly different. This requires that we would have to perform a separate ANOVA t-test for the following pairs of groups:

1. Brand A vs. Brand B
2. Brand A vs. Brand C
3. Brand B vs. Brand C

We will do just one of these pairs of tests, Brand A vs. Brand C, to illustrate the way to perform an ANOVA t-test comparing these two brands of tires. The ANOVA t-test for the other two pairs of groups would be done in the same way.

8.4.1 Comparing Brand A vs. Brand C in Miles Driven Using the ANOVA t-Test

Objective: To compare Brand A vs. Brand C in miles driven using the ANOVA t-test.

The first step is to write the null hypothesis and the research hypothesis for these two brands of tires.

For the ANOVA t-test, the null hypothesis is that the population means of the two groups are equal, while the research hypothesis is that the population means of the two groups are not equal (i.e., there is a significant difference between these two means). Since we are comparing Brand A (Group 1) vs. Brand C (Group 3), these hypotheses would be:

$$H_0: \mu_1 = \mu_3$$

$$H_1: \mu_1 \neq \mu_3$$

For Group 1 vs. Group 3, the formula for the ANOVA t-test is:

$$ANOVA\ t = \frac{\bar{X}_1 - \bar{X}_2}{s.e._{ANOVA}} \quad (8.2)$$

where

$$s.e._{ANOVA} = \sqrt{MS_w \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (8.3)$$

The steps involved in computing this ANOVA t-test are:

1. Find the difference of the sample means for the two groups ($62 - 66.33 = -4.33$).
2. Find $1/n_1 + 1/n_3$ (since both groups have a different number of tires in them, this becomes: $1/5 + 1/6 = 0.20 + 0.17 = 0.37$)
3. Multiply MS_w times the answer for step 2 ($3.24 \times 0.37 = 1.20$)
4. Take the square root of step 3 (SQRT (1.20) = 1.09)
5. Divide Step 1 by Step 4 to find ANOVA t ($-4.33/1.09 = -3.97$)

Note: Since Excel computes all calculations to 16 decimal places, when you use Excel for the above computations, your answer will be -3.98 in two decimal places, but Excel's answer will be much more accurate because it is always in 16 decimal places in its computations.

Now, what do we do with this ANOVA t-test result of -3.97 ? In order to interpret this value of -3.97 correctly, we need to determine the critical value of t for the ANOVA t-test. To do that, we need to find the degrees of freedom for the ANOVA t-test as follows:

8.4.1.1 Finding the Degrees of Freedom for the ANOVA t-Test

Objective: To find the degrees of freedom for the ANOVA t-test.

The degrees of freedom (df) for the ANOVA t-test is found as follows:

df = take **the total sample size of all of the groups** and subtract the number of groups in your study ($n_{TOTAL} - k$ where k = the number of groups)

In our example, **the total sample size of the three groups is 21** since there are 5 tires in Group 1, 10 tires in Group 2, and 6 tires in Group 3, and since there are three groups, $21 - 3$ gives a degrees of freedom for the ANOVA t-test of 18.

If you look up $df = 18$ in the t-table in Appendix E in the degrees of freedom column (df), which is the *second column on the left of this table*, you will find that the critical t-value is 2.101.

Important note: Be sure to use the degrees of freedom column (df) in Appendix E for the ANOVA t-test critical t value

8.4.1.2 Stating the Decision Rule for the ANOVA t-Test

Objective: To learn the decision rule for the ANOVA t-test

Interpreting the result of the ANOVA t-test follows the same decision rule that we used for both the one-group t-test (see Sect. 4.1.6) and the two-group t-test (see Sect. 5.1.8):

If the absolute value of t is less than the critical value of t , we accept the null hypothesis.

or

If the absolute value of t is greater than the critical value of t , we reject the null hypothesis and accept the research hypothesis.

Since we are using a type of t-test, we need to take the absolute value of t . Since the absolute value of -3.98 is greater than the critical t-value of 2.101, we reject the null hypothesis (that the population means of the two groups are equal) and accept the research hypothesis (that the population means of the two groups are significantly different from one another).

This means that our conclusion, in plain English, is as follows:

The average tire mileage for Brand C was significantly greater than the average tire mileage for Brand A (66,000 vs. 62,000).

Note that this difference in average tire mileage of about 4000 miles between Brand A and Brand C might not seem like much, but in practical terms, this means that the average miles driven for Brand C were 7% higher than the average miles driven for Brand A. This, clearly, is an important difference in miles driven based on our hypothetical data.

8.4.1.3 Performing an ANOVA t-Test Using Excel commands

Now, let's do these calculations for the ANOVA t-test using Excel with the file you created earlier in this chapter: TIRE6A

A37: Brand A vs. Brand C

A39: 1/5 + 1/6

A41: s.e. ANOVA

A43: ANOVA t-test

B39: =(1/5 + 1/6) (no spaces between symbols)

B41: =SQRT(D32*B39) (no spaces between symbols)

B43: =(D24 - D26)/B41 (no spaces between symbols)

You should now have the following results in these cells when you round off all these figures in the ANOVA t-test to two decimal points.

B39: 0.37

B41: 1.09

B43: -3.98

Save this final result under the file name: TIRE7

Print out the resulting spreadsheet so that it fits onto one page like Fig. 8.5 (Hint: Reduce the Page Layout/Scale to Fit to 85%).

TIRE MILEAGE TEST						
(Data are in thousands of miles)						
	Brand A (1000 miles)	Brand B (1000 miles)	Brand C (1000 miles)			
	62	61	65			
	61	62	67			
	62	63	71			
	64	60	66			
	61	64	65			
		59	64			
		62				
		63				
		62				
		63				
Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Brand A	5	310	62.00	1.50		
Brand B	10	619	61.90	2.32		
Brand C	6	398	66.33	6.27		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	83.00	2	41.50	12.83	0.0003	3.55
Within Groups	58.23	18	3.24			
Total	141.24	20				
Brand A vs. Brand C						
$1/5 + 1/6$	0.37					
s.e. ANOVA	1.09					
ANOVA t-test	-3.98					

Fig. 8.5 Final Spreadsheet of Tire Mileage for Brand A vs. Brand C

For a more detailed explanation of the ANOVA t -test, see Black (2010).

Important note: You are only allowed to perform an ANOVA t -test comparing the means of two groups when the F -test produces a significant difference between the means of all of the groups in your study.

It is improper to do any ANOVA t -test when the value of F is less than the critical value of F . Whenever F is less than the critical F , this means that there was no difference between the means of the groups, and, therefore, that you cannot test to see if there is a difference between the means of any two groups since this would capitalize on chance differences between these two groups. For more information on this important point, see Gould and Gould (2002).

8.5 End-of-Chapter Practice Problems

1. Let's suppose that an undergraduate Electrical Engineering 101 course at Missouri University of Science and Technology was taught for the spring semester in three different formats: (1) Computer-assisted Instruction (CAI), (2) In-class lectures (LECTURES), and (3) independent study in an online version (INDEPENDENT). Suppose, further, that all students took the same comprehensive final examination (100 points possible).

You have been asked to analyze the data from the final examinations to determine if there was any significant difference in final exam scores between the three types of teaching methods. To test your Excel skills, you have selected a random sample of students from each of these methods (see Fig. 8.6). Note that each group of students can be of a different number of students in order for ANOVA to be used on the data. Statisticians delight in this fact by referring to this characteristic by stating that: "ANOVA is a very robust test." (Statisticians love that term!)

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY		
UNDERGRADUATE ELECTRICAL ENGINEERING 101 COURSE: FINAL EXAM		
CAI	LECTURES	INDEPENDENT
90	85	76
85	89	80
74	83	90
89	79	84
84	74	78
95	75	65
92	86	42
65	87	58
75	86	63
73	88	75
54		66
71		

Fig. 8.6 Worksheet Data for Chap. 8: Practice Problem #1

- (a) Enter these data on an Excel spreadsheet.
- (b) Perform a *one-way ANOVA test* on these data, and show the resulting ANOVA table *underneath* the input data for the three teaching formats.
- (c) If the F-value in the ANOVA table is significant, create an Excel formula to compute the ANOVA t-test comparing the average for LECTURES against INDEPENDENT STUDY and show the results below the ANOVA table on the spreadsheet (put the standard error and the ANOVA t-test value on separate lines of your spreadsheet, and use two decimal places for each value)
- (d) Print out the resulting spreadsheet so that all of the information fits onto one page
- (e) Save the spreadsheet as: STATS8

Now, write the answers to the following questions using your Excel printout:

1. What are the null hypothesis and the research hypothesis for the ANOVA F-test?
2. What is MS_b on your Excel printout?
3. What is MS_w on your Excel printout?
4. Compute $F = MS_b/MS_w$ using your calculator.
5. What is the critical value of F on your Excel printout?
6. What is the result of the ANOVA F-test?
7. What is the conclusion of the ANOVA F-test in plain English?
8. If the ANOVA F-test produced a significant difference between the three types of teaching formats in final examination performance, what is the null hypothesis and the research hypothesis for the ANOVA t-test comparing LECTURES versus INDEPENDENT STUDY?

9. What is the mean (average) for LECTURES on your Excel printout?
10. What is the mean (average) for INDEPENDENT STUDY on your Excel printout?
11. What are the degrees of freedom (df) for the ANOVA t-test comparing LECTURES versus INDEPENDENT STUDY?
12. What is the critical t value for this ANOVA t-test in Appendix E for these degrees of freedom?
13. Compute the $s.e._{ANOVA}$ using your calculator.
14. Compute the ANOVA t-test value comparing LECTURES versus INDEPENDENT STUDY using your calculator.
15. What is the result of the ANOVA t-test comparing LECTURES versus INDEPENDENT STUDY?
16. What is the conclusion of the ANOVA t-test comparing LECTURES versus INDEPENDENT STUDY in plain English?

Note: Since there are three types of teaching formats, you need to do three ANOVA t-tests to determine what the significant differences are between the three types of teaching methods. *Since you have just completed the ANOVA t-test comparing LECTURES versus INDEPENDENT STUDY, you would also need to do the ANOVA t-test comparing LECTURES versus CAI, and also the ANOVA t-test comparing CAI versus INDEPENDENT STUDY in order to write a conclusion summarizing these three types of ANOVA t-tests.*

2. Suppose that you wanted to compare the durability of your company's golf balls (Brand Y) with the durability of two major competitors (X and Z). Suppose that the National Golf Association randomly selected golf balls from each of the three brands. A machine then was used to exert the force of a 250-yard drive from a tee, and the durability of each ball was measured in terms of the number of simulated drives needed to crack or chip each ball. Your task is to determine if there was a significant difference in the number of drives needed to crack or chip the balls of the three brands.

You decide to take a random sample of golf balls from the different brands to test your Excel skills, and the hypothetical data are given in Fig. 8.7.

