Essentials of Biostatistics Workbook

Second Edition

Statistical Computing Using Excel® 2003

Lisa M. Sullivan, PhD

Professor and Chair, Department of Biostatistics
Associate Dean for Education
Boston University School of Public Health
Boston, Massachusetts

JONES & BARTLETT
LEARNING

World Headquarters
Jones & Bartlett Learning
40 Tall Pine Drive
Sudbury, MA 01776
978-443-5000
info@jblearning.com
www.jblearning.com

Jones & Bartlett Learning Canada
6339 Ormindale Way
Mississauga, Ontario
L5V 1J2
Canada

Jones & Bartlett Learning International
Barb House, Barb Mews
London W6 7PA
United Kingdom

Jones & Bartlett Learning books and products are available through most bookstores and online booksellers. To contact Jones & Bartlett Learning directly, call 800-832-0034, fax 978-443-8000, or visit our website www.jblearning.com.

Substantial discounts on bulk quantities of Jones & Bartlett Learning publications are available to corporations, professional associations, and other qualified organizations. For details and specific discount information, contact the special sales department at Jones & Bartlett Learning via the above contact information or send an email to specialsales@jblearning.com.

ISBN-13: 978-0-7637-9531-3
ISBN-10: 0-7637-9531-3

To order this product, use ISBN: 978-1-4496-2394-4

This publication is designed to provide accurate and authoritative information in regard to the Subject Matter covered. It is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional service. If legal advice or other expert assistance is required, the service of a competent professional person should be sought.

Essentials of Biostatistics Workbook: Statistical Computing Using Excel® 2003 is an independent publication and is not affiliated with, nor has it been authorized, sponsored, or otherwise approved by Microsoft Corporation.

Production Credits
Publisher: Michael Brown
Associate Editor: Maro Gartside
Editorial Assistant: Teresa Reilly
Production Assistant: Rebekah Linga
Senior Marketing Manager: Sophie Fleck
Composition: Publishers' Design and Production Services, Inc.
Cover Design: Kate Ternullo
Cover Image: © Kheng Guan Toh/ShutterStock, Inc.

3277

15 14 13 12 11   10 9 8 7 6 5 4 3 2 1

Contents

CHAPTER 1 Basics

In this workbook, we describe how Microsoft Office® Excel® can be used to perform the statistical computations and analyses described in the textbook. Excel is a popular program, often used for organizing and summarizing numerical or financial information. It has substantial graphing capabilities and a statistical analysis module that is designed to perform a number of statistical analyses. One of the primary reasons we use Excel is its accessibility. Whereas other statistical packages (e.g., SAS®, SPSS®, and S-PLUS) offer more advanced analytic techniques and procedures, Excel is suitable for the introductory procedures we present here.[1–3] In fact, Excel offers many more applications than those we present in this workbook. We focus on the concepts and procedures discussed in the textbook. Readers interested in broader applications of Excel should see Dretzke.[4] Before we proceed with specific analyses, we first present some basic terminology and general procedures to get started.

1.1 WORKBOOKS AND WORKSHEETS

Excel files are also called workbooks. A *workbook* is a set of worksheets, where each *worksheet* can be thought of as a table or grid of rows and columns. When we open the Excel program, a workbook with three blank worksheets is presented (this is the default, or preset starting point). Excel calls the new workbook *Book1*. The name can be changed when the workbook is saved (see Section 1.4 for details). The three worksheets are called *Sheet1*, *Sheet2*, and *Sheet3*. The names of the worksheets can also be changed. When Excel is opened, *Sheet1* is shown on the screen and looks like an empty grid of rows and columns; a sample is shown in Figure 1–1. The columns of the worksheet are labeled with letters (A, B, C, and so on) and the rows of the worksheet are numbered. The workbook name appears in the top-left corner (*Book1*), and the tabs along the bottom of the screen show the worksheet names.

It is useful to rename the worksheets to reflect the information stored in each. For example, we rename *Sheet1* as *Data*. This is done using the **Format** option along the top menu bar. Under the **Format** option, we choose the **Sheet** and then the **Rename** options (see Figure 1–2). Once we choose the **Rename** option, Excel places the cursor on the worksheet name at the bottom of the screen (*Sheet1* in this case), where we can enter the new name.

1.2 CELL ADDRESSES

A worksheet can be thought of as a set of cells. Each *cell* is defined by a specific column and row. When we first open Excel, the cursor appears in the top-left cell, making that the current or active cell. Notice in Figure 1–1 and Figure 1–2 that the top-left cell is outlined in a bold black line. The column and row make up the cell's *address*. The top-left cell's address is A1. As we move the cursor around the worksheet into different cells, the address of the current or active cell is shown just below the menu bars in the top-left portion of the screen.

1.3 ENTERING AND EDITING DATA

For statistical analysis, we enter data into the cells of an Excel worksheet. Once the data are entered, we can manipulate the values and perform statistical analyses. Example 1.1 contains data from a small study that we use to illustrate entering and manipulating data.

FIGURE 1–1 New Worksheet

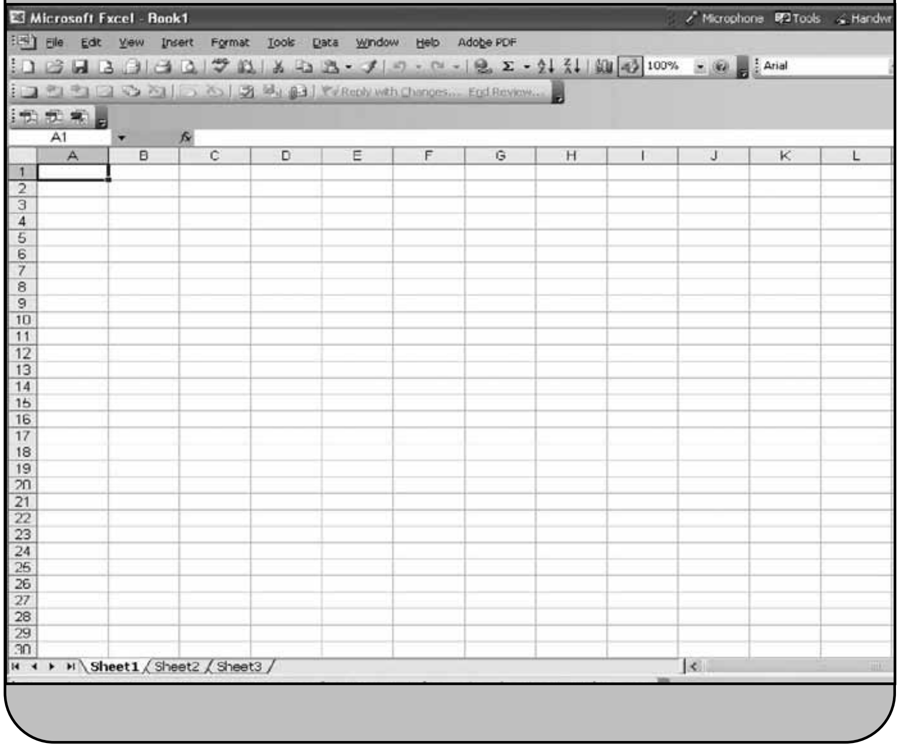


FIGURE 1–2 Renaming the Worksheet

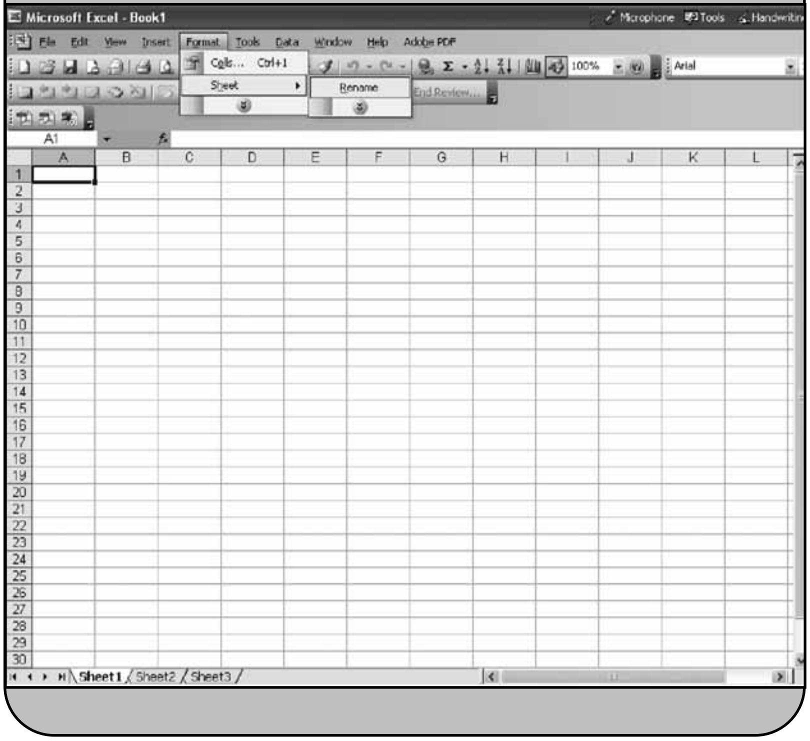


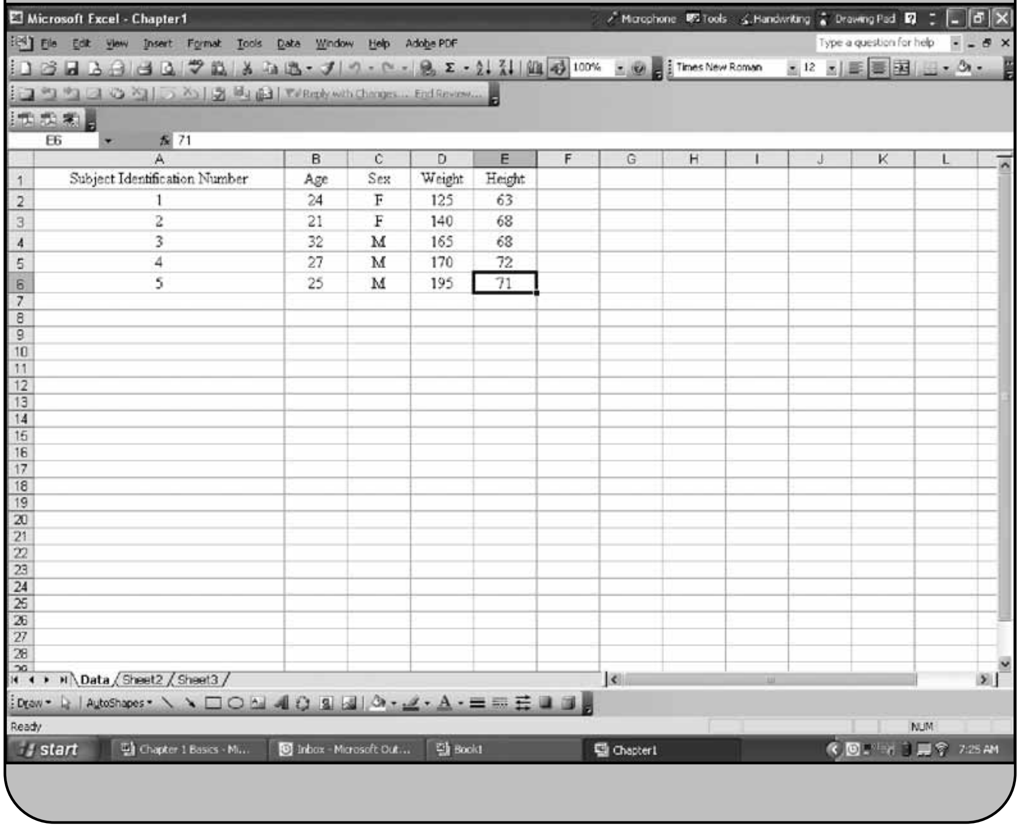
TABLE 1–1 Data from Study of 5 Participants

Subject Identification Number	Age	Sex	Weight (lb)	Height (in)
1	24	F	125	63
2	21	F	140	68
3	32	M	165	68
4	27	M	170	72
5	25	M	195	71

Example 1.1. Suppose we have a sample of $n = 5$ participants and on each participant we measure age, sex, weight, and height. We also assign each participant a unique identification number (shown in the first column of Table 1–1). The identification numbers are not used in statistical analysis but instead are used to keep track of data measured for each participant. The data are shown in Table 1–1.

In Excel, we use the columns to hold the different variables that are measured (e.g., identification number, age, sex, weight, and height) and the rows to hold the observations measured in different participants. We use the first row for the variable names—this is important as the variable names show on the output, making for easier interpretation. The data shown in Table 1–1 are entered into Excel by moving the cursor around the worksheet. Figure 1–3 shows the data entered into the worksheet we named *Data*. Notice that the variable names are contained in row 1 and the data measured on the five participants are shown in row 2 through row 6.

FIGURE 1–3 Study Data Entered into Excel Worksheet


Microsoft Excel - Chapter1
Microphone Tools Handwriting Drawing Pad
File Edit View Insert Format Tools Data Window Help Adobe PDF
Type a question for help
100%
Times New Roman 12
Reply with Changes... End Review...
Data Sheet2 Sheet3
Draw AutoShapes
Ready
NUM
start Chapter 1 Basics - Mi... Inbox - Microsoft Out... Book1 Chapter1
7:25 AM

E6 fx 71

	A	B	C	D	E
1	Subject Identification Number	Age	Sex	Weight	Height
2	1	24	F	125	63
3	2	21	F	140	68
4	3	32	M	165	68
5	4	27	M	170	72
6	5	25	M	195	71

There is no restriction on the format or length of variable names. However, it can be easier to work with shorter names, simply for convenient viewing of the names and data on the worksheet. It is important to choose informative names that reflect the information entered.

In Figure 1–3, the current or active cell is E6. The cell address is shown just below the menu bar in the top-left portion of the screen. The contents of the cell (71 in the example) are shown just to the right of the active cell's address. Once data are entered, we can change or modify entries simply by retyping over the contents of the current cell or by typing into the top row where the active cell's contents are shown (see Figure 1–3). We can move from cell to cell in the worksheet by using the mouse or by using the arrow keys on the keyboard.

There are some instances where the same data are repeated. In Example 1.1, there are two women and three men. Suppose we enter the sex of Participant 1 (i.e., the participant with identification number 1 whose data are in row 2 of the *Data* worksheet) into the C2 cell as "F". Rather than entering the sex of Participant 2 into cell C3 directly, we can copy the data from the C2 cell. First, we make the C2 cell the active cell by moving the cursor to that cell. We then click on the **Copy** icon on the menu bar. To let us know that the contents of the active cell have been selected, Excel shows the borders of the cell with a bold flashing dotted line (as opposed to a bold solid line). We then move the cursor to the destination cell (e.g., C3) and click on the **Paste** icon. The contents of cell C2 are copied and pasted into cell C3. The same idea can be used to copy the contents of one cell to several cells. Suppose we enter the sex of Participant 3 into cell C4 as "M" and want to copy the contents of cell C4 in to cells C5 and C6. We make cell C4 the active cell and click on the **Copy** icon. We then highlight the destination cells—in this case, cells C5 and C6—simultaneously. To do this, we place our cursor on cell C5 (the top or first cell in the range) and, holding the left mouse key down, we drag the cursor to cell C6. This highlights both cells C5 and C6. We then click the **Paste** icon and the contents of cell C4 are copied into cells C5 and C6.

To insert a row into a worksheet, we use the **Insert/Row** command on the menu bar. Once we select the **Insert/Row** option, a row is inserted above the active cell. The same approach can be taken to insert a column. Selecting the **Insert/Column** option from the menu bar inserts a column to the left of the active cell.