Fig. 8.7 Worksheet Data for Chap. 8: Practice Problem #2

GOLF BALL DURABILITY TEST RESULTS		
BRAND X	BRAND Y	BRAND Z
257	310	242
289	315	265
292	324	275
315	336	288
328	324	262
348	318	274
327	321	263
	328	254
		277

- (a) Enter these data on an Excel spreadsheet.
- (b) On your spreadsheet, write the null hypothesis and the research hypothesis for these data
- (c) Perform a **one-way ANOVA test** on these data, and show the resulting ANOVA table underneath the input data for the three brands of golf balls.
- (d) If the F-value in the ANOVA table is significant, create an Excel formula to compute the ANOVA t-test comparing Brand Y versus Brand Z, and show the results below the ANOVA table on the spreadsheet (put the standard error and the ANOVA t-test value on separate lines of your spreadsheet, and use two decimal places for each value)
- (e) Print out the resulting spreadsheet so that all of the information fits onto one page
- (f) On your printout, label by hand the MS (between groups) and the MS (within groups)
- (g) Circle and label the value for F, and the critical value for F, on your printout for the ANOVA of the input data
- (h) Label by hand on the printout the mean Brand Y and the mean for Brand Z that were produced by your ANOVA formulas.
- (i) Save the spreadsheet as: GOLF4

On a separate sheet of paper, now do the following by hand:

- (j) write a summary of the result of the ANOVA test for the input data
 - (k) write a summary of the conclusion of the ANOVA test in plain English for the input data
 - (l) write the null hypothesis and the research hypothesis comparing Brand Y versus Brand Z for the ANOVA t-test, and use Excel to find the standard error
 - (m) compute the ANOVA t-test value with your calculator
 - (n) compute the degrees of freedom for the ANOVA t-test by hand for three brands of golf balls.
 - (o) write the critical value of t for the ANOVA t-test using the table in your Excel book
 - (p) write the result of the ANOVA t-test
 - (q) write the conclusion of the ANOVA t-test in plain English
3. Suppose that you were working as a research scientist and that you wanted to do a research study comparing the highway miles per gallon (mpg) for five types of vehicles: (1) SUBCOMPACTS, (2) COMPACTS, (3) MID-SIZE, (4) LARGE, and (5) SUVs. You want to answer the research question: Is the size of the vehicle related to gasoline usage? You have obtained the cooperation of the owners of each type of vehicle who agree to keep track of their highway mileage over a pre-determined route for three tanks of gasoline. The hypothetical data for this study are given in Fig. 8.8.

HIGHWAY MILES PER GALLON (mpg) COMPARISON OF FIVE TYPES OF CARS				
1	2	3	4	5
SUBCOMPACTS (mpg)	COMPACTS (mpg)	MID-SIZE (mpg)	LARGE (mpg)	SUVs (mpg)
28.1	26.2	24.0	22.0	18.1
30.2	28.3	26.3	23.1	20.2
29.3	29.3	25.2	25.4	22.3
31.6	27.0	27.1	24.3	21.4
33.0	28.0	28.0	25.0	20.5
34.3	29.5	23.6	24.7	19.0
32.1	31.0	29.2	23.1	18.2
35.0	32.3		22.4	19.1
	33.1		26.0	
			21.3	

Fig. 8.8 Worksheet Data for Chap. 8: Practice Problem #3

- Enter these data on an Excel spreadsheet.
- Perform a *one-way ANOVA test* on these data, and show the resulting ANOVA table *underneath* the input data for the five types of vehicles.
- If the F-value in the ANOVA table is significant, create an Excel formula to compute the ANOVA t-test comparing the average mpg for COMPACTS against the average mpg for LARGE, and show the results below the ANOVA table on the spreadsheet (put the standard error and the ANOVA t-test value on separate lines of your spreadsheet, and use two decimal places for each value)
- Print out the resulting spreadsheet so that all of the information fits onto one page
- Save the spreadsheet as: CARS3

Now, write the answers to the following questions using your Excel printout:

- What are the null hypothesis and the research hypothesis for the ANOVA F-test?
- What is MS_b on your Excel printout?
- What is MS_w on your Excel printout?
- Compute $F = MS_b/MS_w$ using your calculator.
- What is the critical value of F on your Excel printout?
- What is the result of the ANOVA F-test?
- What is the conclusion of the ANOVA F-test in plain English?
- If the ANOVA F-test produced a significant difference between the five types of vehicles in their mpg, what is the null hypothesis and the research hypothesis for the ANOVA t-test comparing COMPACTS versus LARGE?
- What is the mean (average) mpg for COMPACTS on your Excel printout?
- What is the mean (average) mpg for LARGE on your Excel printout?

11. What are the degrees of freedom (df) for the ANOVA t-test comparing COMPACTS versus LARGE?
12. What is the critical t value for this ANOVA t-test in Appendix E for these degrees of freedom?
13. Compute the $s.e._{ANOVA}$ using your calculator for COMPACTS versus LARGE.
14. Compute the ANOVA t-test value comparing COMPACTS versus LARGE using your calculator.
15. What is the result of the ANOVA t-test comparing COMPACTS versus LARGE?
16. What is the conclusion of the ANOVA t-test comparing COMPACTS versus LARGE in plain English?

References

- Black K. Business statistics: for contemporary decision making. 6th ed. Hoboken: John Wiley & Sons, Inc.; 2010.
- Gould J, Gould G. Biostats basics: a student handbook. New York: W.H. Freeman and Company; 2002.
- Larson, R. and Farber, B. Elementary Statistics: Picturing the World (6th ed.) Boston, MA: Pearson Education, Inc. 2015.

Appendices

Appendix A: Answers to End-of-Chapter Practice Problems

Chapter 1: Practice Problem #1 Answer (see Fig. A.1)

TV ADVERTISING PILOT TEST							
Panel of male college students (ages 18-24)							
Item #10:	Based on the TV commercial that you just saw, how likely are you to purchase the advertised product?						
	1	2	3	4	5	6	7
	Very Unlikely						Very Likely
		RATING					
		3					
		4	n			21	
		2					
		6					
		3	Mean			3.05	
		5					
		4					
		3	STDEV			1.47	
		6					
		2					
		1	s.e.			0.32	
		2					
		1					
		3					
		4					
		3					
		2					
		4					
		1					
		2					
		3					

Fig. A.1 Answer to Chap. 1: Practice Problem #1

Chapter 1: Practice Problem #2 Answer (see Fig. A.2)

FORD MOTOR COMPANY							
Survey of new-car features							
Panel of female college students (ages 18-24)							
Question #12:	If you were to purchase a new car today, how important to you is the feature that "the car parallel parks itself to the curb" by using a computer?						
	1	2	3	4	5	6	7
	Not Important						Very Important
	<u>RATING</u>						
		5					
		6		n		17	
		4		Mean		5.059	
		3					
		7		STDEV		1.713	
		6					
		5		s.e.		0.415	
		7					
		6					
		7					
		4					
		3					
		1					
		7					
		6					
		4					
		5					

Fig. A.2 Answer to Chap. 1: Practice Problem #2

Chapter 1: Practice Problem #3 Answer (see Fig. A.3)

TELEVISION SPOKESPERSON RATINGS BY FEMALE COLLEGE STUDENTS						
COMMERCIAL No. 512						
Item #12:	Think about the spokesperson you just saw in the commercial. How trustworthy do you think this spokesperson is in terms of the product that was advertised?					
7	6	5	4	3	2	1
Very Trustworthy						Very Untrustworthy
		RATING				
		3				
		4				
		6	n			17
		2				
		1				
		3	Mean			3.94
		5				
		1				
		7	STDEV			1.82
		6				
		5				
		3	s.e.			0.44
		4				
		2				
		6				
		5				
		4				

Fig. A.3 Answer to Chap. 1: Practice Problem #3

Chapter 2: Practice Problem #1 Answer (see Fig. A.4)

Fig. A.4 Answer to Chap. 2: Practice Problem #1

FRAME NUMBERS	Duplicate frame numbers	RANDOM NO.
1	44	0.029
2	33	0.734
3	38	0.885
4	43	0.052
5	13	0.739
6	10	0.574
7	50	0.195
8	1	0.388
9	48	0.584
10	61	0.260
11	4	0.796
12	22	0.234
13	40	0.293
14	37	0.185
15	35	0.484
16	60	0.856
17	59	0.738
18	7	0.394
19	17	1.000
20	20	0.720
21	20	0.720
	11	0.0
56	56	0.434
57	57	0.250
58	54	0.674
59	9	0.761
60	51	0.341
61	39	0.998
62	53	0.893
63	26	0.663

Chapter 2: Practice Problem #2 Answer (see Fig. A.5)

Fig. A.5 Answer to Chap. 2: Practice Problem #2

FRAME NO.	Duplicate frame no.	Random number
1	45	0.185
2	102	0.568
3	16	0.700
4	8	0.143
5	109	0.170
6	64	0.403
7	37	0.857
8	31	0.184
9	27	0.459
10	76	0.016
11	9	0.385
12	70	0.741
13	13	0.946
14	32	0.718
15	56	0.784
16	46	0.957
17	3	0.033
18	98	0.677
19	10	0.687
20	100	0.114
21	29	0.501
100	35	0.796
101	20	0.291
102	73	0.408
103	11	0.364
104	24	0.951
105	82	0.482
106	5	0.877
107	17	0.710
108	34	0.578
109	104	0.976
110	51	0.301
111	6	0.171
112	84	0.106
113	96	0.629
114	67	0.008

Chapter 2: Practice Problem #3 Answer (see Fig. A.6)

Fig. A.6 Answer to Chap. 2: Practice Problem #3

FRAME NUMBERS	Duplicate frame numbers	Random number
1	47	0.459
2	68	0.252
3	15	0.210
4	69	0.204
5	67	0.116
6	38	0.262
7	43	0.533
8	50	0.361
9	65	0.957
10	40	0.321
11	57	0.351
12	37	0.689
13	22	0.469
14	3	0.889
15	17	0.255
16	60	0.894
17	5	0.545
18	29	0.858
19	74	0.131
20	72	0.322
21	14	0.433
22		0.455
	27	0.455
71	46	0.701
72	35	0.460
73	11	0.854
74	7	0.786
75	12	0.547
76	30	0.804

Chapter 3: Practice Problem #1 Answer (see Fig. A.7)

DELTA AIRLINES CUSTOMER SERVICE DEPARTMENT					
FLIGHT CANCELLATION RE-SCHEDULING PHONE SURVEY					
One-item survey:	Would you hire the last representative you spoke with if you owned a customer service company?				
	1	2	3	4	5
	Definitely No				Definitely Yes
	RATING				
	4	Null hypothesis:		$\mu = 3$	
	3	Research hypothesis:		$\mu \neq 3$	
	5	n		16	
	3	Mean		4.06	
	4	STDEV		0.93	
	5	s.e.		0.23	
	4	95% confidence interval			
	5	Lower limit		3.57	
	3	Upper limit		4.56	
	4	Draw a diagram of the confidence interval			
	5	3	3.57	4.06	4.56
		Ref. Value	lower limit	Mean	upper limit
Result:	Since the reference value of 3 is outside of the confidence interval, we reject the null hypothesis and accept the research hypothesis.				
Conclusion:	Customers who called Delta's Customer Service Department to re-schedule a cancelled flight were significantly willing to hire the last customer service representative they spoke with if they owned a customer service company.				

Fig. A.7 Answer to Chap. 3: Practice Problem #1

Chapter 3: Practice Problem #2 Answer (see Fig. A.8)

EXPECTED LIFETIME OF A NEW TYPE OF PASSENGER CAR TIRE			
Research question:	Does this new type of synthetic tire have an expected lifetime of 40,000 miles?		
	LIFETIME IN MILES		
	38,400		
	39,500	Null hypothesis:	$\mu = 40,000$ miles
	39,400		
	42,300	Research hypothesis:	$\mu \neq 40,000$ miles
	46,700		
	45,800	n	15
	44,300		
	38,600	Mean	41020.00
	42,500		
	41,600	STDEV	2785.47
	40,200		
	38,600	s.e.	719.21
	37,900		
	38,900	95% confidence interval	
	40,600		
		lower limit	39477.46
		upper limit	42562.54
		39477 -----40000-----41020-----42563	
		lower limit	Ref. Value
			Mean
			upper limit
Result:	Since the reference value is inside the confidence interval, we accept the null hypothesis.		
Conclusion:	The new type of synthetic passenger car tire does have an expected lifetime of 40,000 miles.		

Fig. A.8 Answer to Chap. 3: Practice Problem #2

Chapter 3: Practice Problem #3 Answer (see Fig. A.9)

FOCUS GROUP PRICING STUDY				
Question #10: "How much would you be willing to pay for this blouse?"				
Groups 1, 2, 3 in S				
62		Null hypothesis:	μ	= \$68
55				
73				
53		Research hypothesis:	μ	≠ \$68
46				
48				
57		n	30	
59				
65				
68		Mean	\$ 63.23	
64				
72				
62		STDEV	\$ 6.75	
67				
59				
71		s.e.	\$ 1.23	
65				
63				
69		95% confidence interval		
71				
70		lower limit	\$ 60.71	
58				
67		upper limit	\$ 65.75	
65				
63				
59	---- \$60.71	-----\$63.23	----- \$65.75	----- \$68 ----
70	lower	Mean	upper	Ref.
67	limit		limit	Value
64				
65				
Result:	Since the reference value is outside of the confidence interval, we reject the null hypothesis and accept the research hypothesis			
Conclusion:	Adult women (ages 25-44) were willing to pay a price significantly less than \$68 , and it was probably closer to \$63			

Fig. A.9 Answer to Chap. 3: Practice Problem #3

Chapter 4: Practice Problem #2 Answer (see Fig. A.11)

First-semester Calculus Standardized math test					
	Score				
87		Null hypothesis:	$\mu =$	88	
90					
85					
94		Research hypothesis:	$\mu \neq$	88	
93					
88					
82	n	25			
85					
96					
89	Mean	89.96			
92					
94					
91	STDEV	3.55			
89					
90					
92	s.e.	0.71			
93					
89					
88	critical t	2.064			
87					
86					
92	t-test	2.76			
94					
95					
88	Result:	Since the absolute value of 2.76 is greater than the critical t of 2.064, we reject the null hypothesis and accept the research hypothesis			
	Conclusion:	This year's beginning calculus students scored significantly higher on the standardized math test than last year's beginning calculus students.			

Fig. A.11 Answer to Chap. 4: Practice Problem #2

Chapter 4: Practice Problem #3 Answer (see Fig. A.12)

BOSTON UNIVERSITY M.S. IN ADVERTISING PROGRAM							
Course: Advertising Management							
Item #12:	"How would you rate the instructor's ability to explain advertising concepts clearly?"						
	1	2	3	4	5	6	7
	Poor					6.05	Excellent
						Mean	
	RATING						
	5		Null hypothesis:			μ	= 4
	6		Research hypothesis:			μ	\neq 4
	4						
	7						
	6						
	5		n		19		
	7						
	6						
	7		Mean		6.05		
	5						
	6						
	7		STDEV		0.91		
	6						
	7						
	5		s.e.		0.21		
	6						
	7						
	6		critical t		2.101		
	7						
			t-test		9.82		
Result:	Since the absolute value of 9.82 is greater than the critical t of 2.101, we reject the null hypothesis and accept the research hypothesis.						
Conclusion:	Students who took Advertising Management this past semester rated the ability of the instructor to explain advertising concepts clearly as significantly positive.						

Fig. A.12 Answer to Chap. 4: Practice Problem #3

Chapter 5: Practice Problem #1 Answer (see Fig. A.13)

UNIVERSITY OF ILLINOIS -- URBANA			
GPA OF MS IN ADVERTISING STUDENTS WHO HAVE COMPLETED ALL ADVERTISING REQUIRED COURSES			
Group	n	Mean	STDEV
1 Males	17	3.15	0.42
2 Females	15	3.45	0.37
Null hypothesis:		$\mu_1 = \mu_2$	
Research hypothesis:		$\mu_1 \neq \mu_2$	
$(n1 - 1) \times \text{STDEV1 squared}$		2.82	
$(n2 - 1) \times \text{STDEV2 squared}$		1.92	
$n1 + n2 - 2$		30	
$1/n1 + 1/n2$		0.13	
s.e.		0.14	
critical t		2.042	
t-test		-2.13	
Result: Since the absolute value of - 2.13 is greater than the critical t of 2.042, we reject the null hypothesis and accept the research hypothesis.			
Conclusion: Female MS in Advertising students who have completed all of the required advertising courses had significantly higher GPAs than male advertising students (3.45 vs. 3.15)			

Fig. A.13 Answer to Chap. 5: Practice Problem #1

Chapter 5: Practice Problem #2 Answer (see Fig. A.14)

Item:		"How interested are you in learning more about how life insurance can provide income for retirement?"																	
	1	2	3	4	5	6	7												
	Not at all interested		3.44 Women		5.16 Men		Very Interested												
Ad: Male model																			
	Men	Women	Null hypothesis: $\mu_1 = \mu_2$																
	5	3	Research hypothesis: $\mu_1 \neq \mu_2$																
	6	4																	
	4	6																	
	7	5																	
	5	2	<table border="1"> <thead> <tr> <th>Group</th> <th>n</th> <th>mean</th> <th>STDEV</th> </tr> </thead> <tbody> <tr> <td>1 Men</td> <td>32</td> <td>5.16</td> <td>1.51</td> </tr> <tr> <td>2 Women</td> <td>34</td> <td>3.44</td> <td>1.31</td> </tr> </tbody> </table>					Group	n	mean	STDEV	1 Men	32	5.16	1.51	2 Women	34	3.44	1.31
Group	n	mean	STDEV																
1 Men	32	5.16	1.51																
2 Women	34	3.44	1.31																
	6	3																	
	5	1																	
	4	3																	
	3	2	STDEV1 squared / n1			0.07													
	6	4	STDEV2 squared / n2			0.05													
	7	3																	
	5	5																	
	6	6																	
	4	3	s.e.			0.35													
	7	4																	
	5	2	critical t			1.96													
	4	5	(df = n1 + n2 - 2 = 64)																
	6	3																	
	3	4																	
	7	5																	
	5	4																	
	6	3	t-test			4.93													
	2	2																	
	6	4																	
	1	3	Result: Since the absolute value of 4.93 is greater than the critical t of 1.96, we reject the null hypothesis and accept the research hypothesis																
	7	5																	
	6	1																	
	5	3																	
	4	2	Conclusion: Adult men (ages 25-39) were significantly more interested than adult women (ages 25-39) in learning more about how life insurance can provide income for retirement when a male model was used in the ad (5.16 vs. 3.44)																
	6	3																	
	5	2																	
	7	5																	
		3																	
		4																	

Fig. A.14 Answer to Chap. 5: Practice Problem #2

Chapter 5: Practice Problem #3 Answer (see Fig. A.15)

FRONT-BUMPER CRASH RESISTANT TEST (2-door passenger cars)				
Honda Civic: 15 mph speed		Null hypothesis:		$\mu_1 = \mu_2$
REPAIR ESTIMATES (\$)		Research hypothesis:		$\mu_1 \neq \mu_2$
BRAND	n	Mean (\$)	STDEV (\$)	
1 BRAND X	11	1,206	83.45	
2 BRAND Y	11	1,333	45.89	
(n1 - 1) x STDEV1 SQUARED			69,639.03	
(n2 - 1) x STDEV2 SQUARED			21,058.92	
n1 + n2 - 2			20	
1/n1 + 1/n2			0.18	
s.e.		28.71		
critical t		2.086		
t-test		-4.42		
Result:	Since the absolute value of -4.42 is greater than the critical t of 2.086, we reject the null hypothesis and accept the research hypothesis.			
Conclusion:	For the Honda Civic (2-door) sedan, the front-bumper for BRAND X cost significantly less to repair than the front-bumper of BRAND Y (\$1,206 vs. \$1,333).			

Fig. A.15 Answer to Chap. 5: Practice Problem #3

Chapter 6: Practice Problem #1 Answer (see Fig. A.16)

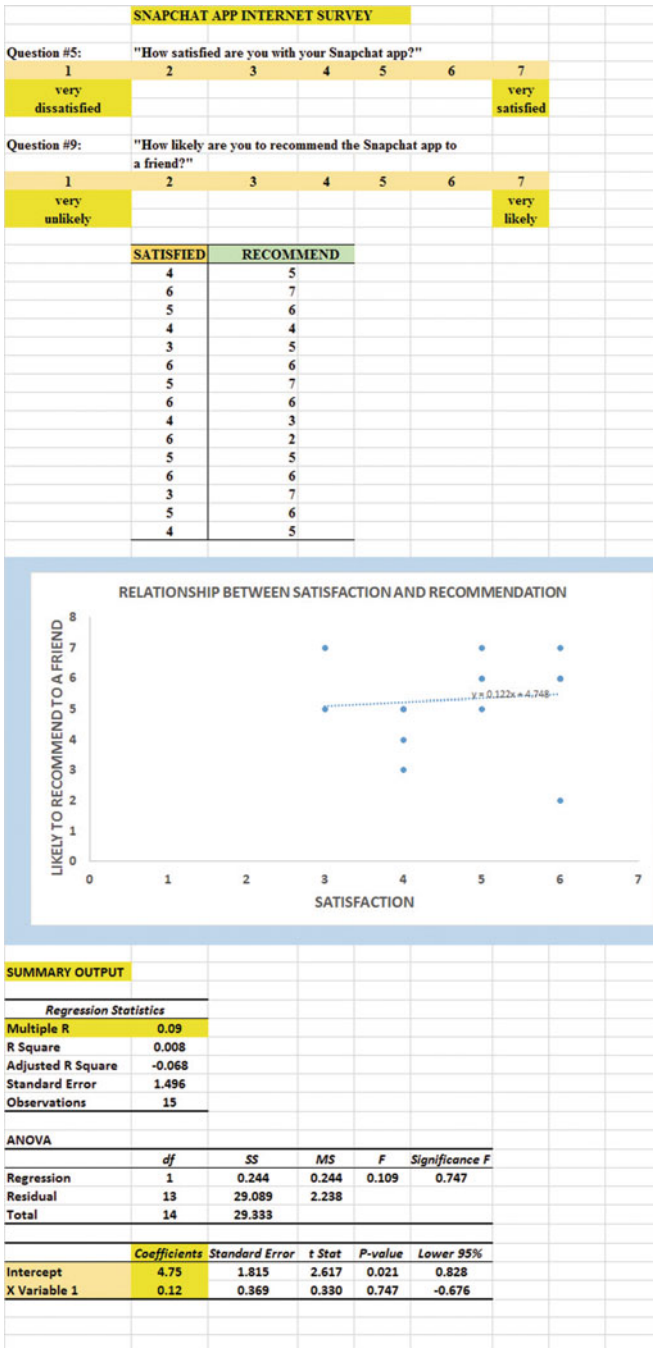


Fig. A.16 Answer to Chap. 6: Practice Problem #1

Chapter 6: Practice Problem #1 (continued)

1. $a = y\text{-intercept} = 4.75$, and $b = \text{slope} = 0.12$
2. $Y = a + b X$
 $Y = 4.75 + 0.12 X$
3. $r = +.09$
4. Be careful here! Because the correlation is so low ($r = .09$), you should not use it to predict Y for any X -value. It is a very poor relationship between these two variables!
5. Same comment as #4 above. Do not use the regression equation to predict Y !