For reporting purposes, we often want to format data or results for a consistent presentation. The **Format** option on the main menu bar can be used to format the contents of any cell or range of cells in a worksheet. For example, we entered weights in pounds, as shown into our *Data* worksheet:

Weight
125
140
165
170
195

Suppose that weights were actually measured to the nearest hundredths place and the data were entered as follows:

Weight
125.45
140.05
165.16
170.39
195.47

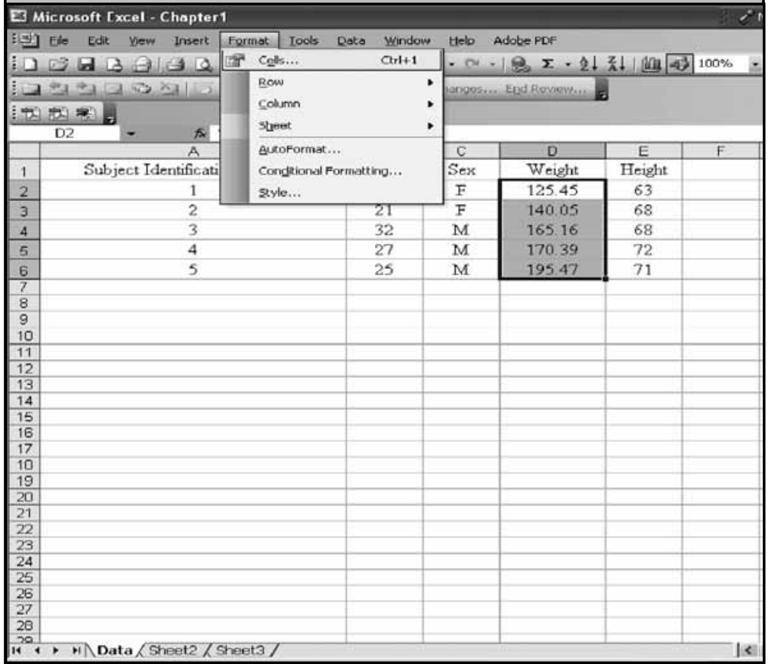
Suppose that we want to present the weights to the nearest tenths place (i.e., round the weights to one decimal place). This can be performed using the **Format** option on the main menu bar. We first highlight the range of cells we want to format (in our example, cells D2 through D6). We then click on the **Format/Cells** option. This is shown in Figure 1–4.

Choosing the **Format/Cells** option brings up the dialog box shown in Figure 1–5, where we specify the format we want for the selected cells. In Figure 1–5, we select the **Number** format from the category list on the left side of the dialog box and specify 1 decimal place. Once we click **OK**, the display in the worksheet changes to the following:

Weight
125.5
140.1
165.2
170.4
195.5

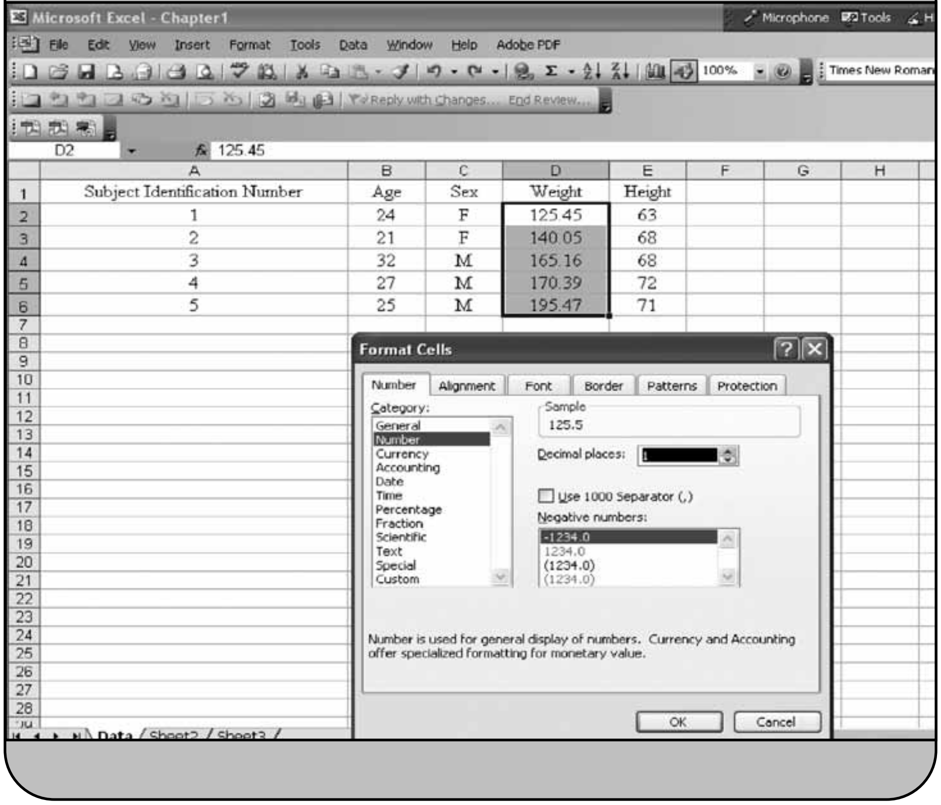
The **Format** option is particularly useful for formatting results. For example, when we compute the mean or standard deviation of a sample, Excel carries more decimal places than we generally want to present. Recall that, as a general rule, we report summary statistics using one more decimal place than the raw data. Reporting too many decimal places implies a false level of precision. We illustrate how to format results in Chapter 4 through Chapter 9 of the Excel workbook.

FIGURE 1–4 Formatting Data


Microsoft Excel - Chapter1
File Edit View Insert Format Tools Data Window Help Adobe PDF
Cells... Ctrl+1
Row
Column
Sheet
AutoFormat...
Conditional Formatting...
Style...
100%
Changes... End Review...
D2
Data Sheet2 Sheet3

	A	B	C	D	E	F
1	Subject Identificati		Sex	Weight	Height	
2	1		F	125.45	63	
3	2	21	F	140.05	68	
4	3	32	M	165.16	68	
5	4	27	M	170.39	72	
6	5	25	M	195.47	71	

FIGURE 1–5 Formatting Numeric Data


Microsoft Excel - Chapter1
Microphone Tools
File Edit View Insert Format Tools Data Window Help Adobe PDF
100%
Times New Roman
Reply with Changes... End Review...
Format Cells
Number Alignment Font Border Patterns Protection
General
Currency
Accounting
Date
Time
Percentage
Fraction
Scientific
Text
Special
Custom
OK Cancel
Data Sheet2 Sheet3

D2 125.45

	A	B	C	D	E	F	G	H
1	Subject Identification Number	Age	Sex	Weight	Height			
2	1	24	F	125.45	63			
3	2	21	F	140.05	68			
4	3	32	M	165.16	68			
5	4	27	M	170.39	72			
6	5	25	M	195.47	71			

Category: Number

Sample
125.5

Decimal places: 1

□ Use 1000 Separator (,)

Negative numbers:
-1234.0
1234.0
(1234.0)
(1234.0)

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

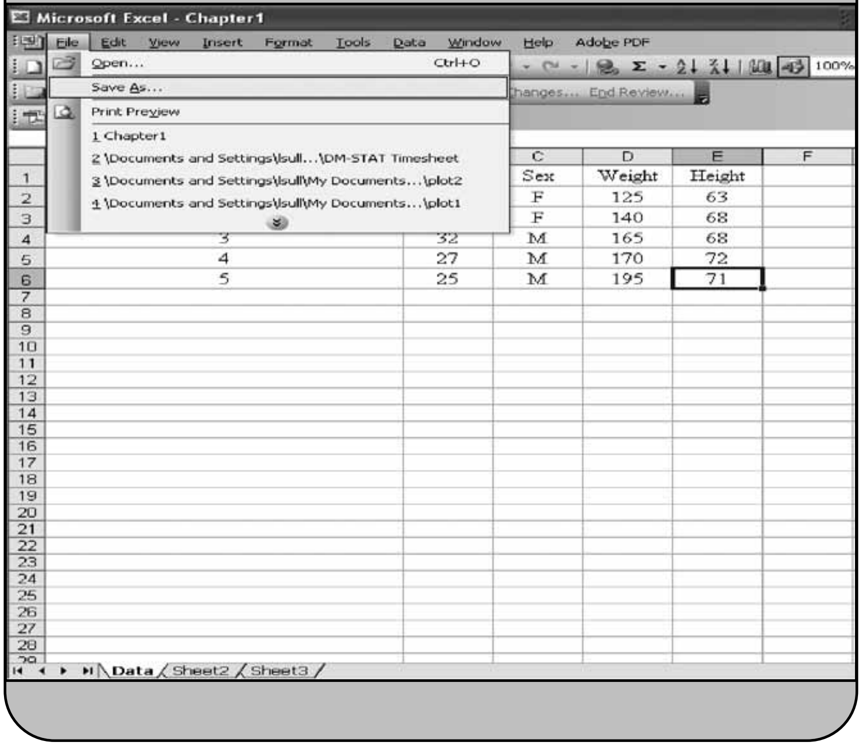
1.4 SAVING FILES

The **File/Save As** option is used to save a workbook (and its associated worksheets). Figure 1–6 shows the commands. After selecting the **File/Save As** option, Excel prompts us to enter a file name to store the workbook. Once a workbook is saved as a file, we can open it using the **File/Open** option for further use.

1.5 PRACTICE PROBLEMS

1. Use Excel to create a worksheet with the data shown in Table 1–2. The data were presented in Table 4–13 in the textbook and were measured in a subsample of *n* = 10 participants who attended the seventh examination of the Framingham Offspring Study. Place the variable names in the first row of the worksheet.

FIGURE 1–6 Saving the Workbook


Microsoft Excel - Chapter1
File
Edit
View
Insert
Format
Tools
Data
Window
Help
Adobe PDF
Open... Ctrl+O
Save As...
Print Preview
1 Chapter1
2 \Documents and Settings\lsull...\DM-STAT Timesheet
3 \Documents and Settings\lsull\My Documents...\plot2
4 \Documents and Settings\lsull\My Documents...\plot1
100%
Changes... End Review...
Data
Sheet2
Sheet3

			C	D	E	F
1			Sex	Weight	Height	
2			F	125	63	
3			F	140	68	
4	3	32	M	165	68	
5	4	27	M	170	72	
6	5	25	M	195	71	

TABLE 1–2 Data for Practice Problem 1

Participant ID	Systolic Blood Pressure	Diastolic Blood Pressure	Total Serum Cholesterol	Weight (lb)	Height (in)
1	141	76	199	138	63.00
2	119	64	150	183	69.75
3	122	62	227	153	65.75
4	127	81	227	178	70.00
5	125	70	163	161	70.50
6	123	72	210	206	70.00
7	105	81	205	235	72.00
8	113	63	275	151	60.75
9	106	67	208	213	69.00
10	131	77	159	142	61.00

2. Rename the worksheet with the data from Problem 1 as *Data*.
3. Save the Excel workbook as a file.

REFERENCES

1. SAS Version 9.1. © 2002–2003 by SAS Institute Inc., Cary, NC.
2. SPSS® Version 15.0. © 2006 by SPSS Inc., Chicago, IL.
3. S-PLUS Version 7.0. © 1999–2006 by Insightful Corp., Seattle, WA.
4. Dretzke, B.J. *Statistics with Microsoft Excel* (3rd ed.). Upper Saddle River, NJ: Pearson Prentice Hall, 2005.

CHAPTER 2

Formulas, Functions, and the Data Analysis ToolPak

Once data are entered into a Microsoft Office® Excel® worksheet, we can create formulas and functions to organize, manipulate, and analyze the data. For example, Excel can be used to create new variables from existing variables (e.g., to convert variables from one scale of measurement to another, to standardize variables into *z* scores, or to create new variables from those that are measured directly) or to compute summary statistics (e.g., the mean, standard deviation, or median of a dataset, or the minimum or maximum values in a set).

2.1 BASIC MATHEMATICAL OPERATIONS

Basic mathematical operations are performed in Excel as they are on a calculator or in other statistical computing packages. In Excel, basic operations are implemented with the operators shown in Table 2–1.

The order of operations is exponentiation first, then multiplication and division, and then addition and subtraction. To implement mathematical operations, we "program" specific operations into the cells of a worksheet. For example, we use these operations to convert variables measured on one scale to another or to create new variables from existing variables.

TABLE 2–1 Mathematical Operators in Excel

Operation	Operator
Multiplication	*
Division	/
Addition	+
Subtraction	-
Exponentiation	^

In Example 1.1 of the Excel workbook, we presented data on $n = 5$ participants. We measured age (in years), sex (male/female), weight (in pounds), and height (in inches). The data for Participant 1 are shown in Table 2–2. Using Excel, we could convert age measured in years to age in months by multiplying age in years by 12 (e.g., $\text{Age}_{\text{months}} = \text{Age}_{\text{years}} \times 12$). We could also convert weight in pounds to weight in kilograms using $\text{Weight}_{\text{kilograms}} = \text{Weight}_{\text{pounds}}/0.4636$. Excel can be used to make these transformations easily. We illustrate how this is done in Section 2.2.

Example 2.1. Consider a study designed to assess the impact of a medication designed to lower systolic blood pressure. Suppose we measure each participant's baseline systolic blood pressure (SBP) and their systolic blood pressure after 6 months on treatment. The data on $n = 3$ participants are entered into Excel and shown in Figure 2–1.

Notice that the distinct variables (e.g., ID and the baseline and 6-month systolic blood pressures) are shown in the columns and the data for each participant are shown in the rows of the worksheet. To analyze these data, we use methods for dependent, matched, or paired samples and focus specifically on differences in blood pressures. For each participant, we need to first compute the differences. We can take differences as

$$\text{Difference} = \text{6-Month SBP} - \text{Baseline SBP}$$

TABLE 2–2 Participant Data

Subject Identification Number	Age	Sex	Weight (lb)	Height (in)
1	24	F	125	63

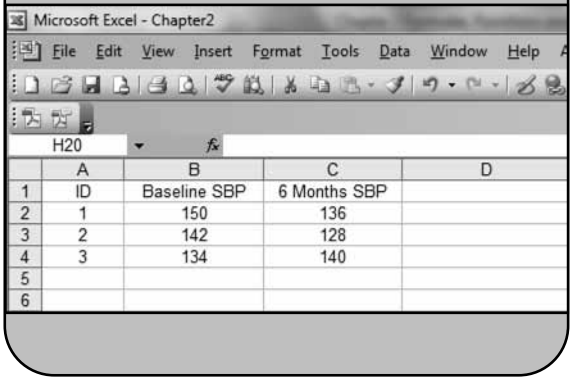
2.2 RELATIVE AND ABSOLUTE CELL REFERENCES

The operations described in Section 2.1 can be implemented in Excel by "programming" the operations into cells in the Excel worksheet. The programming amounts to specifying equations in Excel to perform the desired operations. We are essentially creating new variables as functions of existing variables using specific operations (e.g., converting from one scale to another, creating difference scores). To implement these operations, we first choose a column location for the new variable and specify a name for the new variable. The new variable name is placed in the first row of the worksheet along with the other variable names. We then input the operation or formula to create the new variable. In Excel, these operations are indicated by an equals sign ("="). When Excel sees an equals sign in a cell, it is expecting a formula to follow.

The formula in a cell is implemented to produce the desired result, which is placed into that cell. Figure 2–2 shows the data from Example 1.1 of the Excel workbook in a worksheet. Suppose we want to create a new variable, age in months, and label it "Age, months." We first choose a location for the variable. Suppose we want to place the new variable in column F of the worksheet. We enter the new name into row 1 of column F, as shown in Figure 2–2. Age in months is computed by multiplying age in years (which is contained in column B) by 12. Specifically, the formula to create age in months is age (in years) × 12. This formula is entered into cell F2 as "=B2*12". In this formula, B2 represents the address of the cell containing the age in years for Participant 1. Once the formula is entered, Excel takes the value from B2 (24 years) and multiplies it by 12. The result (288) is placed into cell F2.