Chapter 6: Practice Problem #2 Answer (see Fig. A.17)

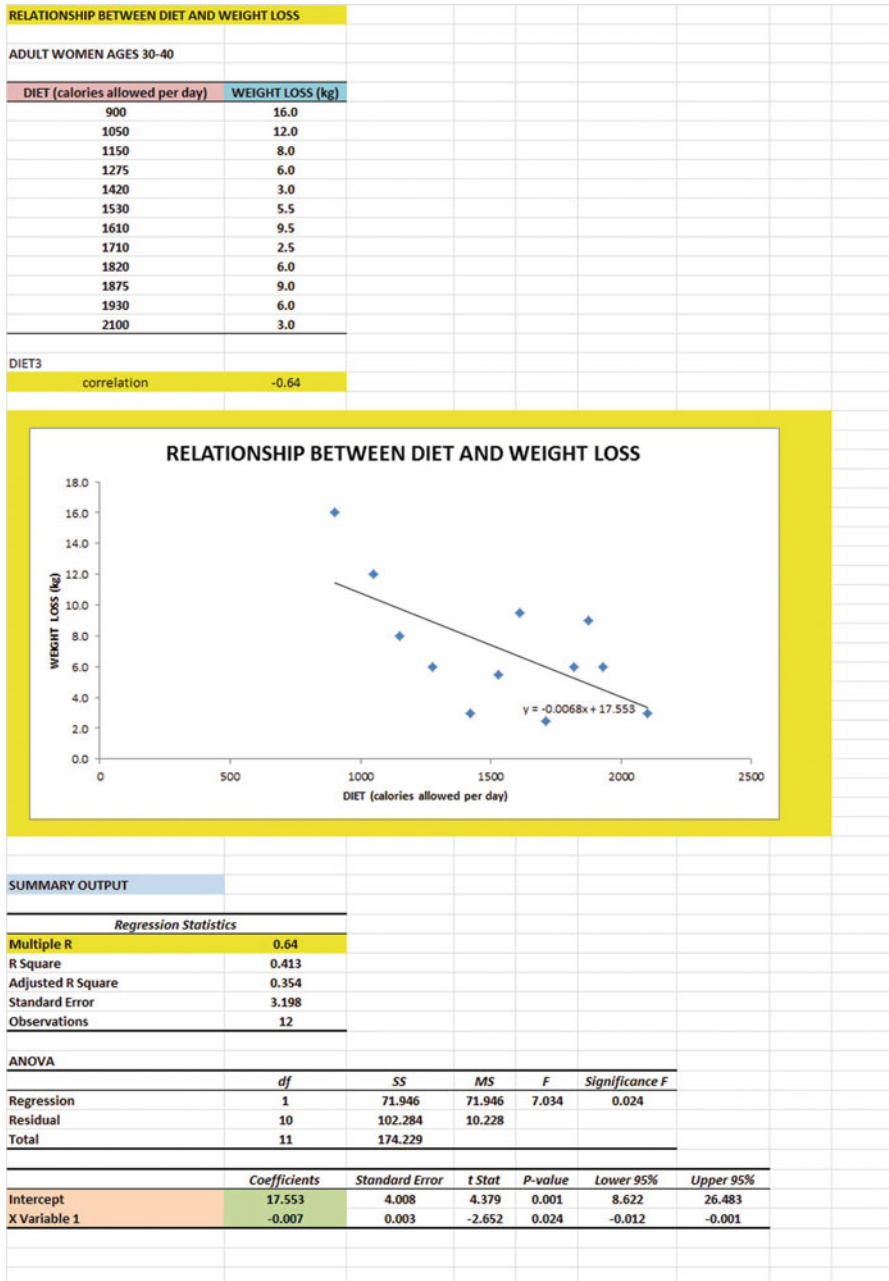


Fig. A.17 Answer to Chap. 6: Practice Problem #2

Chapter 6: Practice Problem #2 (continued)

1. $r = -0.64$ (note the negative correlation!)
2. $a = \text{y-intercept} = 17.553$
3. $b = \text{slope} = -0.007$ (note the minus sign as the slope is negative)
4. $Y = a + b X$
 $Y = 17.553 - 0.007 X$
5. $Y = 17.553 - 0.007 (1500)$
 $Y = 17.553 - 10.5$
 $Y = 7.1 \text{ kg}$

Chapter 6: Practice Problem #3 Answer (see Fig. A.18)

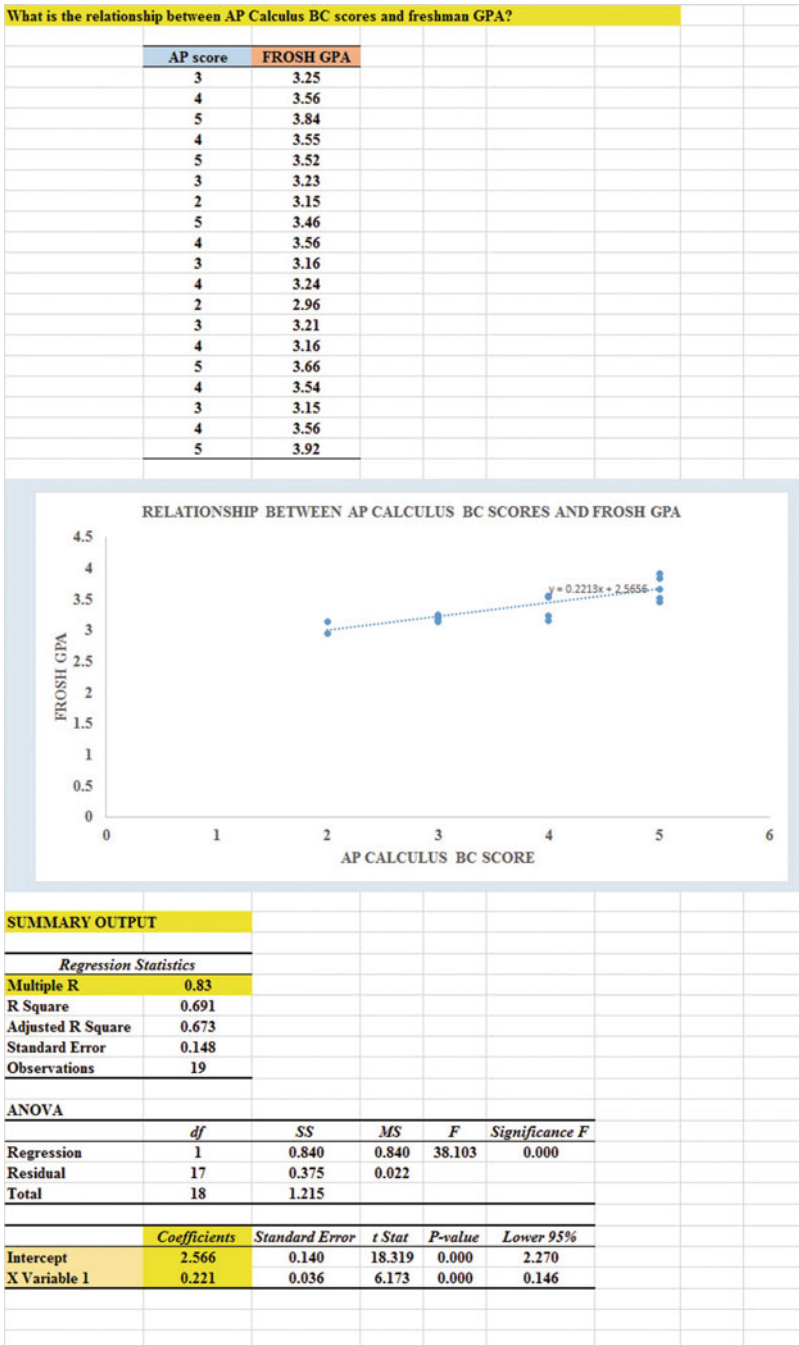


Fig. A.18 Answer to Chap. 6: Practice Problem #3

Chapter 6: Practice Problem #3 (continued)

1. $r = 0.83$
2. $a = y\text{-intercept} = 2.566$
3. $b = \text{slope} = 0.221$
4. $Y = a + b X$
 $Y = 2.566 + 0.221 X$
5. $Y = 2.566 + 0.221 (4)$
 $Y = 2.566 + 0.884$
 $Y = 3.45 \text{ GPA}$

Chapter 7: Practice Problem #1 Answer (see Fig. A.19)

GRADUATE RECORD EXAMINATIONS					
How well does the GRE and the GRE subject area test in Math predict GPA at the end of the first year of a Master's program in Engineering?					
FIRST-YEAR GPA	GRE VERBAL	GRE QUANTITATIVE	GRE WRITING	GRE MATH	
3.25	160	161	5	650	
3.42	156	158	4	600	
2.85	156	157	2	500	
2.65	154	153	1	510	
3.65	166	166	6	630	
3.16	159	160	3	550	
3.56	166	163	4	610	
2.35	155	154	2	430	
2.86	153	154	3	450	
2.95	158	157	4	550	
3.15	158	159	4	580	
3.45	160	160	5	620	
SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.93				
R Square	0.868				
Adjusted R Square	0.792				
Standard Error	0.178				
Observations	12				
ANOVA					
	df	SS	MS	F	Significance F
Regression	4	1.453	0.363	11.491	0.003
Residual	7	0.221	0.032		
Total	11	1.674			
	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	-5.2698	4.173	-1.263	0.247	-15.138
GRE VERBAL	-0.0174	0.041	-0.424	0.684	-0.115
GRE QUANTITATIVE	0.0614	0.057	1.070	0.320	-0.074
GRE WRITING	0.0332	0.075	0.444	0.670	-0.144
GRE MATH	0.0023	0.001	1.592	0.155	-0.001
	FIRST-YEAR GPA	GRE VERBAL	GRE QUANTITATIVE	GRE WRITING	GRE MATH
FIRST-YEAR GPA	1				
GRE VERBAL	0.79	1			
GRE QUANTITATIVE	0.88	0.94	1		
GRE WRITING	0.83	0.72	0.83	1	
GRE MATH	0.89	0.74	0.83	0.81	1

Fig. A.19 Answer to Chap. 7: Practice Problem #1

Chapter 7: Practice Problem #1 (continued)

1. Multiple correlation = .93
2. y-intercept = -5.2698
3. -0.0174
4. 0.0614
5. 0.0332
6. 0.0023
7. $Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$
 $Y = -5.2698 - 0.0174 X_1 + 0.0614 X_2 + 0.0332 X_3 + 0.0023 X_4$
8. $Y = -5.2698 - 0.0174 (150) + 0.0614 (160) + 0.0332 (3) + 0.0023 (610)$
 $Y = -5.2698 - 2.61 + 9.82 + 0.10 + 1.40$
 $Y = 11.32 - 7.88$
 $Y = 3.44$
9. 0.79
10. 0.88
11. 0.83
12. 0.89
13. 0.72
14. 0.74
15. The GRE MATH exam is the best single predictor of First-Year GPA with a correlation $r = +0.89$.
16. The four predictors combined predict the First-Year GPA at $R_{xy} = .93$, and this is much better than the best single predictor's correlation of $r = +.89$.

Chapter 7: Practice Problem #2 Answer (see Fig. A.20)

	Sales (\$ 000)	Radio Ads (\$ 000)	Newspaper Ads (\$ 000)	TV Ads (\$ 000)
	943	22	25	40
	1212	24	26	42
	937	26	27	44
	651	28	29	48
	890	30	28	49
	923	32	24	52
	1243	34	29	51
	1652	36	32	53
	1044	38	31	59

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.86
R Square	0.738
Adjusted R Square	0.580
Standard Error	184.765
Observations	9

ANOVA					
	df	SS	MS	F	Significance F
Regression	3	480166.162	160055.387	4.688	0.065
Residual	5	170689.838	34137.968		
Total	8	650856.000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	2962.68	1244.293	2.381	0.063	-235.875
Radio Ads (\$ 000)	184.02	58.507	3.145	0.026	33.621
Newspaper Ads (\$ 000)	2.82	35.954	0.078	0.941	-89.605
TV Ads (\$ 000)	-154.25	49.580	-3.111	0.027	-281.699

	Sales (\$ 000)	Radio Ads (\$ 000)	Newspaper Ads (\$ 000)	TV Ads (\$ 000)
Sales (\$ 000)	1			
Radio Ads (\$ 000)	0.42	1		
Newspaper Ads (\$ 000)	0.46	0.70	1	
TV Ads (\$ 000)	0.23	0.97	0.63	1

Fig. A.20 Answer to Chap. 7: Practice Problem #2

Chapter 7: Practice Problem #2 (continued)

(e) (1a) $R_{xy} = .86$

(e) (2b) $a = y\text{-intercept} = 2962.68$

$b_1 = \text{Radio ads} = 184.02$

$b_2 = \text{Newspaper ads} = 2.82$

$b_3 = \text{TV ads} = -154.25$

(f) Let Radio ads = X_1 , Newspaper ads = X_2 , and TV ads = X_3

$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3$

(g) $Y = 2962.68 + 184.02 X_1 + 2.82 X_2 - 154.25 X_3$

$Y = 2962.68 + 184.02(28) + 2.82(24) - 154.25(49)$

$Y = 2962.68 + 5152.56 + 67.68 - 7558.25$

$Y = 8182.92 - 7558.25$

$Y = 624.67 (\$000)$

$Y = \$624,670$

- (k) (1a) $r = .42$
- (k) (2b) $r = .46$
- (k) (3c) $r = .23$
- (k) (4d) $r = .97$
- (k) (5e) $r = .70$
- (k) (6f) The best single predictor of Sales was Newspaper Ads ($r = .46$)
- (k) The three predictor variables combined predicted Sales much better ($R_{xy} = .86$)

Chapter 7: Practice Problem #3 Answer (see Fig. A.21)

SAT EXAM					
FROSH GPA	SAT READING	HS GPA	AP CALCULUS BC		
3.23	650	3.55	2		
2.90	490	2.96	2		
2.80	630	3.27	5		
3.42	520	3.45	4		
2.80	560	2.90	3		
2.90	410	2.80	2		
2.35	450	2.63	2		
2.58	420	2.71	1		
3.12	560	3.26	3		
3.47	650	3.45	4		
SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.91				
R Square	0.837				
Adjusted R Square	0.755				
Standard Error	0.176				
Observations	10				
ANOVA					
	df	SS	MS	F	Significance F
Regression	3	0.958	0.319	10.258	0.009
Residual	6	0.187	0.031		
Total	9	1.145			
Coefficients					
	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	-0.209	0.607	-0.344	0.742	-1.695
SAT READING	-0.002	0.001	-1.445	0.199	-0.005
HS GPA	1.327	0.318	4.175	0.006	0.549
AP CALCULUS BC	0.001	0.064	0.012	0.991	-0.156
SAT11					
	FROSH GPA	SAT READING	HS GPA	AP CALCULUS BC	
FROSH GPA	1				
SAT READING	0.59	1			
HS GPA	0.88	0.83	1		
AP CALCULUS BC	0.47	0.65	0.61	1	

Fig. A.21 Answer to Chap. 7: Practice Problem #3

Chapter 7: Practice Problem #3 (continued)

1. Multiple correlation = +.91

2. y-intercept = -0.209

Let X_1 = SAT READING, X_2 = HS GPA, X_3 = AP CALCULUS BC

3. $b_1 = -0.002$

4. $b_2 = 1.327$

5. $b_3 = 0.001$

6. $Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3$

$Y = -0.209 - 0.002 X_1 + 1.327 X_2 + 0.001 X_3$

7. $Y = -0.209 - 0.002 (650) + 1.327 (3.47) + 0.001 (4)$

$Y = -0.209 - 1.3 + 4.6 + 0.004$

$Y = 4.604 - 1.509$

$Y = 3.10$ GPA

8. .59

9. .88

10. .47

11. .83

12. .65

13. .61

14. The best predictor of FROSH GPA was HS GPA ($r = .88$)

15. The three predictors combined predict FROSH GPA much better ($R_{xy} = .91$)

Chapter 8: Practice Problem #1 Answer (see Fig. A.22)

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY		
UNDERGRADUATE ELECTRICAL ENGINEERING 101 COURSE: FINAL EXAM		
CAI	LECTURES	INDEPENDENT
90	85	76
85	89	80
74	83	90
89	79	84
84	74	78
95	75	65
92	86	42
65	87	58
75	86	63
73	88	75
54		66
71		

Anova: Single Factor				
SUMMARY				
Groups	Count	Sum	Average	Variance
CAI	12	947	78.92	151.72
LECTURES	10	832	83.20	28.84
INDEPENDENT	11	777	70.64	183.45

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	867.12	2	433.56	3.46	0.04	3.32
Within Groups	3763.06	30	125.44			
Total	4630.18	32				

LECTURES vs. INDEPENDENT STUDY	
1/n LECTURES + 1/n INDEPENDENT	0.19
s.e. of LECTURES vs. INDEPENDENT	4.89
ANOVA t-test	2.57

Fig. A.22 Answer to Chap. 8: Practice Problem #1

Chapter 8: Practice Problem #1 (continued)

Let CAI = X_1 , LECTURES = X_2 , and INDEPENDENT = X_3

1. $H_0: \mu_1 = \mu_2 = \mu_3$
 $H_1: \mu_1 \neq \mu_2 \neq \mu_3$
2. $MS_b = 433.56$
3. $MS_w = 125.44$
4. $F = 433.56/125.44 = 3.46$
5. 3.32
6. Result: Since 3.46 is greater than 3.32, we reject the null hypothesis and accept the research hypothesis.
7. There was a significant difference between the three teaching formats in final exam scores.

LECTURES vs. INDEPENDENT STUDY

8. $H_0: \mu_2 = \mu_3$
 $H_1: \mu_2 \neq \mu_3$
9. 83.20
10. 70.64
11. $df = 33 - 3 = 30$
12. critical t = 2.042
13. $s.e. = \text{SQRT}(125.44 * [1/10 + 1/11]) = \text{SQRT}(125.44 * 0.19) = \text{SQRT}(23.83) = 4.88$
14. ANOVA t = $(83.20 - 70.64)/4.88 = 12.56/4.88 = 2.57$
15. Result: Since the absolute value of 2.57 is greater than 2.042, we reject the null hypothesis and accept the research hypothesis.
16. Conclusion: Final Exam scores were significantly higher in LECTURES than they were in INDEPENDENT STUDY (83 vs. 71).

Chapter 8: Practice Problem #2 Answer (see Fig. A.23)

GOLF BALL DURABILITY TEST RESULTS						
BRAND X	BRAND Y	BRAND Z				
257	310	242				
289	315	265				
292	324	275				
315	336	288				
328	324	262				
348	318	274				
327	321	263				
	328	254				
		277				
Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
BRAND X	7	2156	308.0	938		
BRAND Y	8	2576	322.0	64.2857		
BRAND Z	9	2400	266.7	186.5		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	14127.33333	2	7063.67	19.60	1.57848E-05	3.47
Within Groups	7570	21	360.476			
Total	21697.33333	23				
BRAND Y vs. BRAND Z						
1/n for Y + 1/n for Z	0.24					
s.e.	9.23					
ANOVA t-test	6.00					

Fig. A.23 Answer to Chap. 8: Practice Problem #2

Chapter 8: Practice Problem #2 (continued)

Let BRAND X = X_1 , BRAND Y = X_2 , and BRAND Z = X_3

- (b) Null hypothesis: $\mu_1 = \mu_2 = \mu_3$
 Research hypothesis: $\mu_1 \neq \mu_2 \neq \mu_3$
- (f) $MS_b = 7063.667$ and $MS_w = 360.4762$
- (g) $F = 19.60$ and critical $F = 3.47$
- (h) Mean BRAND Y = 322 and Mean BRAND Z = 266.7
- (j) Result: Since the F-value of 19.60 is greater than the critical F value of 3.47, we reject the null hypothesis and accept the research hypothesis.
- (k) Conclusion: There was a significant difference between the three brands of golf balls in the number of drives needed to crack or chip the golf balls.

BRAND Y vs. BRAND Z

- (l) $H_0: \mu_2 = \mu_3$
 $H_1: \mu_2 \neq \mu_3$
- (m) ANOVA $t = (322 - 266.7)/9.23 = 55.3/9.23 = 5.99$
- (n) $df = n_{TOTAL} - k = 24 - 3 = 21$
- (o) Critical $t = 2.080$
- (p) Result: Since the absolute value of 5.99 is greater than 2.080, we reject the null hypothesis and accept the research hypothesis.
- (q) Conclusion: BRAND Y had significantly more drives to crack or chip the golf balls than BRAND Z (322 vs. 267).

Chapter 8: Practice Problem #3 Answer (see Fig. A.24)

HIGHWAY MILES PER GALLON (mpg) COMPARISON OF FIVE TYPES OF CARS						
	1	2	3	4	5	
	SUBCOMPACTS (mpg)	COMPACTS (mpg)	MID-SIZE (mpg)	LARGE (mpg)	SUVs (mpg)	
	28.1	26.2	24.0	22.0	18.1	
	30.2	28.3	26.3	23.1	20.2	
	29.3	29.3	25.2	25.4	22.3	
	31.6	27.0	27.1	24.3	21.4	
	33.0	28.0	28.0	25.0	20.5	
	34.3	29.5	23.6	24.7	19.0	
	32.1	31.0	29.2	23.1	18.2	
	35.0	32.3		22.4	19.1	
		33.1		26.0		
				21.3		
Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
SUBCOMPACTS (mpg)	8	253.60	31.70	5.78		
COMPACTS (mpg)	9	264.70	29.41	5.48		
MID-SIZE (mpg)	7	183.40	26.20	4.28		
LARGE (mpg)	10	237.30	23.73	2.48		
SUVs (mpg)	8	158.80	19.85	2.29		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	718.23	4	179.56	44.80	0.00	2.63
Within Groups	148.29	37	4.01			
Total	866.52	41				
COMPACTS vs. LARGE						
1/9 + 1/10	0.21					
s.e. ANOVA	0.92					
ANOVA t-test	6.18					

Fig. A.24 Answer to Chap. 8: Practice Problem #3

Chapter 8: Practice Problem #3 (continued)

1. Null hypothesis: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
 Research hypothesis: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$
2. $MS_b = 179.56$
3. $MS_w = 4.01$
4. $F = 179.56/4.01 = 44.78$
5. critical $F = 2.63$
6. Result: Since the F-value of 44.78 is greater than the critical F value of 2.63, we reject the null hypothesis and accept the research hypothesis.
7. Conclusion: There was a significant difference between the five types of cars in highway miles per gallon.

COMPACTS vs. LARGE

8. Null hypothesis: $\mu_2 = \mu_4$
 Research hypothesis: $\mu_2 \neq \mu_4$
9. Mean COMPACTS = 29.41
10. Mean LARGE = 23.73
11. degrees of freedom = $42 - 5 = 37$
12. critical $t = 2.026$
13. $s.e._{ANOVA} = \text{SQRT}(4.01 \times 0.21) = \text{SQRT}(0.84) = 0.92$
14. ANOVA $t = (29.41 - 23.73)/0.92 = 5.68/0.92 = 6.17$
15. Result: Since the absolute value of 6.17 is greater than the critical t of 2.026, we reject the null hypothesis and accept the research hypothesis.
16. Conclusion: COMPACTS had significantly more highway miles per gallon than LARGE (29 vs. 24).

Appendix B: Practice Test

Chapter 1: Practice Test

Suppose that you have been asked by the manager of the Webster Groves Subaru dealer in St. Louis to analyze the data from a recent survey of its customers. Subaru of America mails a “SERVICE EXPERIENCE SURVEY” to customers who have recently used the Service Department for their car. Let’s try your Excel skills on Item #10e of this survey (see Fig. B.1).

Question #10e: "Your overall rating of the quality of work performed on your vehicle."									
1	2	3	4	5	6	7	8	9	10
Unacceptable									Extraordinary
	Data								
	8								
	5								
	6								
	5								
	4								
	8								
	7								
	7								
	8								
	6								
	7								
	5								
	4								
	8								
	7								
	5								
	7								
	5								
	7								
	6								

Fig. B.1 Worksheet Data for Chap. 1 Practice Test (Practical Example)

- (a) Create an Excel table for these data, and then use Excel to the right of the table to find the sample size, mean, standard deviation, and standard error of the mean for these data. Label your answers, and round off the mean, standard deviation, and standard error of the mean to two decimal places.
- (b) Save the file as: SUBARU8

Chapter 2: Practice Test

Suppose that you wanted to do a personal interview with a random sample of 12 of your company's 42 salespeople as part of a company morale survey.

- (a) Set up a spreadsheet of frame numbers for these salespeople with the heading: FRAME NUMBERS
- (b) Then, create a separate column to the right of these frame numbers which duplicates these frame numbers with the title: Duplicate frame numbers.
- (c) Then, create a separate column to the right of these duplicate frame numbers called RAND NO. and use the =RAND() function to assign random numbers to all of the frame numbers in the duplicate frame numbers column, and change this column format so that three decimal places appear for each random number.
- (d) Sort the *duplicate frame numbers and random numbers* into a random order.
- (e) Print the result so that the spreadsheet fits onto one page.
- (f) Circle on your printout the I.D. number of the first 12 salespeople that you would interview.
- (g) Save the file as: RAND15

Important note: Note that everyone who does this problem will generate a different random order of salespeople ID numbers since Excel assign a different random number each time the RAND() command is used. For this reason, the answer to this problem given in this Excel Guide will have a completely different sequence of random numbers from the random sequence that you generate. This is normal and what is to be expected.