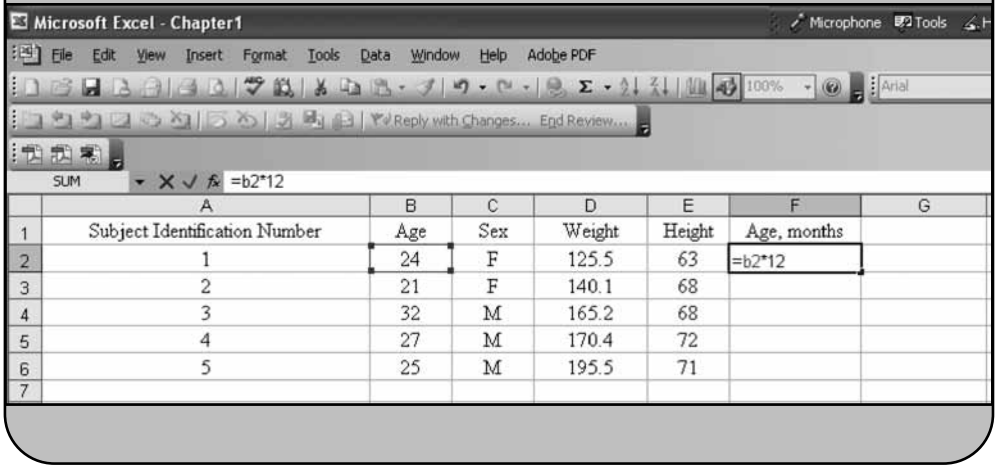
To perform this operation for each participant, we copy the formula in cell F2 and paste it into cell F3 through cell F6. This is done by making F2 the active cell and clicking the **Copy** icon. The border of cell F2 is shown in a bold, flashing dotted line. We next highlight cell F3 through cell F6 and click the **Paste** icon. The formula is copied from cell F2 into cell F3 through cell F6, and Excel automatically updates the cell referencing (i.e., the lo-

FIGURE 2–1 Systolic Blood Pressure Measured at Baseline and 6 Months Later



	A	B	C	D
1	ID	Baseline SBP	6 Months SBP	
2	1	150	136	
3	2	142	128	
4	3	134	140	
5				
6				

FIGURE 2–2 Creating New Variables


Microsoft Excel - Chapter1
Microphone
Tools
File Edit View Insert Format Tools Data Window Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
SUM

=b2*12

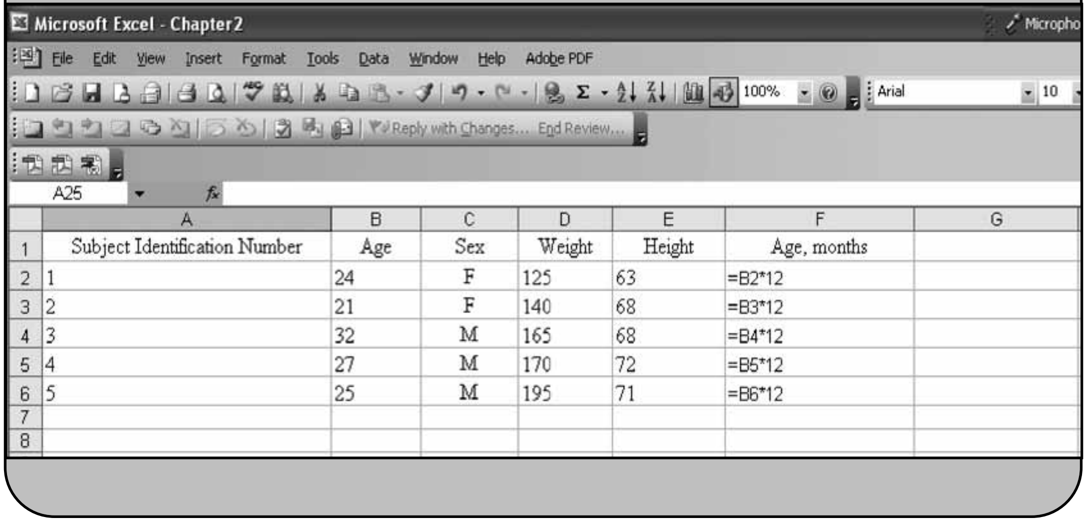
	A	B	C	D	E	F	G
1	Subject Identification Number	Age	Sex	Weight	Height	Age, months	
2	1	24	F	125.5	63	=b2*12	
3	2	21	F	140.1	68		
4	3	32	M	165.2	68		
5	4	27	M	170.4	72		
6	5	25	M	195.5	71		
7							

cations of the cells that contain the ages in years for each participant). Specifically, when we enter the formula to compute age in months for Participant 1 into F2, we specify that Excel should take the data in cell B2 and multiply it by 12. We want to do the corresponding for the remaining participants: for each participant, we want to multiply their age in years by 12. When we copy and paste the formula from cell F2 into cell F3 through cell F6, Excel updates the cell references as shown in Figure 2–3.

Excel automatically updates the formula to compute age in months for Participant 2 through Participant 5 by updating the cell references (i.e., to cell B3 through cell B6). These references are called *relative cell references*. The formula to compute age in months for each participant uses the relevant information, the age in years for that participant, contained in column B.

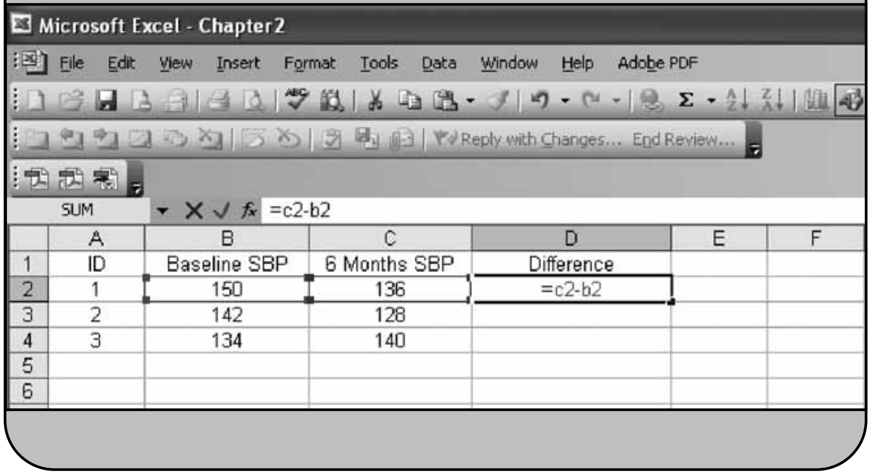
Consider again the data in Example 2.1 shown in Figure 2–1. Suppose we now want to create the difference variable. Figure 2–4 shows the new variable (column D) and the formula to compute it in cell D2. The difference score is computed by subtracting the baseline SBP (column B) from the SBP measured at 6 months (column C). The formula for Participant 1 is "=C2−B2". If we copy the contents of cell D2

FIGURE 2–3 Using Relative Cell References


Microsoft Excel - Chapter2
Micropho
File Edit View Insert Format Tools Data Window Help Adobe PDF
100%
Arial
10
Reply with Changes... End Review...
A25

	A	B	C	D	E	F	G
1	Subject Identification Number	Age	Sex	Weight	Height	Age, months	
2	1	24	F	125	63	=B2*12	
3	2	21	F	140	68	=B3*12	
4	3	32	M	165	68	=B4*12	
5	4	27	M	170	72	=B5*12	
6	5	25	M	195	71	=B6*12	
7							
8							

FIGURE 2–4 Computing Differences


Microsoft Excel - Chapter2
File Edit View Insert Format Tools Data Window Help Adobe PDF
Reply with Changes... End Review...

SUM =c2-b2

	A	B	C	D	E	F
1	ID	Baseline SBP	6 Months SBP	Difference		
2	1	150	136	=c2-b2		
3	2	142	128			
4	3	134	140			
5						
6						

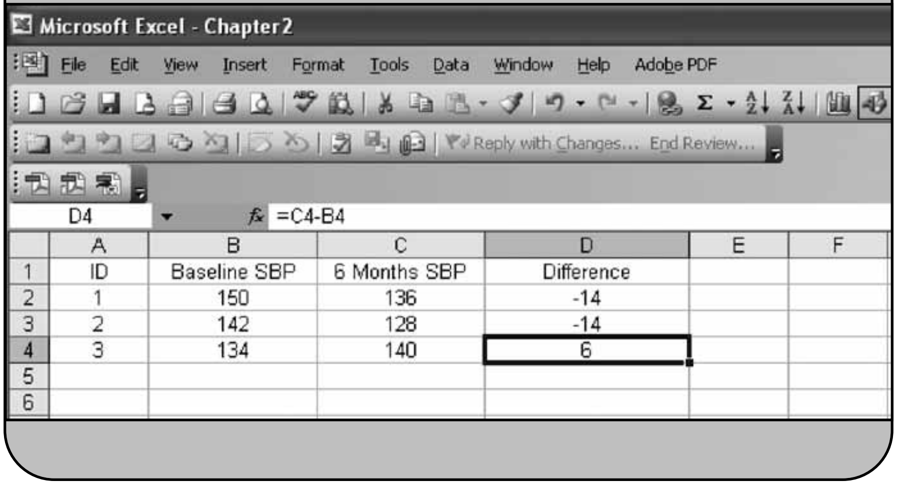
into cell D3 and cell D4, Excel automatically updates the cell referencing. The new formulas are "=C3−B3" and "=C4−B4", respectively. Once the formulas are entered, Excel computes the differences and the results are shown in Figure 2–5.

Example 2.2. Suppose we measure the lengths (in centimeters) of $n = 6$ infant boys who are 12 months of age. Suppose we want to standardize the lengths by subtracting the mean and dividing by the standard deviation. The mean length for 12–month-old boys is 75 centimeters and the standard deviation is 2.1 centimeters. The data are entered into an Excel worksheet as shown in Figure 2–6.

The lengths in centimeters are shown in column A. We standardize the lengths by subtracting the mean and dividing by the standard deviation: $Z = (\text{Length} - 75)/ 2.1$. To create the new variable Z, we enter the formula as shown in Figure 2–7. If we copy the formula from cell B2 into cell B3 through cell B7, Excel updates the reference to cell A2 to use the lengths in centimeters in cell A3 through cell A7, respectively.

There is a second way to perform the standardization. Suppose we enter the data into an Excel worksheet as shown in Figure 2–8. The mean and standard deviation are now shown in cell B9 and cell B10, respectively. We again create Z scores by

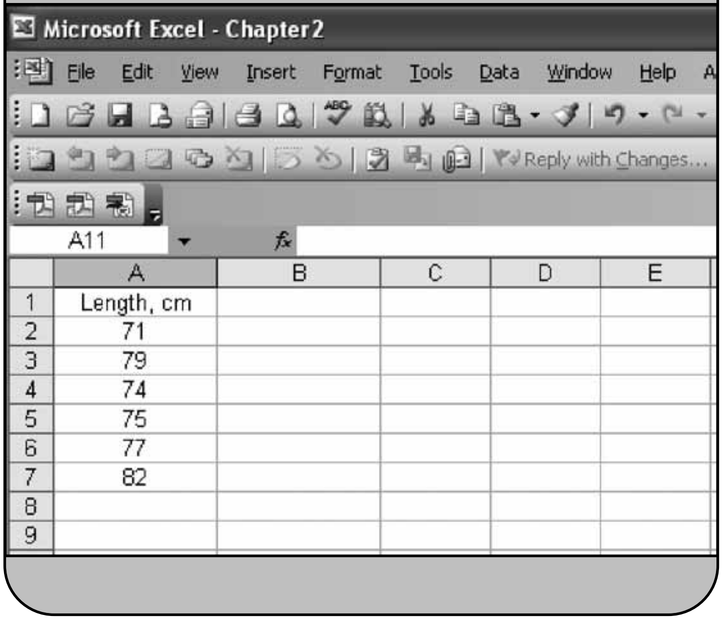
FIGURE 2–5 Difference Scores


Microsoft Excel - Chapter2
File Edit View Insert Format Tools Data Window Help Adobe PDF
Reply with Changes... End Review...

D4 =C4-B4

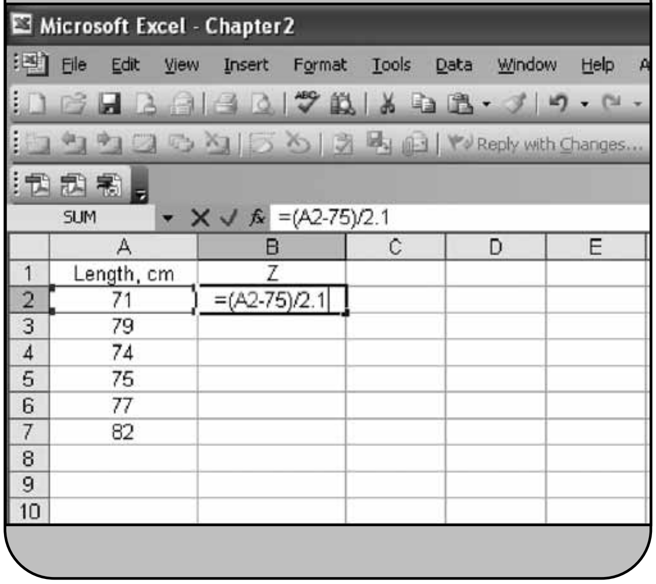
	A	B	C	D	E	F
1	ID	Baseline SBP	6 Months SBP	Difference		
2	1	150	136	-14		
3	2	142	128	-14		
4	3	134	140	6		
5						
6						

FIGURE 2–6 Lengths of Boys 12 Months of Age


Microsoft Excel - Chapter2
File Edit View Insert Format Tools Data Window Help
Reply with Changes...
A11
fx

	A	B	C	D	E
1	Length, cm				
2	71				
3	79				
4	74				
5	75				
6	77				
7	82				
8					
9					

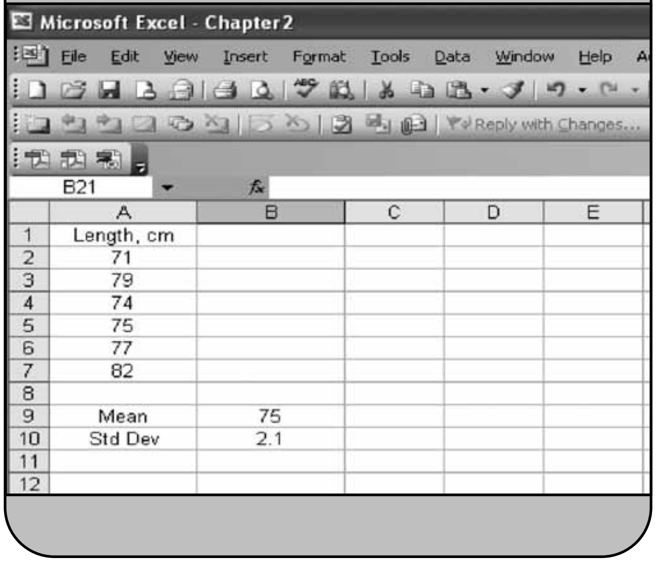
FIGURE 2–7 Standardizing the Lengths


Microsoft Excel - Chapter2
File Edit View Insert Format Tools Data Window Help
Reply with Changes...
SUM
fx

=(A2-75)/2.1

	A	B	C	D	E
1	Length, cm	Z			
2	71	=(A2-75)/2.1			
3	79				
4	74				
5	75				
6	77				
7	82				
8					
9					
10					

FIGURE 2–8 Lengths of Boys with Mean and Standard Deviation

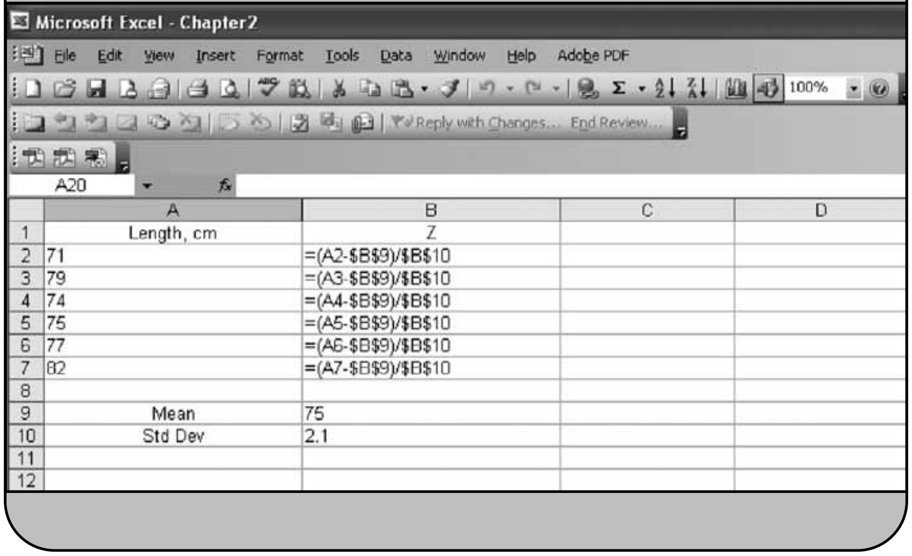

Microsoft Excel - Chapter2
File Edit View Insert Format Tools Data Window Help
Reply with Changes...
B21
fx

	A	B	C	D	E
1	Length, cm				
2	71				
3	79				
4	74				
5	75				
6	77				
7	82				
8					
9	Mean	75			
10	Std Dev	2.1			
11					
12					

taking each length in column A, subtracting the mean of 75 and dividing by the standard deviation of 2.1. What we do here is refer Excel to the cells containing the mean and standard deviation (i.e., cell B9 and cell B10) in the worksheet. If we again place the *Z* scores in column B, the formula entered in cell B2 is "=(A2−$B$9)/$B$10". A dollar sign before a column or row in a cell address freezes or fixes that column or row so that, as opposed to relative addresses as in the previous examples, Excel is not allowed to update it. In this example, we are fixing both the columns and rows of the addresses of the mean and standard deviation. These are called *absolute cell references*. Figure 2–9 displays the formulas that are copied into cell B3 through cell B7; notice that the cell addresses for the mean and standard deviation do not change from cell to cell.