Chapter 3: Practice Test

Suppose that you have been asked to analyze the data from a flight on Southwest Airlines from St. Louis to Boston. Southwest sent an online customer satisfaction survey to a random sample of its frequent fliers the day after the flight and asked them to rate their flight on 10-point scales with 1 = extremely dissatisfied, and 10 = extremely satisfied. The hypothetical data for Item #2c appear in Fig. B.2.

SOUTHWEST AIRLINES ONLINE SURVEY										
Item #2c:	Please tell us your overall satisfaction with your gate area experience at the airport (gate agent service, facilities, boarding process, and departure time)									
	1	2	3	4	5	6	7	8	9	10
	extremely dissatisfied									extremely satisfied
				STL-BOS						
				6						
				3						
				8						
				5						
				9						
				10						
				4						
				7						
				6						
				9						
				8						
				7						
				9						
				10						
				7						
				6						
				8						

Fig. B.2 Worksheet Data for Chap. 3 Practice Test (Practical Example)

- (a) Create an Excel table for these data, and use Excel to the right of the table to find the sample size, mean, standard deviation, and standard error of the mean for these data. Label your answers, and round off the mean, standard deviation, and standard error of the mean to two decimal places in number format.
- (b) By hand, write the null hypothesis and the research hypothesis on your printout.
- (c) Use Excel’s *TINV* function to find the 95% confidence interval about the mean for these data. Label your answers. Use two decimal places for the confidence interval figures in number format.
- (d) On your printout, draw a diagram of this 95% confidence interval by hand, including the reference value.
- (e) On your spreadsheet, enter the *result*.
- (f) On your spreadsheet, enter the *conclusion in plain English*.
- (g) Print the data and the results so that your spreadsheet fits onto one page.
- (h) Save the file as: south3

Chapter 4: Practice Test

Suppose that you have been asked by the American Marketing Association to analyze the data from the Summer Educators' conference in Boston. In order to check your Excel formulas, you have decided to analyze the data for one of these questions before you analyze the data for the entire survey, one item at a time. The conference used five-point scales with 1 = Definitely Would Not, and 5 = Definitely Would. A random sample of the hypothetical data for Item #3 is given in Fig. B.3.

American Marketing Association				
Summer Educators' Conference in Boston, MA				
Item #3: "How likely are you to recommend the Conference to a friend or colleague?"				
1	2	3	4	5
Definitely Would Not				Definitely Would
Rating				
4				
5				
3				
4				
2				
5				
4				
5				
3				
5				
4				
5				
3				
2				
1				
4				
5				
4				
5				
3				
5				
5				

Fig. B.3 Worksheet Data for Chap. 4 Practice Test (Practical Example)

- (a) Write the null hypothesis and the research hypothesis on your spreadsheet.
- (b) Create a spreadsheet for these data, and then use Excel to find the sample size, mean, standard deviation, and standard error of the mean to the right of the data set. Use number format (three decimal places) for the mean, standard deviation, and standard error of the mean.
- (c) Type the *critical t* from the t-table in Appendix E onto your spreadsheet, and label it.
- (d) Use Excel to compute the t-test value for these data (use three decimal places) and label it on your spreadsheet.
- (e) Type the *result* on your spreadsheet, and then type the *conclusion in plain English* on your spreadsheet.
- (f) Save the file as: BOS2

Chapter 5: Practice Test

Massachusetts Mutual Financial Group (2010) placed a full-page color ad in *The Wall Street Journal* in which it used a male model hugging a two-year old daughter. The ad had the headline and sub-headline:

WHAT IS THE SIGN OF A GOOD DECISION?

It's knowing your life insurance can help provide income for retirement. And peace of mine until you get there.

Since the majority of the subscribers to *The Wall Street Journal* are men, an interesting research question would be the following:

Research question: "Does the gender of the model affect adult men's willingness to learn more about how life insurance can provide income for retirement?"

Suppose that you have shown two groups of adult males (ages 25–44) a mockup of an ad such one group of males saw the ad with a male model, while another group of males saw the identical ad except that it had a female model in the ad. (You randomly assigned these males to one of the two experimental groups.) The two groups were kept separate during the experiment and could not interact with one another.

At the end of a one-hour discussion of the mockup ad, the respondents were asked the question given in Fig. B.4.

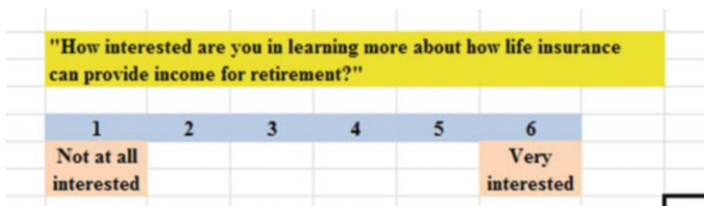


Fig. B.4 Survey Item for a Mockup Ad (Practical Example)

The resulting data for this one item appear in Fig. B.5.

Fig. B.5 Worksheet Data for Chap. 5 Practice Test (Practical Example)

MASS MUTUAL FINANCIAL GROUP					
Item: "How interested are you in learning more about how life insurance can provide income for retirement?"					
1	2	3	4	5	6
Not at all interested					Very interested
			Male model	Female model	
			3	4	
			2	6	
			4	5	
			5	3	
			1	4	
			6	6	
			2	6	
			4	5	
			3	3	
			5	5	
			2	4	
			4	3	
			3	5	
			5	4	
			1	6	
			2	5	
			3	5	
			1	6	
			4	4	
			5	6	
			6	3	
			2	4	
			3	6	
			1	5	
			4	6	
			3	4	
			5	4	

- (a) Write the null hypothesis and the research hypothesis.
- (b) Create an Excel table that summarizes these data.
- (c) Use Excel to find the standard error of the difference of the means.
- (d) Use Excel to perform a *two-group t-test*. What is the value of *t* that you obtain (use two decimal places)?
- (e) On your spreadsheet, type the *critical value of t* the t-table in Appendix E.
- (f) Type the *result* of the test on your spreadsheet.
- (g) Type your *conclusion in plain English* on your spreadsheet.
- (h) Save the file as: lifeinsur3
- (i) Print the final spreadsheet so that it fits onto one page.

Chapter 6: Practice Test

Is there a relationship between the amount of money spent on TV ads for a local supermarket chain in St. Louis, Missouri USA and the weekly sales dollars for this chain? You have decided to use the cost of the TV ads as the independent variable (predictor) and weekly sales for the supermarket chain as the dependent variable (criterion). Use simple linear regression for the hypothetical data given in Fig. B.6.

Fig. B.6 Worksheet Data for Chap. 6 Practice Test (Practical Example)

TV ads (\$)	Weekly Sales (\$)
1,600	24,000
2,400	26,000
2,000	20,000
2,600	23,000
3,400	35,000
2,100	33,000
2,600	24,000
2,700	22,000
2,000	10,000
2,200	31,000
2,600	25,000

Create an Excel spreadsheet, and enter the data.

- (a) create an *XY scatterplot* of these two sets of data such that:
 - top title: RELATIONSHIP BETWEEN TV AD COST AND WEEKLY SALES
 - x-axis title: TV ads (\$)
 - y-axis title: Weekly sales (\$)
 - move the chart below the table

- re-size the chart so that it is 7 columns wide and 25 rows long
- delete the legend
- delete the gridlines

- (b) Create the *least-squares regression line* for these data on the scatterplot.
- (c) Use Excel to run the regression statistics to find the *equation for the least-squares regression line* for these data and display the results below the chart on your spreadsheet. Use number format (two decimal places) for the correlation and for the coefficients

Print *just the input data and the chart* so that this information fits onto one page in portrait format.

Then, print *just the regression output table* on a separate page so that it fits onto that separate page in portrait format.

By hand:

- (d) Circle and label the value of the *y-intercept* and the *slope* of the regression line on your printout.
- (e) Write the regression equation *by hand* on your printout for these data (use two decimal places for the *y-intercept* and the *slope*).
- (f) Circle and label the *correlation* between the two sets of scores in the regression analysis summary output table on your printout.
- (g) Underneath the regression equation you wrote by hand on your printout, use the regression equation to predict the weekly sales you would expect for a TV ad cost of \$2500 for that week.
- (h) *Read from the graph*, the weekly sales you would predict for a TV ad cost of \$3000 and write your answer in the space immediately below:

-
- (i) save the file as: TV33

Chapter 7: Practice Test

Suppose that you wanted to estimate the total number of gallons required for 4-door sedans when they were driven on a specific route of 200 miles between St. Louis, Missouri, and Indianapolis, Indiana, at specified speeds using drivers that were about the same weight. You have decided to use two predictors: (1) weight of the car (measured in thousands of pounds), and (2) the car's engine horsepower. To check your skills in Excel, you have created the hypothetical data given in the hypothetical table in Fig. B.7.

TOTAL GALLONS USED TO DRIVE FROM ST. LOUIS TO INDIANAPOLIS		
FOUR-DOOR SEDANS		
TOTAL GALLONS USED	WEIGHT (1000 lbs)	HORSEPOWER
6.1	3.8	130
6.3	3.7	150
4.8	4.0	140
4.2	2.4	125
3.8	2.9	98
4.7	3.0	115
3.5	2.1	121
5.5	2.9	123
5.9	3.1	110
3.4	2.1	96

Fig. B.7 Worksheet Data for Chap. 7 Practice Test (Practical Example)

- (a) create an Excel spreadsheet using TOTAL GALLONS USED as the criterion (Y), and the other variables as the two predictors of this criterion ($X_1 = \text{WEIGHT (1000 lbs)}$, and $X_2 = \text{HORSEPOWER}$).
- (b) Use Excel's *multiple regression* function to find the relationship between these three variables and place the SUMMARY OUTPUT below the table.
- (c) Use number format (two decimal places) for the multiple correlation on the Summary Output, and use two decimal places for the coefficients in the SUMMARY OUTPUT.
- (d) Save the file as: GALLONS9
- (e) Print the table and regression results below the table so that they fit onto one page.

Answer the following questions using your Excel printout:

1. What is the multiple correlation R_{xy} ?
 2. What is the y-intercept a ?
 3. What is the coefficient for WEIGHT b_1 ?
 4. What is the coefficient for HORSEPOWER b_2 ?
 5. What is the multiple regression equation?
 6. Predict the TOTAL GALLONS USED you would expect for a WEIGHT of 3800 pounds and a car that had 126 HORSEPOWER.
- (f) Now, go back to your Excel file and create a correlation matrix for these three variables, and place it underneath the SUMMARY OUTPUT.
 - (g) Re-save this file as: GALLONS9
 - (h) Now, print out *just this correlation matrix* on a separate sheet of paper.

Answer to the following questions using your Excel printout. (Be sure to include the plus or minus sign for each correlation):

7. What is the correlation between WEIGHT and TOTAL GALLONS USED?
8. What is the correlation between HORSEPOWER and TOTAL GALLONS USED?
9. What is the correlation between WEIGHT and HORSEPOWER?
10. Discuss which of the two predictors is the better predictor of total gallons used.
11. Explain in words how much better the two predictor variables combined predict total gallons used than the better single predictor by itself.

Chapter 8: Practice Test

Suppose that you have been asked to determine the best selling price for one of the products that your company sells in 7Eleven stores, and that you are undecided about whether to charge \$10, \$12, or \$14 per unit. Suppose that you have completed a test market by randomly assigning the selling price to stores owned by your company, and that you have recorded the number of units sold for each price point over a two-month period. Your task is to determine if there was a significant difference in the number of units sold of this product at these three prices during test marketing using the hypothetical data in Fig. B.8.

HOW DOES PRICE AFFECT UNITS SOLD IN A TEST MARKET STUDY			
	\$10 Price	\$12 Price	\$14 Price
	185	191	118
	161	180	110
	195	176	166
	132	132	165
	114	201	157
	151	143	143
	120	110	114
	182	184	99
	171	138	142
	170	170	170
	148	161	108
	149	152	98
	162	179	152
	135	162	154
	181	138	133
	201	214	180
	191	211	118
	151		128
	136		105
	129		

Fig. B.8 Worksheet Data for Chap. 8 Practice Test (Practical Example)

- (a) Enter these data on an Excel spreadsheet.
- (b) On your spreadsheet, write the null hypothesis and the research hypothesis for these data
- (c) Perform a *one-way ANOVA test* on these data, and show the resulting ANOVA table underneath the input data for the three prices.
- (d) If the F-value in the ANOVA table is significant, create an Excel formula to compute the ANOVA t-test comparing the number of units sold when the \$12 price was emphasized versus the number of units sold when the \$14 price was emphasized, and show the results below the ANOVA table on the spreadsheet (put the standard error and the ANOVA t-test value on separate lines of your spreadsheet, and use two decimal places for each value)
- (e) Print out the resulting spreadsheet so that all of the information fits onto one page.
- (f) On your printout, label by hand the MS (between groups) and the MS (within groups).
- (g) Circle and label the value for F on your printout for the ANOVA of the input data.
- (h) Label by hand on the printout the mean for the \$12 price and the mean for the \$14 price that were produced by your ANOVA formulas.
- (i) Save the spreadsheet as: Price9

On a separate sheet of paper, now do the following by hand:

- (j) What is the value of F that you obtained on your spreadsheet?
- (k) Find the critical value of F in the ANOVA Single Factor results table.
- (l) Write a summary of the *result* of the ANOVA test for the input data.
- (m) Write a summary of the *conclusion* of the ANOVA test in plain English for the input data.
- (n) Write the null hypothesis and the research hypothesis comparing the \$12 price versus the \$14 price.
- (o) Using your calculator, compute the standard error of the difference of the means.
- (p) Using your calculator, compute the ANOVA t-test.
- (q) Compute the degrees of freedom for the *ANOVA t-test* by hand for the two prices.
- (r) write the *critical value of t* for the ANOVA t-test using the table in Appendix E.
- (s) write a summary of the *result* of the ANOVA t-test.
- (t) write a summary of the *conclusion* of the ANOVA t-test in plain English.

Reference

Mass Mutual Financial Group. What is the Sign of a Good Decision? (Advertisement) *The Wall Street Journal*, September 29, 2010, p. A22.

Appendix C: Answers to Practice Test

Practice Test Answer: Chapter 1 (see Fig. C.1)

Question #10e:	"Your overall rating of the quality of work performed on your vehicle."	
	Data	
	8	
	5	n 20
	6	
	5	
	4	Mean 6.25
	8	
	7	
	7	STDEV 1.33
	8	
	6	
	7	s.e. 0.30
	5	
	4	
	8	
	7	
	5	
	7	
	5	
	7	
	6	

Fig. C.1 Practice Test Answer to Chap. 1 Problem

Practice Test Answer: Chapter 2 (see. Fig. C.2)

FRAME NUMBERS	Duplicate frame numbers	RAND NO.
1	8	0.335
2	22	0.732
3	31	0.802
4	42	0.163
5	4	0.933
6	29	0.053
7	3	0.379
8	21	0.965
9	37	0.187
10	17	0.523
11	34	0.608
12	25	0.650
13	10	0.705
14	41	0.995
15	30	0.136
16	36	0.501
17	13	0.884
18	15	0.857
19	20	0.709
20	14	0.306
21	9	0.969
22	12	0.282
23	38	0.370
24	26	0.610
25	1	0.215
26	5	0.520
27	35	0.007
28	28	0.368
29	24	0.759
30	32	0.427
31	27	0.373
32	19	0.258
33	6	0.265
34	39	0.018
35	2	0.959
36	18	0.932
37	7	0.311
38	11	0.422
39	16	0.802
40	40	0.710
41	33	0.202
42	23	0.021

Fig. C.2 Practice Test Answer to Chap. 2 Problem

Practice Test Answer: Chapter 3 (see. Fig. C.3)

SOUTHWEST AIRLINES ONLINE SURVEY										
Item #2c:	Please tell us your overall satisfaction with your gate area experience at the airport (gate agent service, facilities, boarding process, and departure time)									
1	2	3	4	5	6	7	8	9	10	
extremely dissatisfied									extremely satisfied	
			STL-BOS							
			6			Null hypothesis:			$\mu = 5.5$	
			3			Research hypothesis:			$\mu \neq 5.5$	
			8							
			5			n	17			
			9							
			10			Mean	7.18			
			4			STDEV	2.01			
			7							
			6			s.e.	0.49			
			9							
			7							
			9							
			10			95% confidence interval				
			7							
			6			lower limit		6.14		
			8			upper limit		8.21		
Draw a diagram of the confidence interval										
			5.5		6.14		7.18		8.21	
			Ref. Value		lower limit		Mean		upper limit	
Result:	Since the reference value of 5.5 is outside of the confidence interval, we reject the null hypothesis and accept the research hypothesis.									
Conclusion:	Frequent flier passengers on Southwest Airlines flight from St. Louis to Boston were significantly satisfied with their gate experience at the St. Louis airport.									

Fig. C.3 Practice Test Answer to Chap. 3 Problem

Practice Test Answer: Chapter 4 (see Fig. C.4)

Chapter 4: Practice Test answer (see Fig. C.4)	
American Marketing Association	
Summer Educators' Conference in Boston, MA	
Item #3:	"How likely are you to recommend the Conference to a friend or colleague?"
Rating	Null hypothesis: $\mu = 3$
4	
5	Research hypothesis: $\mu \neq 3$
3	
4	
2	n 22
5	
4	
5	Mean 3.909
3	
5	
4	STDEV 1.192
5	
3	
2	s.e. 0.254
1	
4	
5	critical t 2.080
4	
5	
3	t-test 3.578
5	
5	
Result:	Since the absolute value of 3.578 is greater than the critical t of 2.080, we reject the null hypothesis and accept the research hypothesis.
Conclusion:	Attendees at the Summer Educators' Conference of the American Marketing Association in Boston were significantly likely to recommend the Conference to a friend or colleague.

Fig. C.4 Practice Test Answer to Chap. 4 Problem

Practice Test Answer: Chapter 5 (see. Fig. C.5)

MASS MUTUAL FINANCIAL GROUP						
Item:	"How interested are you in learning more about how life insurance can provide income for retirement?"					
1	2	3	4	5	6	
Not at all interested		3.30		4.70		Very interested
Male model	Female model	Group	n	Mean	STDEV	
3	4	1 Male model	27	3.30	1.54	
2	6	2 Female model	27	4.70	1.07	
4	5	Null hypothesis: $\mu_1 = \mu_2$				
5	3	Research hypothesis: $\mu_1 \neq \mu_2$				
1	4					
6	6					
2	6	$1/n_1 + 1/n_2$		0.07		
4	5					
3	3					
5	5	$(n_1 - 1) \times S_1^2$		61.63		
2	4					
4	3					
3	5	$(n_2 - 1) \times S_2^2$		29.63		
5	4					
1	6					
2	5	$n_1 + n_2 - 2$ (degrees of freedom)		52		
3	5					
1	6					
4	4	s.e.		0.36		
5	6					
6	3					
2	4	critical t		1.96		
3	6					
1	5					
4	6	t-test		-3.90		
3	4					
5	4					
Result:	Since the absolute value of - 3.90 is greater than the critical t of 1.96, we reject the null hypothesis and accept the research hypothesis.					
Conclusion:	Adult men (ages 25-44) were significantly more interested in learning more about how life insurance can provide income for retirement when a female model was used than when a male model was used in the ad (4.70 vs. 3.30)					

Fig. C.5 Practice Test Answer to Chap. 5 Problem

Practice Test Answer: Chapter 6 (see. Fig. C.6)

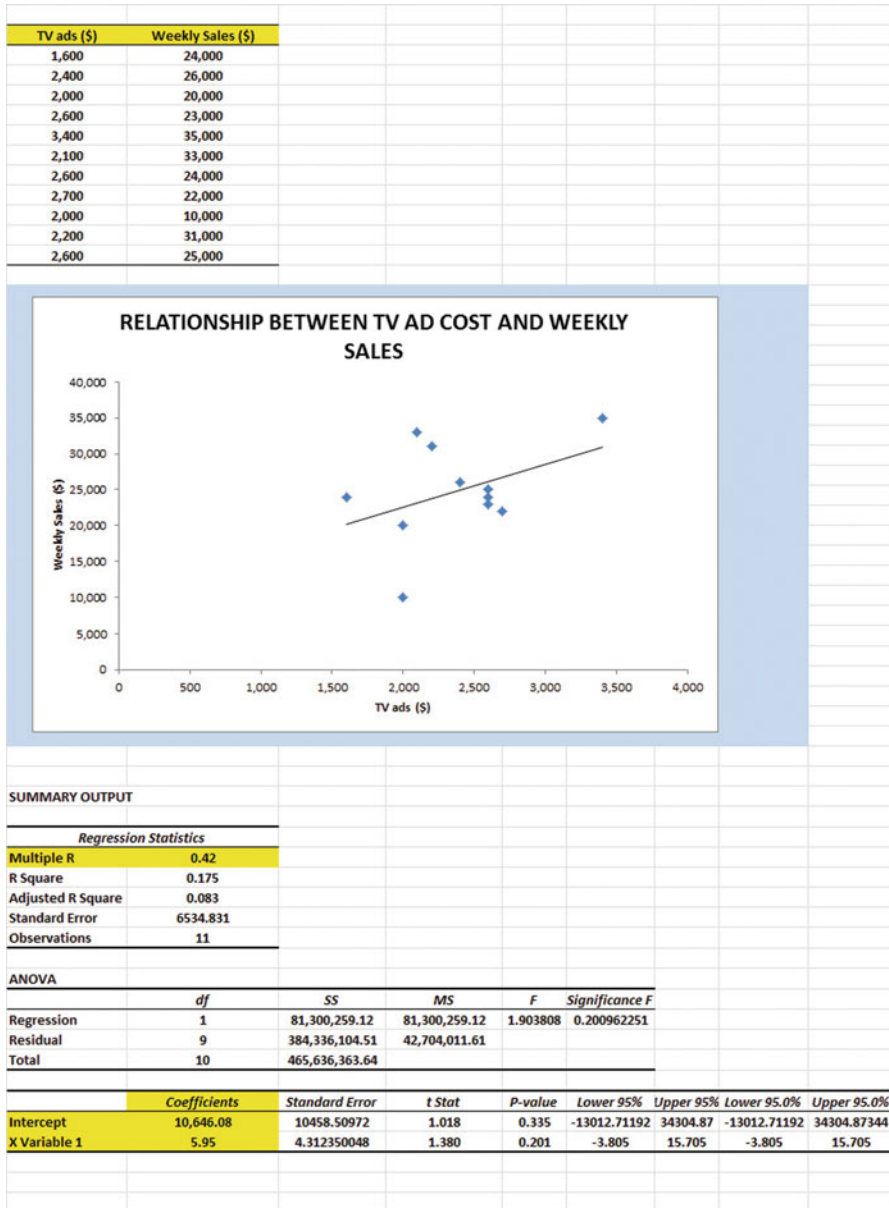


Fig. C.6 Practice Test Answer to Chap. 6 Problem

Practice Test Answer: Chapter 6: (continued)

(d) a = y-intercept = 10,646.08

b = slope = +5.95

(e) $Y = a + b X$

$$Y = 10,646.08 + 5.95 X$$

(f) r = correlation = +.42

(g) $Y = 10,646.08 + 5.95 (2500)$

$$Y = 10,646.08 + 14,875$$

$$Y = \$25,521.08$$

(h) About \$28,000–\$29,000

Practice Test Answer: Chapter 7 (see. Fig. C.7)