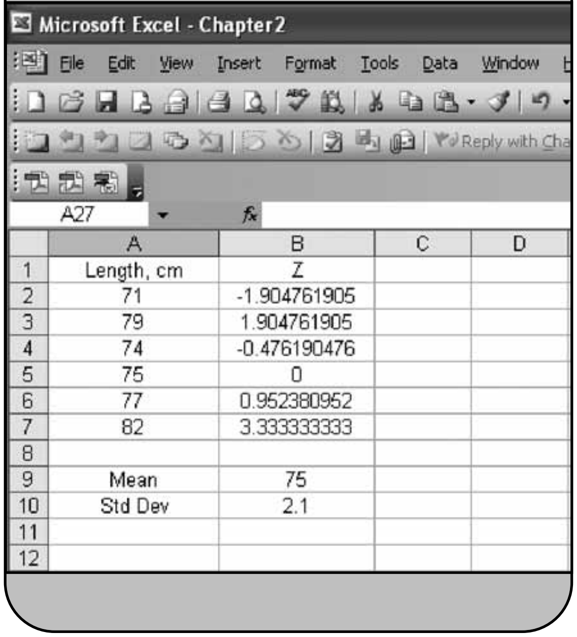
When we enter the formulas, the results are shown in column B. Both of the methods illustrated in Figure 2–7 and Figure 2–9 produce the results shown in Figure 2–10. The boy of length 71 cm is 1.9 standard deviations below the mean, whereas the boy of length 79 cm is 1.9 standard deviations above the mean. For presentation purposes, we format the cells in column B to two decimal places (using the **Format/Cells** option; see Figure 1–4 and Figure 1–5).

FIGURE 2–9 Using Absolute Cell References


Microsoft Excel - Chapter2
File Edit View Insert Format Tools Data Window Help Adobe PDF
100%
Reply with Changes... End Review...
A20
fx

	A	B	C	D
1	Length, cm	Z		
2	71	=(A2-$B$9)/$B$10		
3	79	=(A3-$B$9)/$B$10		
4	74	=(A4-$B$9)/$B$10		
5	75	=(A5-$B$9)/$B$10		
6	77	=(A6-$B$9)/$B$10		
7	82	=(A7-$B$9)/$B$10		
8				
9	Mean	75		
10	Std Dev	2.1		
11				
12				

FIGURE 2–10 Standardizing Lengths


Microsoft Excel - Chapter2
File Edit View Insert Format Tools Data Window
Reply with Cha
A27
fx

	A	B	C	D
1	Length, cm	Z		
2	71	-1.904761905		
3	79	1.904761905		
4	74	-0.476190476		
5	75	0		
6	77	0.952380952		
7	82	3.333333333		
8				
9	Mean	75		
10	Std Dev	2.1		
11				
12				

2.3 CREATING FORMULAS AND FUNCTIONS

We now describe how Excel is used to compute summary statistics (e.g., $\overline{X}$, s, median) using formulas and functions.

Example 2.3. In many studies of cardiovascular disease (e.g., the Framingham Heart Study), body mass index is assessed as a risk factor. Body mass index (BMI) is defined as

$$\text{BMI} = \text{weight}_{\text{kilograms}} \,/\, \text{height}^2_{\text{meters}}$$

Often weights are measured in pounds and heights in inches. Thus, the observed measurements must be converted to kilograms and meters, respectively, and then divided to produce BMI scores. In Example 1.1 of the Excel workbook, we measured weight in pounds and height in inches in $n = 5$ participants. Suppose we now want to create a BMI for each participant. The conversion from pounds to kilograms is 1 pound = 0.4536 kilograms, and the conversion from inches to meters is 1 inch = 0.0254 meters. The formula to compute BMI from weight in pounds and height in inches is

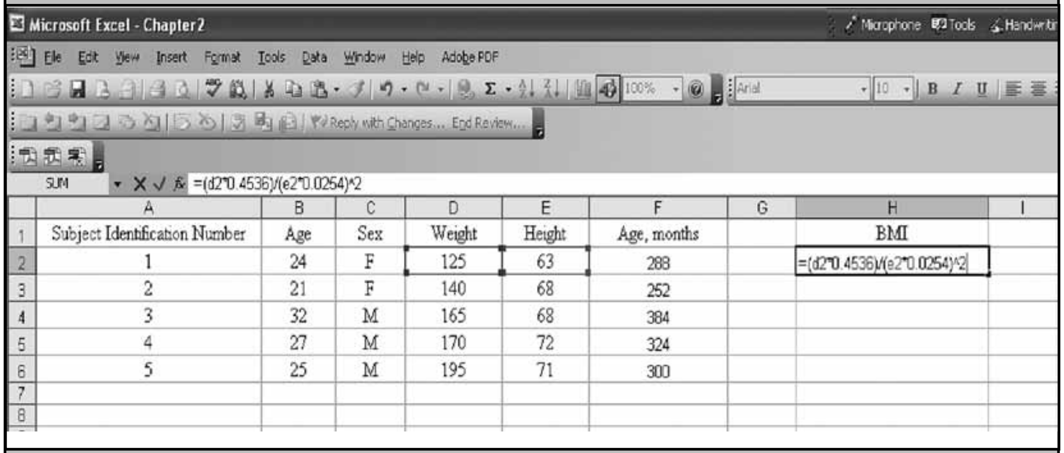
$$\text{BMI} = \frac{\text{weight}_{\text{pounds}} \times 0.4536}{(\text{height}_{\text{inches}} \times 0.0254)^2}$$

Figure 2–11 shows the computation of BMI. The formula in cell H2 can be copied to cell H3 through cell H6 to compute BMI for each participant. Notice the power operator "^2", used to square the height in the denominator of the formula. Once the formula is copied, the BMI scores are computed as shown in Figure 2–12. Before we compute summary statistics on the BMI data, we first use the **Format/Cells** option to format the BMIs to two decimal places (Figure 2–13).

Excel has a number of functions and formulas that can be used to summarize and analyze data. We now use Excel to compute the sample mean BMI (i.e., $\overline{X} = \Sigma X \,/\, n$) using these functions. We first sum the BMI scores and place the sum in cell H8. This is done with the SUM function. In cell H8, we enter the formula "=SUM(H2:H6)". The SUM function sums the data in the cells listed in the range shown in parentheses. In this example, we want to sum the data in cell H2 through cell H6. We then compute the sample size using the COUNT function and place the sample size into cell H9. In cell H9, we enter the formula "=COUNT(H2:H6)". The COUNT function tallies the number of cells with non-missing data. The sample mean is computed by dividing the sum by the sample size, and we place the sample mean into cell H10. Specifically, in cell H10 we enter the formula "=H8/H9". We use column G for labels (Figure 2–13).

Suppose we now wish to compute the standard deviation of BMI (i.e., $s = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n-1}}$). We need to first subtract the mean BMI (in cell H10) from each BMI and square the difference. We place the squared differences in column I, and the formula entered in cell I2 is "=(H2−$H$10)^2" and then copied into cell I3 through cell I6. We then sum the squared differences using the SUM function (i.e., "=SUM(I2:I6)") and place the result in cell I8. The variance is computed by dividing the

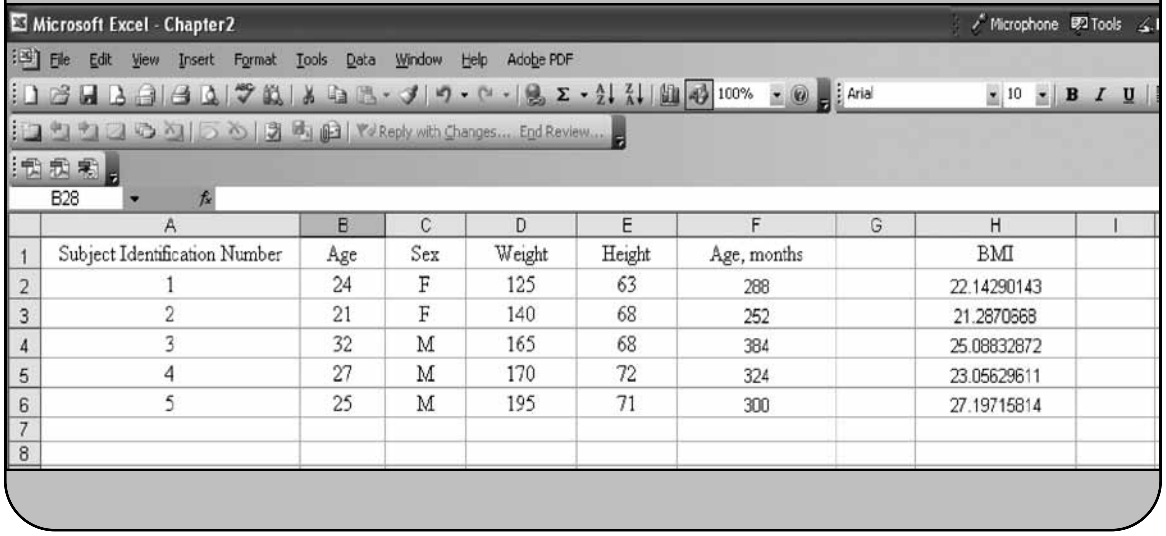
FIGURE 2–11 Computing BMI From Height and Weight



=(d2*0.4536)/(e2*0.0254)^2

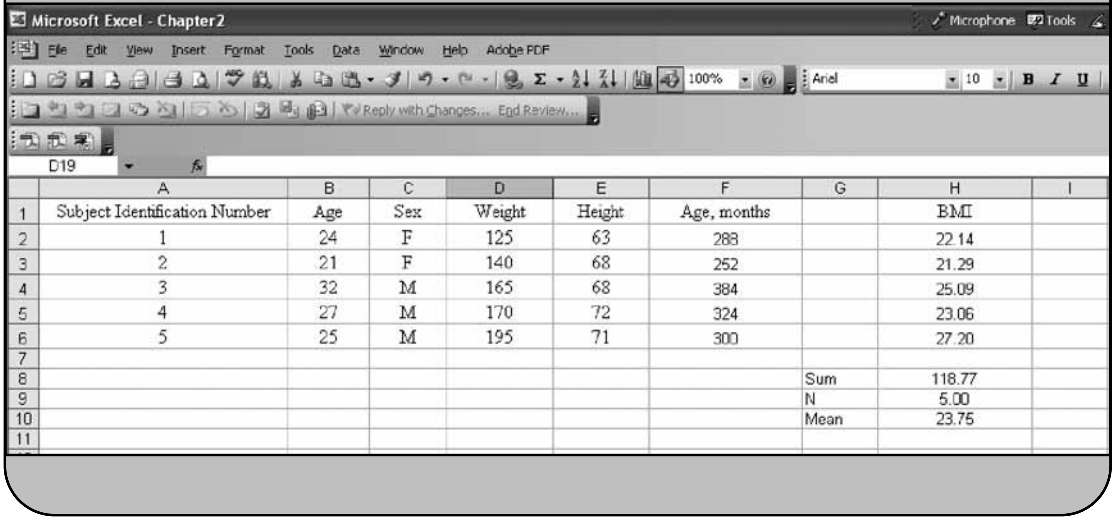
	A	B	C	D	E	F	G	H	I
1	Subject Identification Number	Age	Sex	Weight	Height	Age, months		BMI	
2	1	24	F	125	63	288		=(d2*0.4536)/(e2*0.0254)^2	
3	2	21	F	140	68	252			
4	3	32	M	165	68	384			
5	4	27	M	170	72	324			
6	5	25	M	195	71	300			
7									
8									

FIGURE 2–12 Computing Body Mass Index



	A	B	C	D	E	F	G	H	I
1	Subject Identification Number	Age	Sex	Weight	Height	Age, months		BMI	
2	1	24	F	125	63	288		22.14290143	
3	2	21	F	140	68	252		21.2870668	
4	3	32	M	165	68	384		25.08832872	
5	4	27	M	170	72	324		23.05629611	
6	5	25	M	195	71	300		27.19715814	
7									
8									

FIGURE 2–13 Computing the Sample Mean

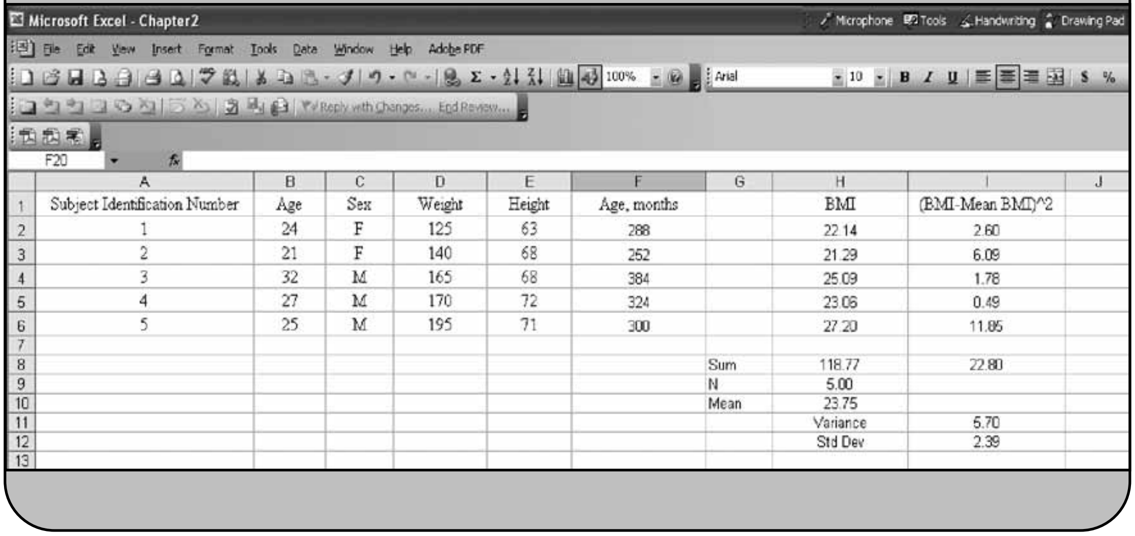


	A	B	C	D	E	F	G	H	I
1	Subject Identification Number	Age	Sex	Weight	Height	Age, months		BMI	
2	1	24	F	125	63	288		22.14	
3	2	21	F	140	68	252		21.29	
4	3	32	M	165	68	384		25.09	
5	4	27	M	170	72	324		23.06	
6	5	25	M	195	71	300		27.20	
7									
8							Sum	118.77	
9							N	5.00	
10							Mean	23.75	
11									

sum of the squared differences by $(n-1)$. We compute the variance in cell I11 as "=I8/(H9−1)". The standard deviation is then computed as "=SQRT(I11)". Figure 2–14 displays the results.

Like many statistical computing packages, there are several ways to compute summary statistics in Excel. One way is to use the mathematical operations to program the formulas. A second method is to use one of Excel's many built-in functions that directly compute summary statistics on a continuous variable. For example, Excel has an AVERAGE function that computes a mean. First, we select a cell for the result. Suppose we wish to compute the mean age for the data shown in Figure 2–14, and we want to place the mean age in cell B8. We enter the following into cell B8: "=AVERAGE(B2:B6)". Once the formula is entered, the mean of the observations contained in cell B2 through cell B6 is computed and placed in cell B8. The AVERAGE function sums the data specified in parentheses (in the worksheet, the age data are in cell B2 through cell B6) and divides by the sample size (i.e., the total number of non-

FIGURE 2–14 Computing the Sample Variance and Sample Standard Deviation


Microsoft Excel - Chapter2
File Edit View Insert Format Tools Data Window Help Adobe PDF
F20

	A	B	C	D	E	F	G	H	I	J
1	Subject Identification Number	Age	Sex	Weight	Height	Age, months		BMI	(BMI-Mean BMI)^2	
2	1	24	F	125	63	288		22.14	2.60	
3	2	21	F	140	68	252		21.29	6.09	
4	3	32	M	165	68	384		25.09	1.78	
5	4	27	M	170	72	324		23.06	0.49	
6	5	25	M	195	71	300		27.20	11.85	
7										
8							Sum	118.77	22.80	
9							N	5.00		
10							Mean	23.75		
11								Variance	5.70	
12								Std Dev	2.39	
13										

missing values). There are other functions available in Excel that are useful for computing summary statistics; these are discussed in Chapter 4 of the Excel workbook.

A third option for computing summary statistics is through the Excel Data Analysis ToolPak. The ToolPak offers a number of modules designed to perform various statistical analyses. We introduce the ToolPak here and use it extensively in Chapter 4 through Chapter 9 of the Excel workbook to perform statistical analysis.

2.4 THE DATA ANALYSIS TOOLPAK

The Excel Data Analysis ToolPak is an additional module (or add-in) that must be loaded either at the time of installation of Excel or at a later date. If the ToolPak was loaded at installation, it will be available as an option on the **Tools** menu (Figure 2–15).

If the Data Analysis ToolPak is not available, it can be loaded at any time. This is done using the **Tools/Add-Ins** option (Figure 2–15). Once the **Add-Ins** option is selected, a dialog box appears with several additional modules, one of which is the Data Analysis ToolPak. If we check the box next to "Analysis ToolPak" (Figure 2–16) and click on **OK**, the ToolPak will be added and available under the **Tools** menu.