TOTAL GALLONS USED TO DRIVE FROM ST. LOUIS TO INDIANAPOLIS					
FOUR-DOOR SEDANS					
TOTAL GALLONS USED	WEIGHT (1000 lbs)	HORSEPOWER			
6.1	3.8	130			
6.3	3.7	150			
4.8	4.0	140			
4.2	2.4	125			
3.8	2.9	98			
4.7	3.0	115			
3.5	2.1	121			
5.5	2.9	123			
5.9	3.1	110			
3.4	2.1	96			
SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.77				
R Square	0.593				
Adjusted R Square	0.477				
Standard Error	0.787				
Observations	10				
ANOVA					
	df	SS	MS	F	Significance F
Regression	2	6.320	3.160	5.102	0.043
Residual	7	4.336	0.619		
Total	9	10.656			
	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	0.29	1.877	0.154	0.882	-4.150
WEIGHT (1000 lbs)	1.01	0.509	1.984	0.088	-0.194
HORSEPOWER	0.01	0.020	0.614	0.559	-0.035
	TOTAL GALLONS USED	WEIGHT (1000 lbs)	HORSEPOWER		
TOTAL GALLONS USED	1				
WEIGHT (1000 lbs)	0.76	1			
HORSEPOWER	0.60	0.65	1		

Fig. C.7 Practice Test Answer to Chap. 7 Problem

Practice Test Answer: Chapter 7 (continued)

1. $R_{xy} = +0.77$
2. y-intercept = 0.29
3. $b_1 = 1.01$
4. $b_2 = 0.01$
5. $Y = a + b_1 X_1 + b_2 X_2$
6. $Y = 0.29 + 1.01 X_1 + 0.01 X_2$
 $Y = 0.29 + 1.01 (3.8) + 0.01 (126)$
 $Y = 0.29 + 3.84 + 1.26$
 $Y = 5.39$ gallons
7. +.76
8. +.60
9. +.65
10. The better predictor of TOTAL GALLONS USED was WEIGHT with a correlation of $r_{xy} = +.76$.
11. The two predictors combined predict TOTAL GALLONS USED with a correlation of $R_{xy} = .77$ which is only very slightly better than the better single predictor by itself.

Let \$10 price = X_1 , \$12 price = X_2 , and \$14 price = X_3

- (b) $H_0: \mu_1 = \mu_2 = \mu_3$
 $H_1: \mu_1 \neq \mu_2 \neq \mu_3$
- (f) $MS_b = 5147.29$ and $MS_w = 728.33$
- (g) $F = 7.07$
- (h) Mean of \$12 Price = 167.18 and Mean of \$14 price = 134.74
- (j) $F = 7.07$
- (k) critical $F = 3.172$
- (l) Result: Since 7.07 is greater than 3.172, we reject the null hypothesis and accept the research hypothesis
- (m) Conclusion: There was a significant difference in the number of units sold between the three prices.

\$12 PRICE vs. \$14 PRICE

- (n) $H_0: \mu_2 = \mu_3$
 $H_1: \mu_2 \neq \mu_3$
- (o) $s.e. = \text{SQRT}(728.33 \times [1/17 + 1/19]) = \text{SQRT}(728.33 \times [.06 + .05]) = \text{SQRT}(728.33 \times 0.11) = \text{SQRT}(80.11) = 8.95$
- (p) ANOVA $t = (167.18 - 134.74)/8.95 = 32.44/8.95 = 3.62$
- (q) $df = n_{\text{TOTAL}} - k = 56 - 3 = 53$
- (r) critical $t = 1.96$

Practice Test Answer: Chapter 8 (continued)

- (s) Result: Since the absolute value of 3.62 is greater than the critical t of 1.96, we reject the null hypothesis and accept the research hypothesis
- (t) A \$12 Price sold significantly more units than the \$14 Price (167 vs. 135)

Appendix D: Statistical Formulas

Mean $\bar{X} = \frac{\Sigma X}{n}$

Standard Deviation $STDEV = S = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}}$

Standard error of the mean $s.e. = S_{\bar{X}} = \frac{S}{\sqrt{n}}$

Confidence interval about the mean $\bar{X} \pm tS_{\bar{X}}$
 where $S_{\bar{X}} = \frac{S}{\sqrt{n}}$

One-group t-test $t = \frac{\bar{X} - \mu}{S_{\bar{X}}}$
 where $S_{\bar{X}} = \frac{S}{\sqrt{n}}$

Two-group t-test

(a) when both groups have a sample size greater than 30

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}$$

$$\text{where } S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

and where $df = n_1 + n_2 - 2$

(b) when one or both groups have a sample size less than 30

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}$$

$$\text{where } S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

and where $df = n_1 + n_2 - 2$

Correlation

$$r = \frac{\frac{1}{n-1} \Sigma(X - \bar{X})(Y - \bar{Y})}{S_x S_y}$$

where S_x = standard deviation of X

and where S_y = standard deviation of Y

Simple linear regression

$$Y = a + b X$$

where a = y-intercept and b = slope of the line

Multiple regression equation

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \text{etc.}$$

where a = y-intercept

One-way ANOVA F-test

$$F = MS_b / MS_w$$

ANOVA t-test

$$ANOVA t = \frac{\bar{X}_1 - \bar{X}_2}{s.e. ANOVA}$$

$$\text{where } s.e. ANOVA = \sqrt{MS_w \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

and where $df = n_{TOTAL} - k$

where $n_{TOTAL} = n_1 + n_2 + n_3 + \text{etc.}$

and where k = the number of groups

Appendix E: t-Table

Critical t-values needed for rejection of the null hypothesis (see Fig. E.1)

Fig. E.1 Critical t-values
Needed for Rejection of the
Null Hypothesis

sample size n	degrees of freedom df	critical t
10	9	2.262
11	10	2.228
12	11	2.201
13	12	2.179
14	13	2.160
15	14	2.145
16	15	2.131
17	16	2.120
18	17	2.110
19	18	2.101
20	19	2.093
21	20	2.086
22	21	2.080
23	22	2.074
24	23	2.069
25	24	2.064
26	25	2.060
27	26	2.056
28	27	2.052
29	28	2.048
30	29	2.045
31	30	2.042
32	31	2.040
33	32	2.037
34	33	2.035
35	34	2.032
36	35	2.030
37	36	2.028
38	37	2.026
39	38	2.024
40	39	2.023
infinity	infinity	1.960

Index

A

- Absolute value of a number, 66–67
- Analysis of Variance
 - ANOVA t-test formula, 176
 - degrees of freedom, 177, 182, 183, 185, 230
 - Excel commands, 178–180
 - formula, 174
 - interpreting the Summary Table, 174
 - s.e. formula for ANOVA t-test, 176
- ANOVA, *see* Analysis of Variance
- ANOVA t-test, *see* Analysis of Variance
- Average function, *see* Mean

C

- Centering information within cells, 7
- Chart
 - adding the regression equation, 141–143
 - changing the width and height, 5–6
 - creating a chart, 119–129
 - drawing the regression line onto the chart, 119–129
 - moving the chart, 127–128
 - printing the spreadsheet, 13–15, 129–131
 - reducing the scale, 130
 - scatter chart, 121
 - titles, 121–123, 125
- Column width (changing), 5–6, 155
- Confidence interval about the mean
 - drawing a picture, 45, 53
 - formula, 41, 53
 - lower limit, 38–42, 45, 46, 53, 62, 63
 - 95% confident, 38–39, 41, 42, 54, 75, 222
 - upper limit, 38–42, 46, 53, 62, 63

- Correlation, 107–154, 160, 161, 163, 165, 166, 168, 204, 209, 227, 229, 237, 238, 241
 - formula, 111, 112, 114, 116, 118, 140, 160
 - negative correlation, 107, 109, 110, 139, 144, 150, 206
 - positive correlation, 107–109, 118, 144, 150, 163
 - 9 steps for computing r , 112–114
- CORREL function, *see* Correlation
- COUNT function, 9, 53
- Critical t-value, 59, 177, 178, 243

D

- Data Analysis ToolPak, 132–135, 154, 169
- Data/Sort commands, 28
- Degrees of freedom, 85–88, 90, 100, 177, 182, 183, 185, 219, 230

F

- Fill/Series/Columns commands, 4–5
 - step value/stop value commands, 5, 24
- Formatting numbers
 - currency format, 15–17
 - decimal format, 137

H

- Home/Fill/Series commands, 4
- Hypothesis testing
 - decision rule, 53, 66–67, 84
 - null hypothesis, 49–52
 - rating scale hypotheses, 49–52, 56, 70

- Hypothesis testing (*cont.*)
 research hypothesis, 49–53, 55–58, 60, 62, 63, 66, 69, 70, 76, 79
 7 steps for hypothesis testing, 52–58, 65–69
 stating the conclusion, 58
 stating the result, 58
- M**
 Mean, 1–22, 37–65, 67, 68, 70–76, 78, 79, 81–111, 114–120, 139, 150, 159, 160, 163, 166, 169, 174–176, 178, 180, 182–184, 217, 219, 220, 222, 224, 226, 230, 240, 241
 formula (1.1), 1
 Multiple correlation
 correlation matrix, 160–163
 Excel commands, 160–163
 Multiple regression
 correlation matrix, 160–163
 equation, 153, 160
 Excel commands, 156–160
 predicting Y, 153, 160
- N**
 Naming a range of cells, 8–9
 Null hypothesis, *see* Hypothesis testing
- O**
 One-group t-test for the mean
 absolute value of a number, 66–67
 formula, 65, 67
 hypothesis testing, 75
 s.e. formula, 67
 7 steps for hypothesis testing, 65–69
- P**
 Page Layout/Scale to Fit commands, 32
 Population mean, 37–40, 48, 50, 65, 67, 84, 91, 169, 174–176, 178
 Printing a spreadsheet
 entire worksheet, 144–146
 part of the worksheet, 144–146
 printing a worksheet to fit onto one page, 32–34, 129–131
- R**
 RAND, *see* Random number generator
 Random number generator
 duplicate frame numbers, 25–27, 35, 36, 221
 frame numbers, 23–26, 35, 36, 221
 sorting duplicate frame numbers, 28–30, 35, 36, 221
 Regression, 107–152, 156–159, 162, 164, 166, 167, 226–228, 242
 equation
 adding it to the chart, 141–143
 formula, 140
 negative correlation, 144
 predicting Y from X, 140–141
 slope, b, 139
 writing the regression equation using the Summary Output, 135–139
 y-intercept, a, 139
 line, 119–129, 139–144, 147–149, 151, 152, 227
 Research hypothesis, *see* Hypothesis testing
- S**
 Sample size, 1–22, 38, 41–43, 45, 48, 53, 60, 62, 63, 65, 68, 70, 71, 76, 78, 79, 81–85, 87, 90–93, 98, 100, 104, 106, 111, 112, 116, 117, 171, 177, 220, 222, 224, 241
 COUNT function, 9, 53
 Saving a spreadsheet, 12–13
 Scale to Fit commands, 32, 46
 s.e., *see* Standard error of the mean
 Standard deviation (STDEV), 1–22, 38, 39, 42, 43, 45, 46, 53, 62, 65, 67, 70, 71, 76, 78, 79, 83, 87, 88, 91–93, 99, 100, 104, 106, 111, 113, 116, 117, 220, 222, 224, 241
 formula, 2
 Standard error of the mean (s.e.), 1–22, 38–43, 45, 46, 53, 60, 62, 63, 65, 67, 71, 72, 76, 78, 79, 90, 91, 93, 101, 215, 220, 222, 224, 240, 241
 formula, 3
 STDEV, *see* Standard deviation
- T**
 t-table, 41, 67–69, 72, 76, 78, 79, 85–86, 102, 106, 177, 224, 226, 243
 Two-group t-test
 basic table, 83
 degrees of freedom, 85–86
 drawing a picture of the means, 89
 formula, 90
 formula #1, 90–98
 formula #2 (5.5), 98–103
 hypothesis testing, 84, 88–90
 9 steps in hypothesis testing, 82–90
 s.e. formula, 90, 100