The Data Analysis ToolPak can be used to perform many statistical computations. Table 2–3 lists the analyses that are available through the Data Analysis ToolPak that are discussed in the textbook. Table 2–3 lists the procedures alphabetically, along with the chapters from the textbook where these procedures are discussed in more detail. (Excel offers other procedures, but we restrict our attention to only those discussed in the textbook.)

There are some analyses (e.g., chi-square tests) that are not available in the Data Analysis ToolPak. In addition to those in

FIGURE 2–15 The Data Analysis ToolPak

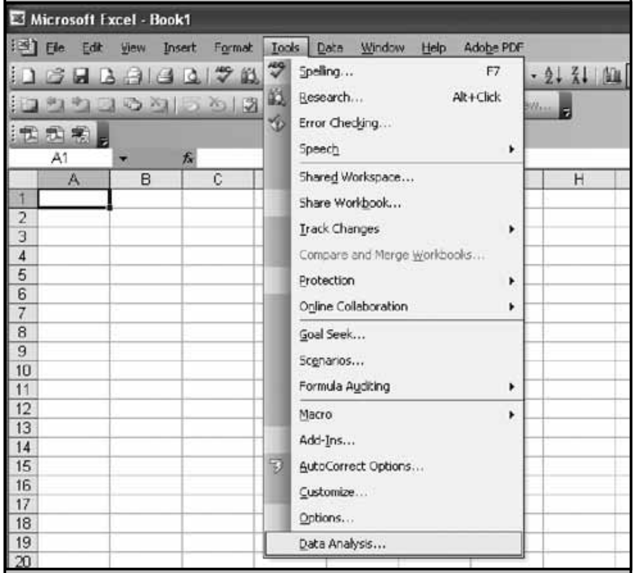
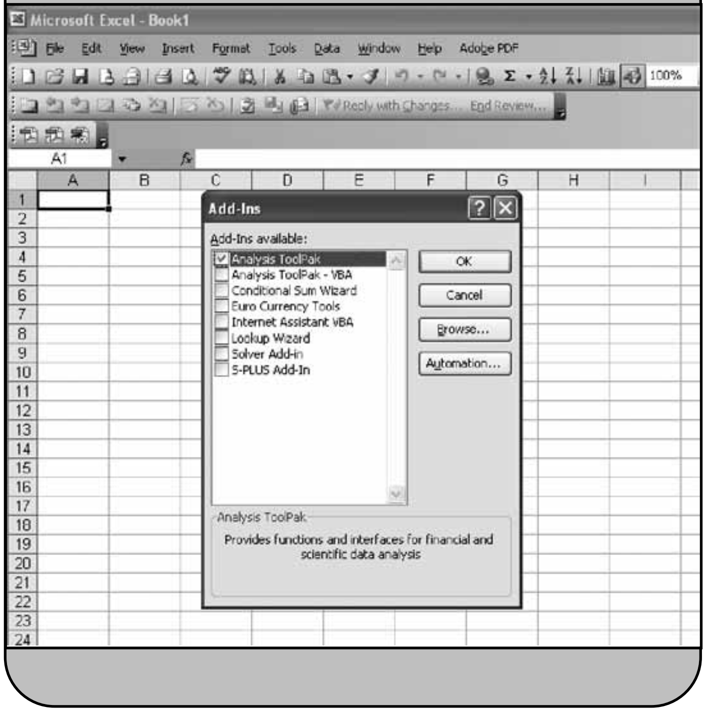


FIGURE 2–16 Adding the Data Analysis ToolPak


Microsoft Excel - Book1
File Edit View Insert Format Tools Data Window Help Adobe PDF
100%
Reply with Changes... End Review...
A1
A B C D E F G H I
Add-Ins
OK
Cancel
Browse...
Automation...

Add-Ins available:

☑ Analysis ToolPak
□ Analysis ToolPak - VBA
□ Conditional Sum Wizard
□ Euro Currency Tools
□ Internet Assistant VBA
□ Lookup Wizard
□ Solver Add-in
□ S-PLUS Add-In

Analysis ToolPak

Provides functions and interfaces for financial and scientific data analysis

TABLE 2–3 Analyses Available in the Data Analysis ToolPak

Analysis	Chapter in Textbook
Analysis of variance	7
Descriptive statistics	4, 6
Histogram	4
Regression	9
t Test: Paired two sample for means	7
t Test: Two sample assuming equal variances	7
z Test: Two sample for means	7

the ToolPak, Excel also offers many built-in statistical functions (e.g., CHITEST for a chi-square test) that can be used to perform specific tests and procedures. As we discuss specific analyses in Chapter 4 through Chapter 11 of the Excel workbook, we present options for analysis using the ToolPak and Excel's statistical functions.

2.5 PRACTICE PROBLEMS

1. Use Excel to create a worksheet, called *Data*, with the data in Table 2–4. The data were presented in Table 4–13 and were measured in a subsample of $n = 10$ participants who attended the seventh examination of the Framingham Offspring Study. Place the variable names in the first row of the worksheet.
2. Compute two new variables for each participant, body mass index (BMI) and mean arterial pressure (MAP). The formulas for the variables are:

$$\text{BMI} = \frac{\text{weight}_{\text{pounds}} \times 0.4536}{(\text{height}_{\text{inches}} \times 0.0254)^2}$$

$$\text{MAP} = \frac{(2 \times \text{Diastolic Blood Pressure}) + \text{Systolic Blood Pressure}}{3}$$

3. Compute the sample size for the MAP data in Problem 2 using the COUNT function and store the result in the *Data* worksheet.
4. Compute the mean MAP in Problem 2 by programming the formula for the mean (i.e., $\overline{X} = \frac{\Sigma X}{n}$) and store the result in the *Data* worksheet.
5. Compute the standard deviation of the MAP values in Problem 2 by programming the formula for the standard deviation (i.e., $s = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}}$) and store the result in the *Data* worksheet.

TABLE 2–4 Data for Practice Problems

Participant ID	Systolic Blood Pressure	Diastolic Blood Pressure	Total Serum Cholesterol	Weight (lb)	Height (in)
1	141	76	199	138	63.00
2	119	64	150	183	69.75
3	122	62	227	153	65.75
4	127	81	227	178	70.00
5	125	70	163	161	70.50
6	123	72	210	206	70.00
7	105	81	205	235	72.00
8	113	63	275	151	60.75
9	106	67	208	213	69.00
10	131	77	159	142	61.00

CHAPTER 3

Creating Tables and Graphs

Microsoft Office® Excel® is very useful for creating tables and graphs to summarize and present statistical information. Many investigators use Excel to prepare tables and graphs for reports, presentations, and manuscripts. Often, presentations and manuscripts are prepared in other packages (e.g., PowerPoint®, Word®) but tables and charts are prepared in Excel and then imported into those presentations or manuscripts.

3.1 CREATING AND FORMATTING TABLES

In Chapter 4 of the textbook, we presented a number of statistics to summarize continuous, ordinal, and categorical variables. Investigators must determine which statistics most accurately and completely describe sample data. For example, for continuous variables we can compute the sample mean, median, and mode to describe central tendency and the sample range, interquartile range, variance, and standard deviation to describe variability. For ordinal and categorical variables, we can compute frequencies, relative frequencies, and cumulative relative frequencies (appropriate for ordinal variables). For presentation purposes, we must decide which statistics to present and how.

In almost all research reports, investigators include a description of the study sample. The description usually includes socio-demographic or background characteristics (e.g., age, gender, educational level) and might include data to describe clinical history (e.g., prevalent disease, symptom severity at the start of the study). With a cross-sectional or cohort study, the description is often based on the full sample or cohort. In clinical trials, descriptions are usually provided for each treatment group, considered separately. Regardless of whether data are presented for a single group or for separate groups, the investigator must select the most appropriate statistics to summarize key information.

Example 3.1. Suppose we conduct a cross-sectional study of 125 undergraduate students to estimate the prevalence of cigarette smoking. Before presenting data on the prevalence of smoking, the investigators wish to provide a description of the study sample. Suppose that several background variables are analyzed (with Excel or with another statistical computing package). For each continuous variable, sample means and standard deviations are produced, and for each ordinal and categorical variable, frequencies and relative frequencies are produced. The results are shown in Table 3–1.

Table 3–2 was developed in Excel to present the information shown here. For the continuous variables, we present means and standard deviations rounded to one decimal place. For ordinal and categorical variables, we present the frequencies and relative frequencies. The table can be copied from Excel into a Word document for presentation.

It is always important to include a clear and concise title in a table. It is also important to specify clear variable names with appropriate units. Finally, the data (i.e., summary statistics) presented in the table must be clearly defined. Table 3–2 was prepared in Excel, as shown in Figure 3–1.

The title is entered in the first row of the table. To accommodate the length of the title and lengths of variable names, the widths of the columns (e.g., A, B) in the worksheet can be increased. It is not necessary to enlarge column A to accommodate all of the long title because the title is the only item in the first row. Thus, to accommodate the length of the title, we merge cell A1 and cell B1 into one larger cell. This is

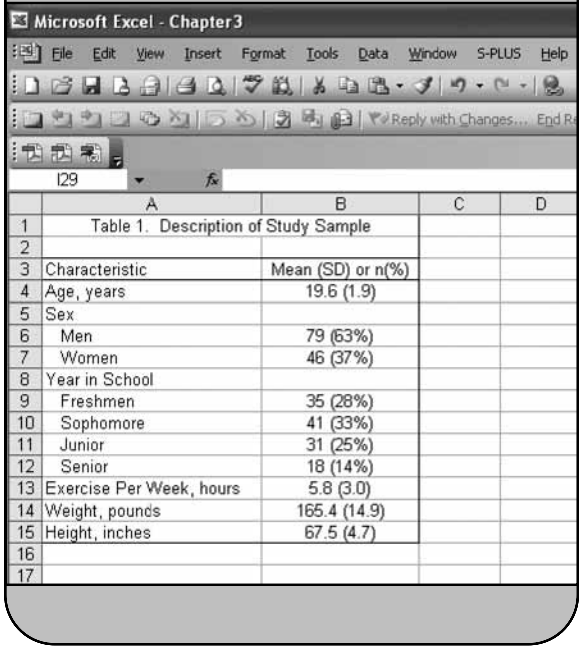
TABLE 3–1 Summary Statistics on Background Variables

Age	$\overline{X} = 19.567$		$s = 1.867$	
Sex	Men		Women	
	79 (63.2%)		46 (36.8%)	
Year in school	Freshmen	Sophomore	Junior	Senior
	35 (28.0%)	41 (32.8%)	31 (24.8%)	18 (14.4%)
Number of hours of exercise per week	$\overline{X} = 5.821$		$s = 2.989$	
Weight	$\overline{X} = 165.352$		$s = 14.857$	
Height	$\overline{X} = 67.463$		$s = 4.655$	

TABLE 3–2 Description of Study Sample

Characteristic	Mean (SD) or *n* (%)
Age (years)	19.6 (1.9)
Gender	
Men	79 (63%)
Women	46 (37%)
Year in School	
Freshmen	35 (28%)
Sophomore	41 (33%)
Junior	31 (25%)
Senior	18 (14%)
Exercise per week (h)	5.8 (3.0)
Weight (lb)	165.4 (14.9)
Height (in)	67.5 (4.7)

FIGURE 3–1 Description of Study Sample



	A	B	C	D
1	Table 1. Description of Study Sample			
2				
3	Characteristic	Mean (SD) or n(%)		
4	Age, years	19.6 (1.9)		
5	Sex			
6	Men	79 (63%)		
7	Women	46 (37%)		
8	Year in School			
9	Freshmen	35 (28%)		
10	Sophomore	41 (33%)		
11	Junior	31 (25%)		
12	Senior	18 (14%)		
13	Exercise Per Week, hours	5.8 (3.0)		
14	Weight, pounds	165.4 (14.9)		
15	Height, inches	67.5 (4.7)		
16				
17				

done by first entering the title into cell A1. Because the title is longer than the width of the A1 cell, the title runs across into cell B1 through cell D1, as shown in Figure 3–2.

Before we merge cells in the first row to accommodate the title, we first resize column A and column B to accommodate the variable names, units, and the descriptive statistics. To increase the width of column A, we place the cursor on the vertical line between column A and column B. The cursor changes shape to a bold cross with arrows running right and left. Clicking and holding the mouse while moving the cursor to the right or to the left then increases or decreases the width of column A. We do the same for column B, using the vertical line between column B and column C. Figure 3–3 shows the widened column A and column B.

We now merge cell A1 and cell B1 to accommodate the title. This is done by highlighting cell A1 and cell B1 and clicking on the **Merge and Center** icon shown on the menu bar (see Figure 3–3). The result is shown in Figure 3–4. We now enter the variable names, units, and the summary statistics into column A and column B.

Excel can be used to create tables for presentation purposes, and it offers options similar to those offered in Word for formatting text in terms of font size, type, justification, and so on. Excel also has options to round numeric information—for example, to round statistics to one or two decimal places.

FIGURE 3–2 Entering the Title for the Table

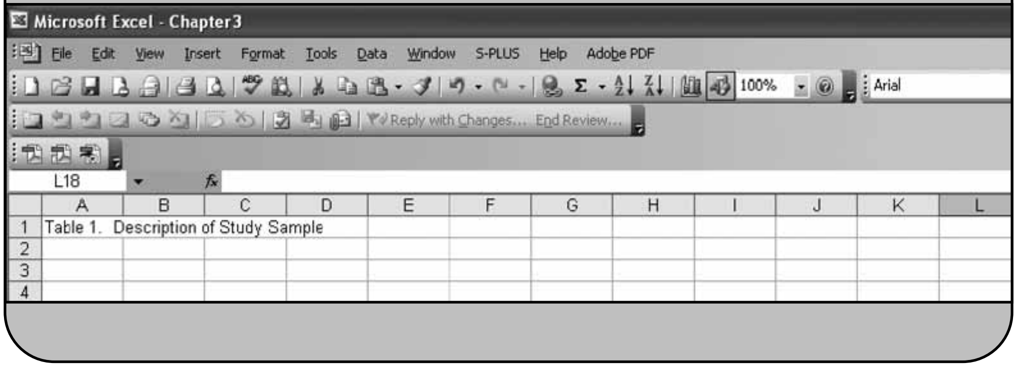

Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
L18
A B C D E F G H I J K L
1 2 3 4

Table 1. Description of Study Sample

FIGURE 3–3 Increasing the Widths of the Columns

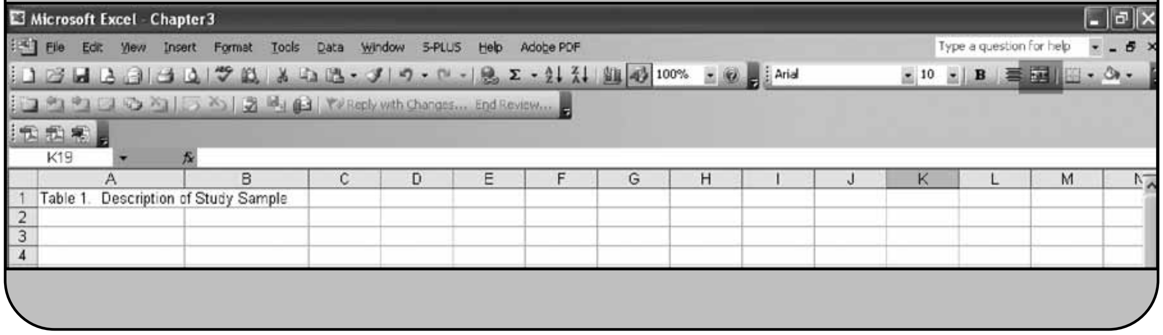
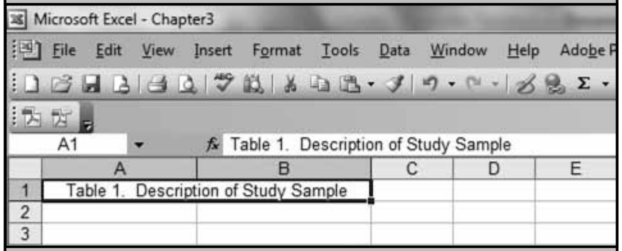

Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a question for help
100%
Arial
10
B
Reply with Changes... End Review...
K19
A B C D E F G H I J K L M N
1 2 3 4

Table 1. Description of Study Sample

FIGURE 3–4 Merging Cell A1 and Cell B1


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window Help Adobe P
A1
A B C D E
1 2 3

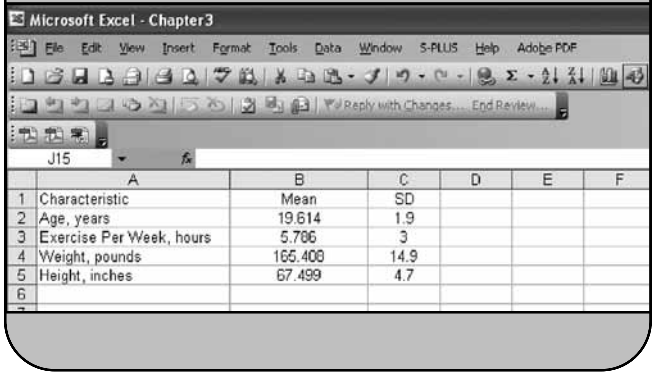
fx Table 1. Description of Study Sample

Table 1. Description of Study Sample

Suppose we compute the mean and standard deviation of several continuous variables as shown in Figure 3–5. For presentation, we wish to display the summary statistics to one decimal place. This can be done by first highlighting the desired range of cells (B2 through C5), and then selecting the **Format/Cells** option (Figure 3–6).

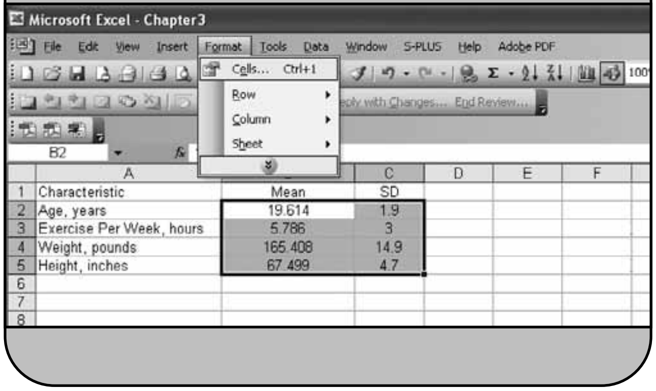
Once the **Format/Cells** menu option is selected, a dialog box appears. Various options are available for formatting. To format the numerical values to one decimal place, we select the "Number" category and then the desired number of decimal places. Here we choose 1 decimal place, as shown in Figure 3–7. Once we click **OK**, the data are formatted to one decimal place.

FIGURE 3–5 Summary Statistics for Continuous Variables


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...
J15

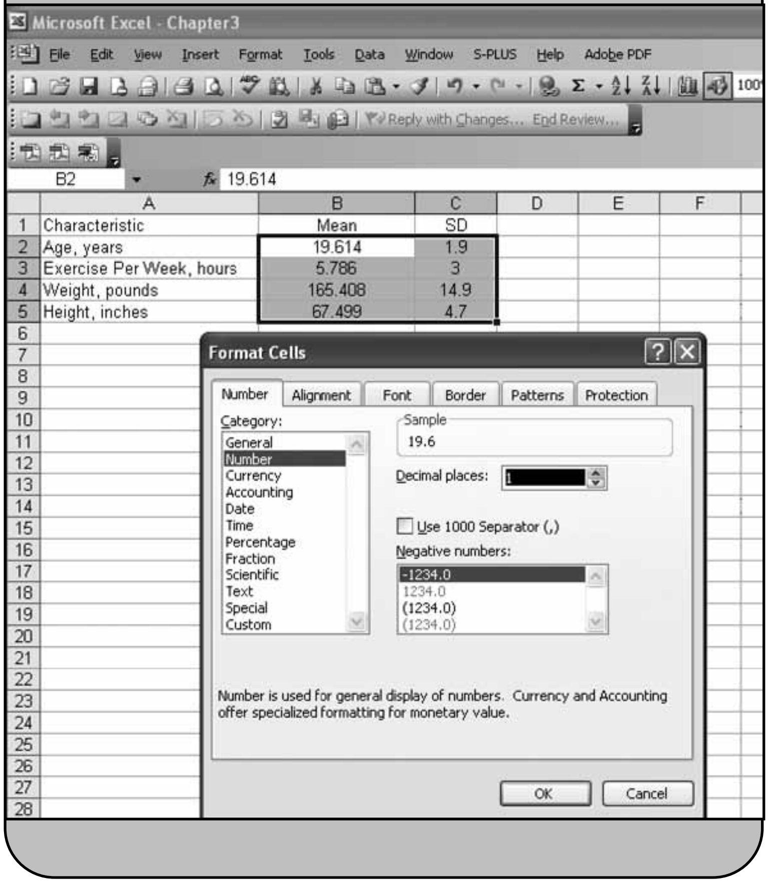
	A	B	C	D	E	F
1	Characteristic	Mean	SD			
2	Age, years	19.614	1.9			
3	Exercise Per Week, hours	5.786	3			
4	Weight, pounds	165.408	14.9			
5	Height, inches	67.499	4.7			
6						

FIGURE 3–6 Formatting Summary Statistics


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Cells... Ctrl+1
Row
Column
Sheet
Reply with Changes... End Review...
B2

	A	B	C	D	E	F
1	Characteristic	Mean	SD			
2	Age, years	19.614	1.9			
3	Exercise Per Week, hours	5.786	3			
4	Weight, pounds	165.408	14.9			
5	Height, inches	67.499	4.7			
6						
7						
8						

FIGURE 3–7 Formatting Numeric Data


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...
Format Cells
Number Alignment Font Border Patterns Protection
General
Number
Currency
Accounting
Date
Time
Percentage
Fraction
Scientific
Text
Special
Custom
-1234.0
1234.0
(1234.0)
(1234.0)
OK
Cancel

B2 = 19.614

	A	B	C	D	E	F
1	Characteristic	Mean	SD			
2	Age, years	19.614	1.9			
3	Exercise Per Week, hours	5.786	3			
4	Weight, pounds	165.408	14.9			
5	Height, inches	67.499	4.7			

Category: Number

Sample: 19.6

Decimal places: 1

□ Use 1000 Separator (,)

Negative numbers: -1234.0

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

3.2 FREQUENCY DISTRIBUTION TABLES

Excel has a built-in menu option that can be used to create frequency tables for presenting information. This is especially useful for ordinal and categorical variables. The option is illustrated below.

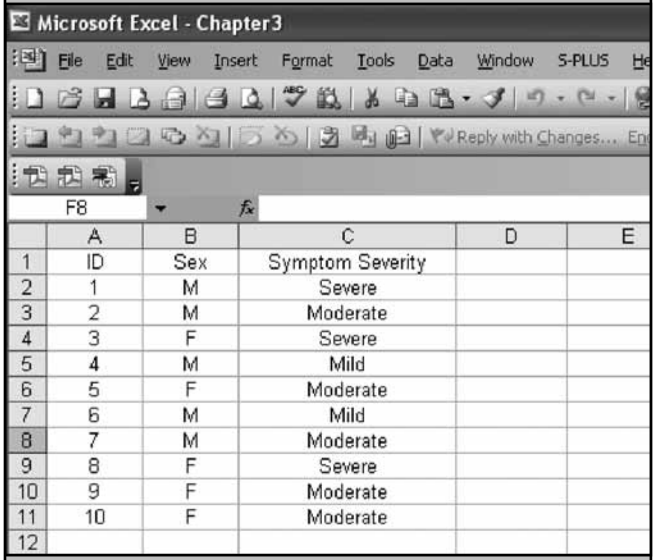
Example 3.2. Suppose we have a small study of $n = 10$ patients with rheumatoid arthritis, and we record their gender and the severity of their symptoms of arthritis. The data are shown in Figure 3–8. Suppose we want to present the gender distribution of the sample in a frequency distribution table. We first select the **Data** option on the main menu bar and then the **Pivot Table and Pivot Chart Report** option, as shown in Figure 3–9. Once we select this option, Excel opens the Pivot Table and Pivot Chart Wizard, as shown in Figure 3–10.

The wizard essentially asks for the details necessary to generate the frequency distribution table. The first detail concerns the data. Excel asks where the data reside, and the default response is a Microsoft Office Excel list or database, and this response is already checked. This response applies when the data are in an Excel worksheet, as is the case here. Thus, we do not need to modify the default response. We would choose an alternate response if, for example, the data were stored in a different file. The second detail concerns the type of report we wish to produce. The default response is checked and is a pivot table. This again is the appropriate response for a frequency distribution table. Once we click **Next**, we are presented with a second dialog box, as shown in Figure 3–11.

Excel then asks specifically for the range of cells containing the study data. We want to generate a frequency distribution table for gender, so we specify cell B1 through cell B11. Notice that we include cell B1, which actually contains the name of the variable as opposed to data (Figure 3–12).

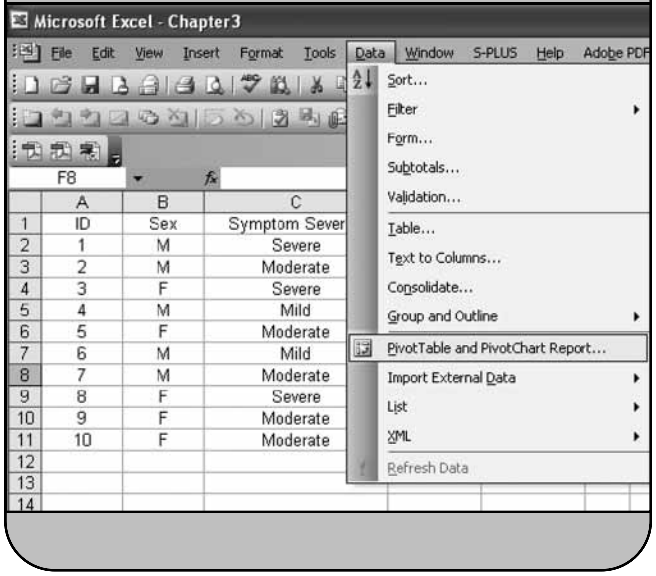
Once we click **Next**, we are presented with a third dialog box where Excel asks for the location for the resultant frequency distribution table (see Figure 3–13). We need only to specify the cell address for the top-left corner of the frequency distribution table. We request that Excel place the top-left corner of the frequency distribution table in cell E1. (Note that we could have checked the first option and requested that the frequency distribution table be placed in a different worksheet.) When we click **Finish**, Excel sets up a template in the current worksheet for the frequency distribution table. The template is shown in Figure 3–14. Notice that the top-left corner of the template is in cell E1.

FIGURE 3–8 Data on n=10 Patients with Rheumatoid Arthritis


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS He
Reply with Changes... En
F8
fx

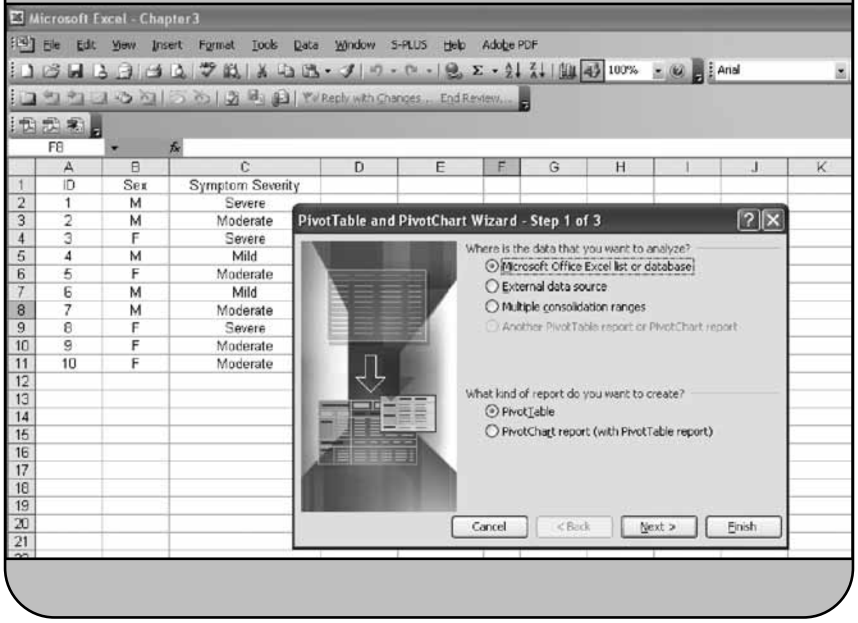
	A	B	C	D	E
1	ID	Sex	Symptom Severity		
2	1	M	Severe		
3	2	M	Moderate		
4	3	F	Severe		
5	4	M	Mild		
6	5	F	Moderate		
7	6	M	Mild		
8	7	M	Moderate		
9	8	F	Severe		
10	9	F	Moderate		
11	10	F	Moderate		
12					

FIGURE 3–9 Creating a Frequency Distribution Table



	A	B	C
1	ID	Sex	Symptom Sever
2	1	M	Severe
3	2	M	Moderate
4	3	F	Severe
5	4	M	Mild
6	5	F	Moderate
7	6	M	Mild
8	7	M	Moderate
9	8	F	Severe
10	9	F	Moderate
11	10	F	Moderate
12			
13			
14			

FIGURE 3–10 Creating a Frequency Distribution Table


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
F8
Cancel
< Back
Next >
Finish

ID	Sex	Symptom Severity
1	M	Severe
2	M	Moderate
3	F	Severe
4	M	Mild
5	F	Moderate
6	M	Mild
7	M	Moderate
8	F	Severe
9	F	Moderate
10	F	Moderate

PivotTable and PivotChart Wizard - Step 1 of 3

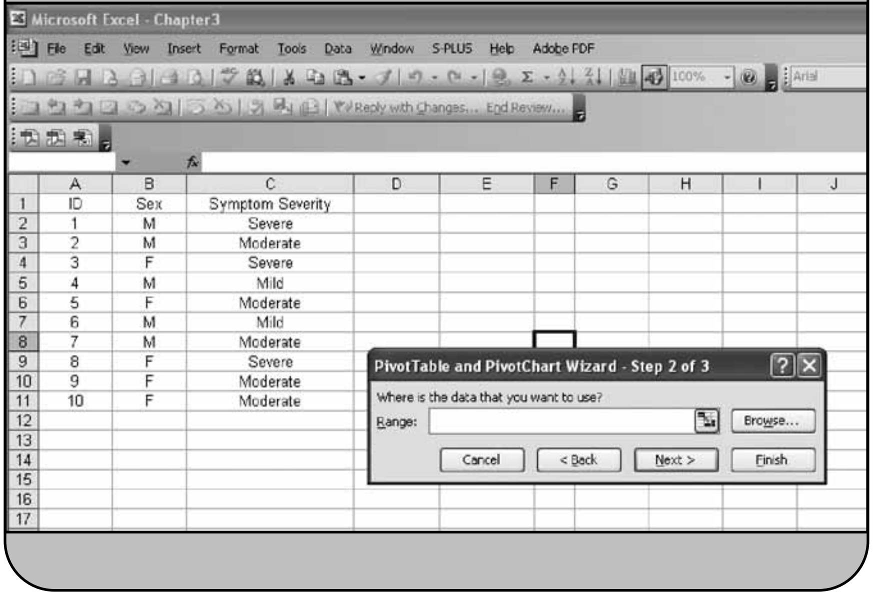
Where is the data that you want to analyze?

● Microsoft Office Excel list or database
○ External data source
○ Multiple consolidation ranges
○ Another PivotTable report or PivotChart report

What kind of report do you want to create?

● PivotTable
○ PivotChart report (with PivotTable report)

FIGURE 3–11 Creating a Frequency Distribution Table


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
Browse...
Cancel
< Back
Next >
Finish

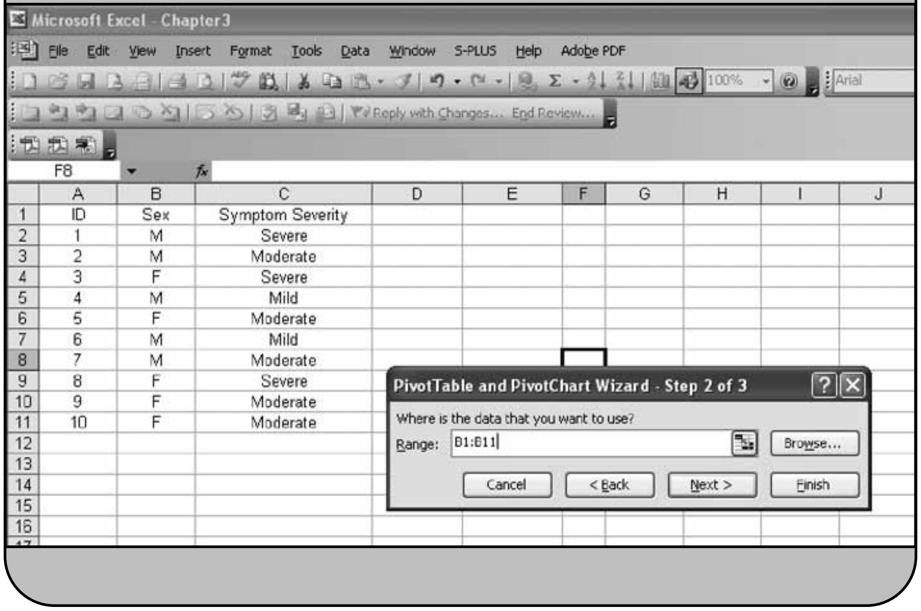
ID	Sex	Symptom Severity
1	M	Severe
2	M	Moderate
3	F	Severe
4	M	Mild
5	F	Moderate
6	M	Mild
7	M	Moderate
8	F	Severe
9	F	Moderate
10	F	Moderate

PivotTable and PivotChart Wizard - Step 2 of 3

Where is the data that you want to use?

Range:

FIGURE 3–12 Creating a Frequency Distribution Table


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
F8
fx

	A	B	C	D	E	F	G	H	I	J
1	ID	Sex	Symptom Severity							
2	1	M	Severe							
3	2	M	Moderate							
4	3	F	Severe							
5	4	M	Mild							
6	5	F	Moderate							
7	6	M	Mild							
8	7	M	Moderate							
9	8	F	Severe							
10	9	F	Moderate							
11	10	F	Moderate							

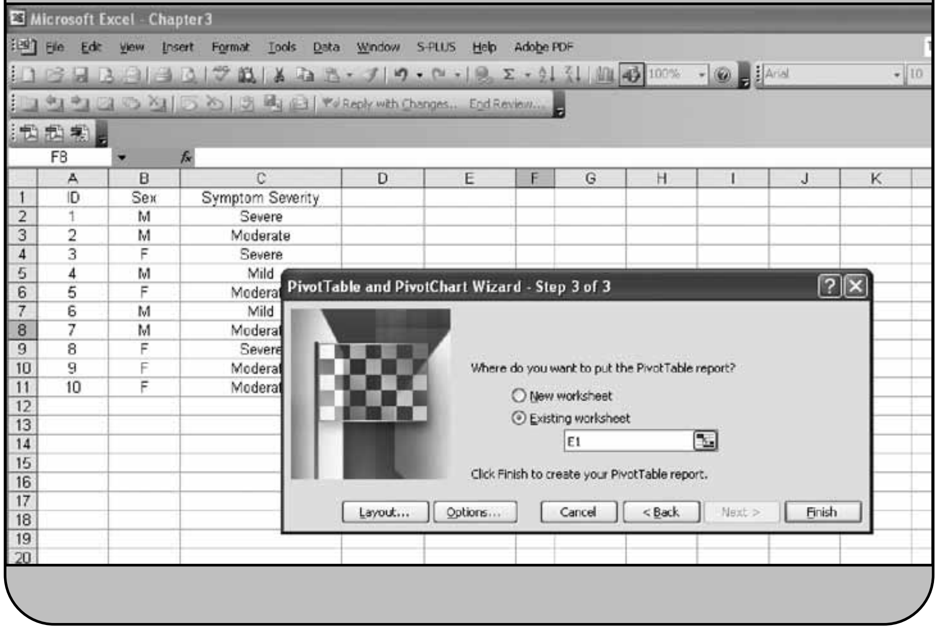
PivotTable and PivotChart Wizard - Step 2 of 3

Where is the data that you want to use?

Range: B1:B11

Browse... Cancel < Back Next > Finish

FIGURE 3–13 Creating a Frequency Distribution Table


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
10
Reply with Changes... End Review...
F8
fx

	A	B	C	D	E	F	G	H	I	J	K
1	ID	Sex	Symptom Severity								
2	1	M	Severe								
3	2	M	Moderate								
4	3	F	Severe								
5	4	M	Mild								
6	5	F	Moderat								
7	6	M	Mild								
8	7	M	Moderat								
9	8	F	Severe								
10	9	F	Moderat								
11	10	F	Moderat								

PivotTable and PivotChart Wizard - Step 3 of 3

Where do you want to put the PivotTable report?

○ New worksheet

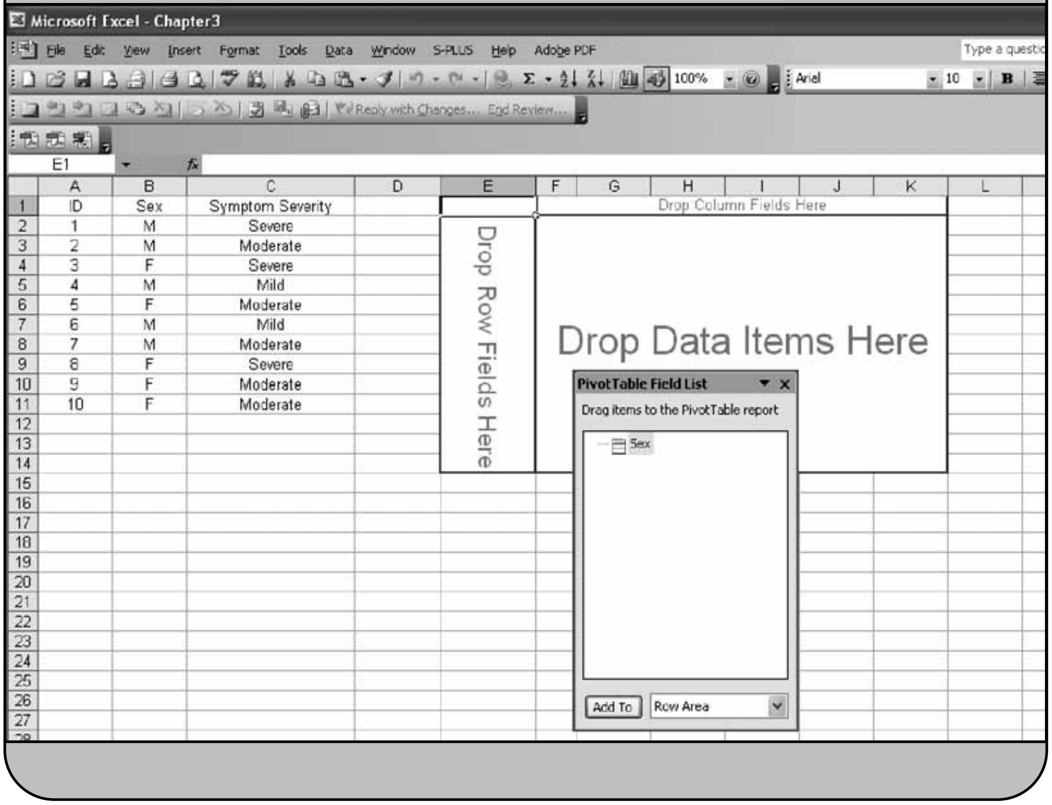
● Existing worksheet

E1

Click Finish to create your PivotTable report.

Layout... Options... Cancel < Back Next > Finish

FIGURE 3–14 Creating a Frequency Distribution Table


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a questio
100%
Arial
10
Reply with Changes... End Review...
E1
Drop Column Fields Here
Drop Row Fields Here
Drop Data Items Here
PivotTable Field List
Drag items to the PivotTable report
Sex
Add To
Row Area

ID	Sex	Symptom Severity
1	M	Severe
2	M	Moderate
3	F	Severe
4	M	Mild
5	F	Moderate
6	M	Mild
7	M	Moderate
8	F	Severe
9	F	Moderate
10	F	Moderate

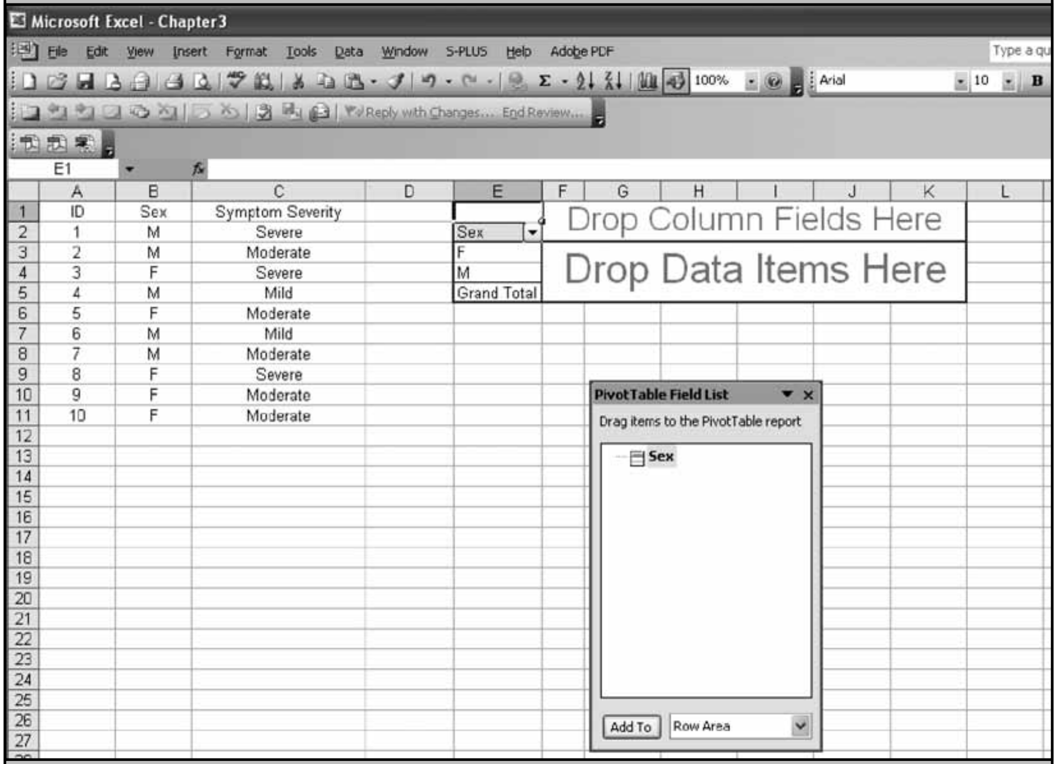
In addition to the template (shown in the background), Excel also displays the variable we specified for the analysis (in this case, sex) in the pivot table field list. To produce the frequency distribution table, we must specify that sex is the row variable for the table and, in addition, that sex is the data variable. We first specify that sex is the row variable by clicking the Add To button in the dialog box. The sex variable is specified in the list and we are requesting that Excel use sex in the row area. Once we click **Add To**, Excel automatically makes sex the row variable in the frequency distribution table, as shown in Figure 3–15.

We now need to specify that sex is also the data variable. The data variable is the variable that will be summarized. For a frequency distribution table on one variable, the row and data variable are the same. In other instances, it may be of interest to summarize a second variable (e.g., compute frequencies of symptom severity by sex, in which case the data variable would be different from the row variable). This is done by selecting **Data Area** from the drop-down list at the bottom of the dialog box and clicking **Add To**. This is shown in Figure 3–16. Once we add sex to the data area, the frequency distribution table is generated by Excel (Figure 3–17).

Using the same sequence of steps, we can also generate a frequency distribution table for symptom severity. The results are shown in Figure 3–18. In the specifications, we requested that Excel place the top-left corner of the frequency distribution table in cell E7.

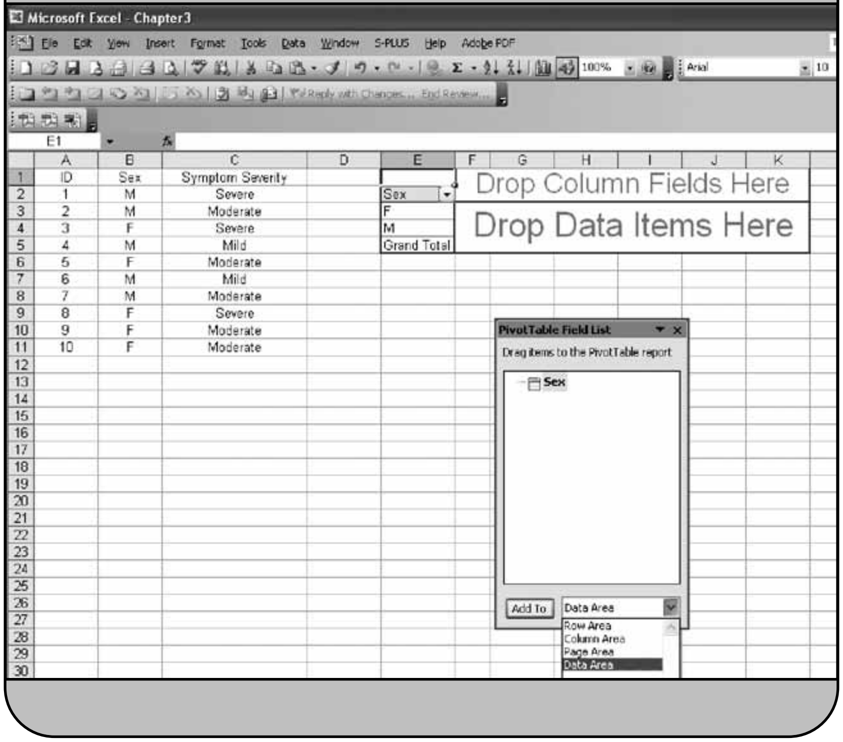
Suppose we also want to present relative frequencies. The relative frequencies of females and males are computed by entering "=F3/$F$5" and "=F4/$F$5" into cell G3 and cell G4, respectively. The same can be done for the symptom severity data using "=F9/$F$12" in cell G9 and copying to cell G10 and cell G11. The results are shown in Figure 3–19.

FIGURE 3–15 Specifying the Variable for the Frequency Distribution Table


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a qu
100%
Arial
10
B
Reply with Changes... End Review...
E1
Drop Column Fields Here
Drop Data Items Here
PivotTable Field List
Drag items to the PivotTable report
Sex
Add To
Row Area

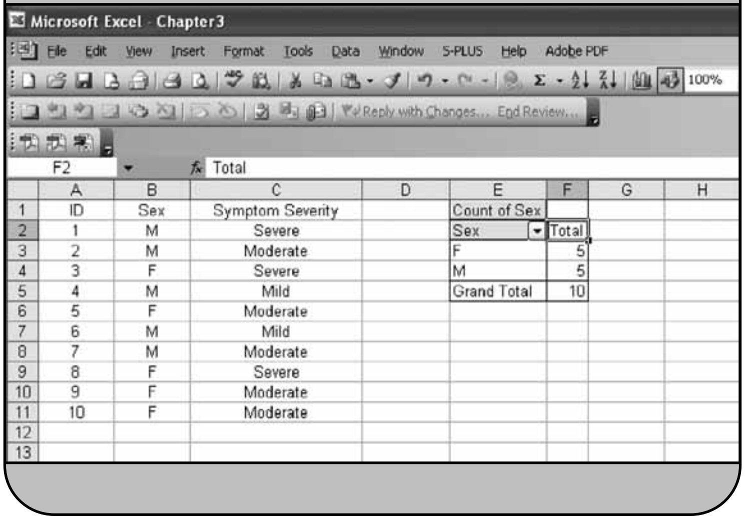
	A	B	C	D	E
1	ID	Sex	Symptom Severity		
2	1	M	Severe		Sex
3	2	M	Moderate		F
4	3	F	Severe		M
5	4	M	Mild		Grand Total
6	5	F	Moderate		
7	6	M	Mild		
8	7	M	Moderate		
9	8	F	Severe		
10	9	F	Moderate		
11	10	F	Moderate		

FIGURE 3–16 Selecting the Data Variable


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
10
Reply with Changes... End Review...
E1
Drop Column Fields Here
Drop Data Items Here
PivotTable Field List
Drag items to the PivotTable report
Sex
Add To
Data Area
Row Area
Column Area
Page Area
Data Area

	A	B	C	D	E
1	ID	Sex	Symptom Severity		
2	1	M	Severe		Sex
3	2	M	Moderate		F
4	3	F	Severe		M
5	4	M	Mild		Grand Total
6	5	F	Moderate		
7	6	M	Mild		
8	7	M	Moderate		
9	8	F	Severe		
10	9	F	Moderate		
11	10	F	Moderate		

FIGURE 3–17 Frequency Distribution Table for Sex


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

F2 *fx* Total

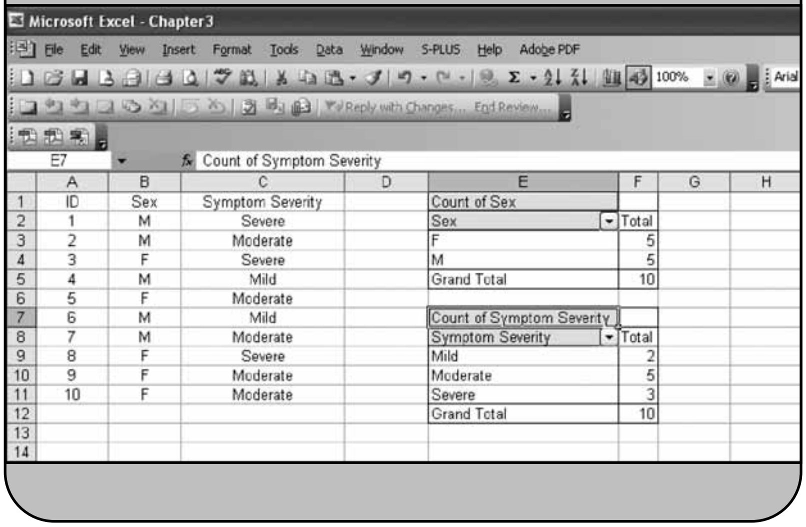
	A	B	C	D	E	F	G	H
1	ID	Sex	Symptom Severity		Count of Sex			
2	1	M	Severe		Sex	Total		
3	2	M	Moderate		F	5		
4	3	F	Severe		M	5		
5	4	M	Mild		Grand Total	10		
6	5	F	Moderate					
7	6	M	Mild					
8	7	M	Moderate					
9	8	F	Severe					
10	9	F	Moderate					
11	10	F	Moderate					
12								
13								

3.3 HISTOGRAMS AND BAR CHARTS

Excel is very powerful for generating graphical displays. Investigators often run statistical analyses in other packages, such as SAS®, and then use Excel to produce graphical displays of the statistical results. There are several ways to generate histograms and bar charts for ordinal and categorical variables, respectively. We describe two techniques. The first is a follow-on to the **Data/Pivot Table and Pivot Chart Report** menu option, and the second uses the Excel Chart Wizard.

Similar to the wizard we used to create frequency distribution tables in the preceding section, Excel has a Chart Wizard that is very useful for generating graphical displays. The Chart Wizard can be accessed through the graphic icon on the main toolbar, as shown in Figure 3–20. Clicking on the Chart Wizard opens the dialog box shown in Figure 3–20, which offers various options for graphical displays. Excel offers a number of standard graphical displays as well as some custom displays (under the respective tabs). We first illustrate how to produce

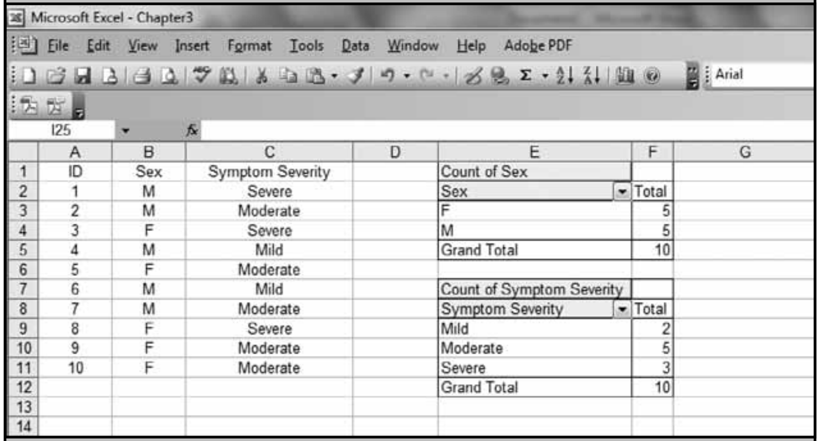
FIGURE 3–18 Frequency Distribution Table for Symptom Severity


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...

E7 *fx* Count of Symptom Severity

	A	B	C	D	E	F	G	H
1	ID	Sex	Symptom Severity		Count of Sex			
2	1	M	Severe		Sex	Total		
3	2	M	Moderate		F	5		
4	3	F	Severe		M	5		
5	4	M	Mild		Grand Total	10		
6	5	F	Moderate					
7	6	M	Mild		Count of Symptom Severity			
8	7	M	Moderate		Symptom Severity	Total		
9	8	F	Severe		Mild	2		
10	9	F	Moderate		Moderate	5		
11	10	F	Moderate		Severe	3		
12					Grand Total	10		
13								
14								

FIGURE 3–19 Adding Relative Frequencies



Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
I25

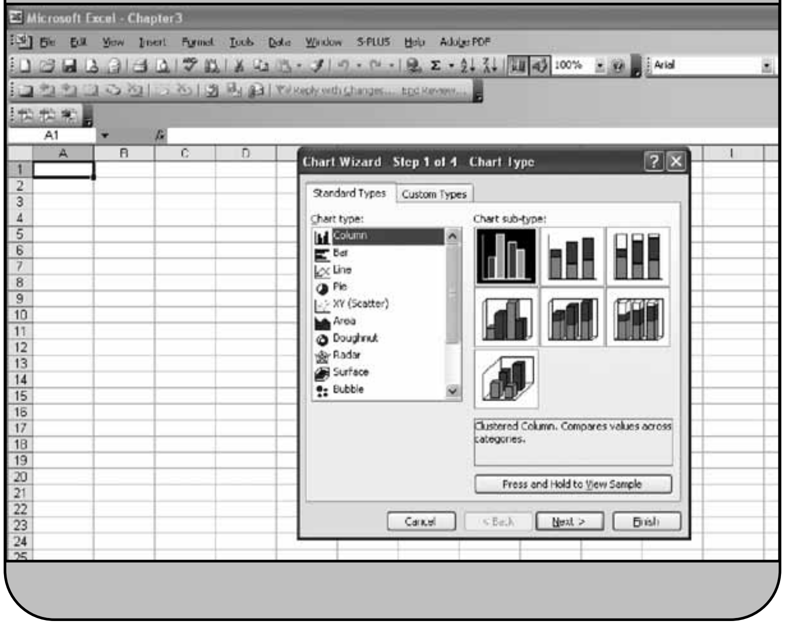
	A	B	C	D	E	F	G
1	ID	Sex	Symptom Severity		Count of Sex		
2	1	M	Severe		Sex	Total	
3	2	M	Moderate		F	5	
4	3	F	Severe		M	5	
5	4	M	Mild		Grand Total	10	
6	5	F	Moderate				
7	6	M	Mild		Count of Symptom Severity		
8	7	M	Moderate		Symptom Severity	Total	
9	8	F	Severe		Mild	2	
10	9	F	Moderate		Moderate	5	
11	10	F	Moderate		Severe	3	
12					Grand Total	10	
13							
14							

a graphical display following the **Data/Pivot Table and Pivot Chart Report** option, and then using the Chart Wizard.

In Example 3.1 of the Excel workbook, we used the **Data/Pivot Table and Pivot Chart Report** to generate frequency distribution tables for participant's sex and symptom severity. Once a frequency distribution table is produced, we can easily generate a graphical display. Suppose we want to generate a histogram for the distribution of symptom severity. This is done by selecting (highlighting) the frequency distribution for symptom severity and then clicking on the Chart Wizard (Figure 3–21). Once we click on the Chart Wizard, a new worksheet is generated, as shown in Figure 3–22. The new worksheet is called *Chart1.* (This is the default name, but it can be changed.)

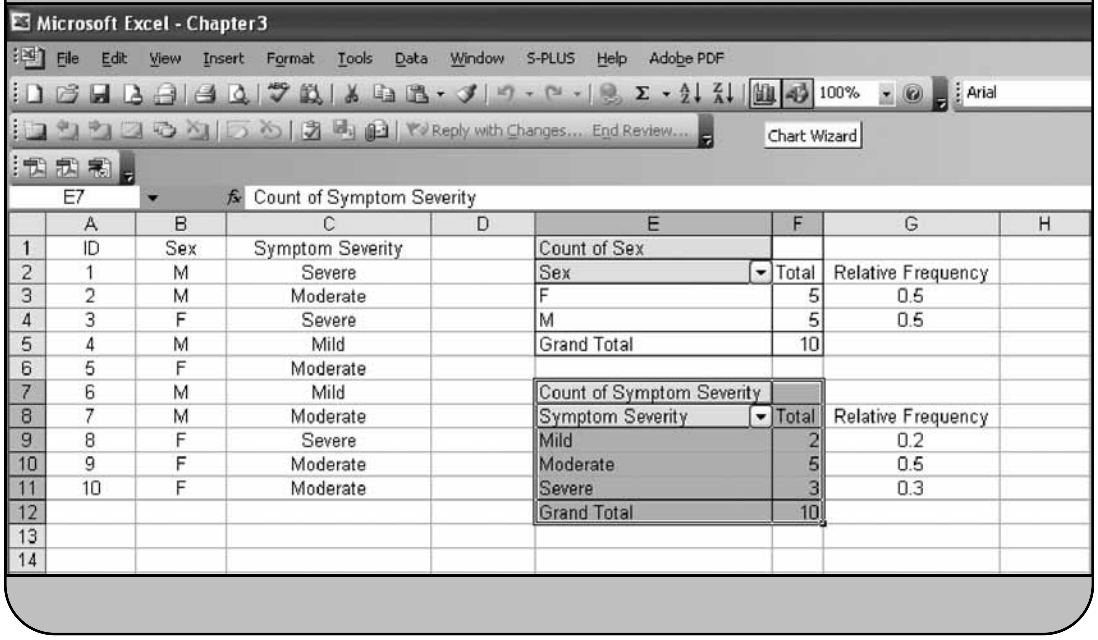
There are a number of options available for formatting the display. First, we can hide some of the templates and labels that Excel has placed on the display. This can be done by

FIGURE 3–20 The Chart Wizard



Clustered Column. Compares values across categories.

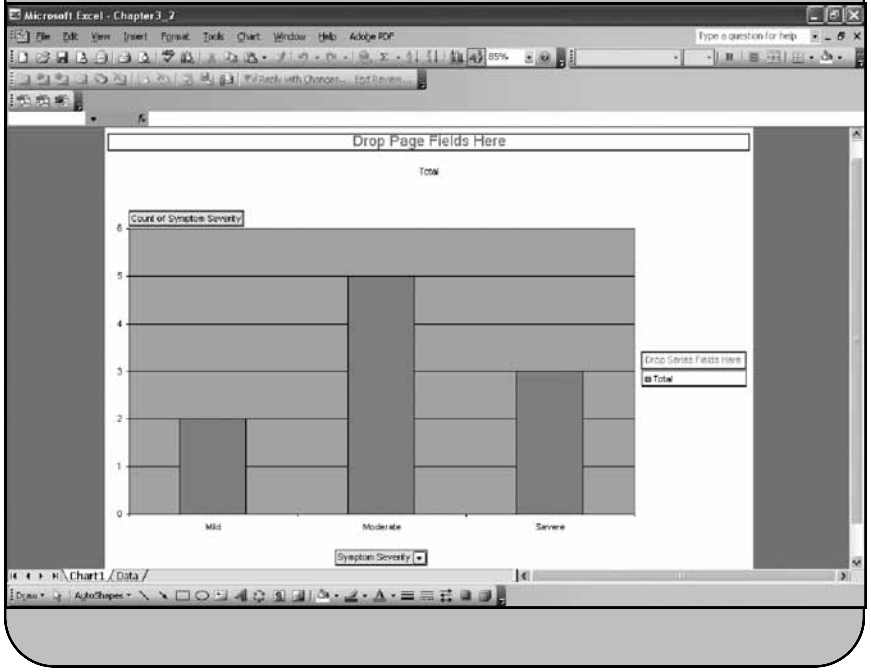
FIGURE 3–21 Creating a Histogram with the Chart Wizard


Microsoft Excel - Chapter3
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
Chart Wizard

E7 Count of Symptom Severity

	A	B	C	D	E	F	G	H
1	ID	Sex	Symptom Severity		Count of Sex			
2	1	M	Severe		Sex	Total	Relative Frequency	
3	2	M	Moderate		F	5	0.5	
4	3	F	Severe		M	5	0.5	
5	4	M	Mild		Grand Total	10		
6	5	F	Moderate					
7	6	M	Mild		Count of Symptom Severity			
8	7	M	Moderate		Symptom Severity	Total	Relative Frequency	
9	8	F	Severe		Mild	2	0.2	
10	9	F	Moderate		Moderate	5	0.5	
11	10	F	Moderate		Severe	3	0.3	
12					Grand Total	10		
13								
14								

FIGURE 3–22 Display with Dafault Options


Microsoft Excel - Chapter3_2
File Edit View Insert Format Tools Chart Window Help Adobe PDF
Type a question for help
85%
Reply with Changes... End Review...
Drop Page Fields Here
Total
Count of Symptom Severity
6
5
4
3
2
1
0
Mild
Moderate
Severe
Drop Series Fields Here
Total
Symptom Severity
Chart1
Data
Draw
AutoShapes

placing the cursor over the label, "Count of Symptom Severity." If we right-click while over this label, several options are available (Figure 3–23). When we click on **Hide Pivot Chart Field Buttons**, the field labels are removed, producing the display shown in Figure 3–24.

Notice in Figure 3–24 that we also changed the title. This is done by double-clicking on the default title ("Total"); when the default title is highlighted, we enter a new title ("Frequency Histogram of Symptom Severity"). By default, Excel generates a bar chart. Because symptom severity is an ordinal variable

FIGURE 3–23 Formatting Options

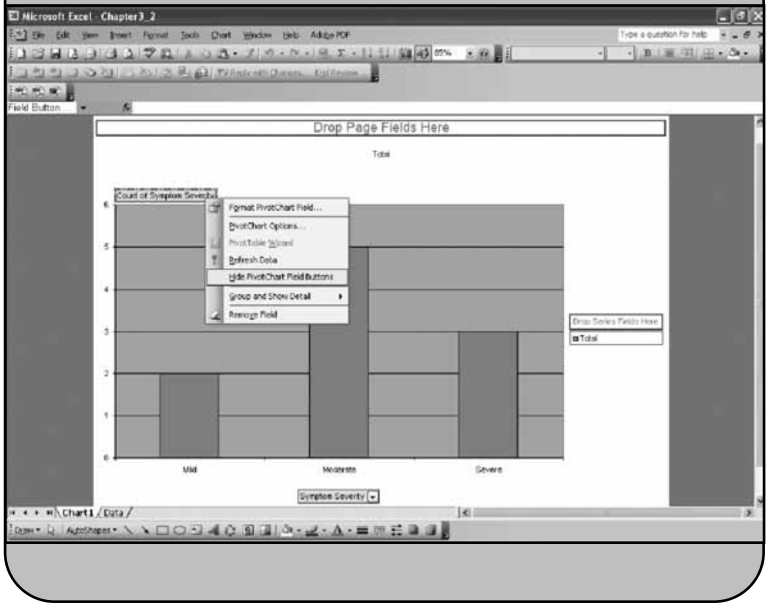
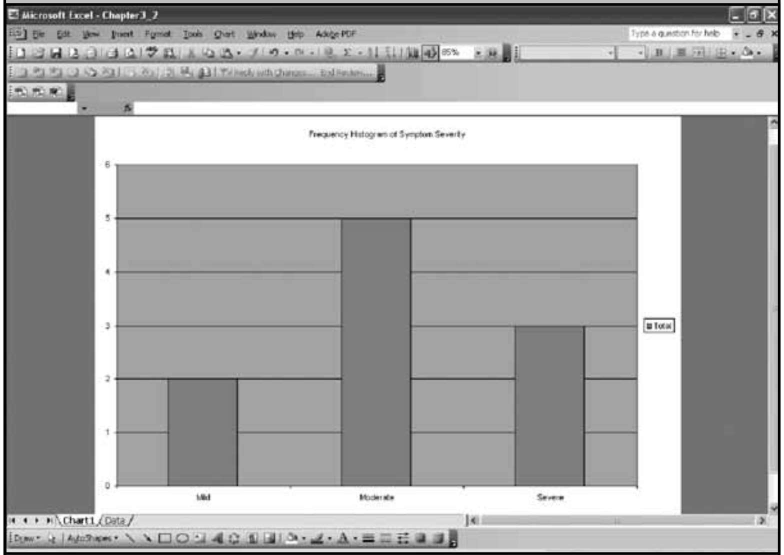


FIGURE 3–24 Removing Field Labels


Microsoft Excel - Chapter3_2
File Edit View Insert Format Tools Chart Window Help Adobe PDF
Type a question for help
Frequency Histogram of Symptom Severity
0
1
2
3
4
5
6
Mild
Moderate
Severe
Total
Chart1
Data
Draw
AutoShapes

and not a categorical variable, we want to display a histogram (i.e., the adjacent bars should run together). To change the bar chart into a histogram, we double-click on one of the three bars (double-clicking on any of the three will produce the same result). Double-clicking opens the menu of formatting options shown in Figure 3–25.

To convert the bar chart to a histogram, we select the **Options** tab, shown in Figure 3–26. The default gap width between bars is 150. To produce a histogram, we change the gap width to 0 (Figure 3–27). Once we click **OK**, Excel generates the histogram shown in Figure 3–28.

Excel offers many options for formatting graphical displays. We describe only a few here. A second method for generating graphical displays uses the Chart Wizard directly (i.e., does not involve first using the **Data/Pivot Table and Pivot Chart Report**).

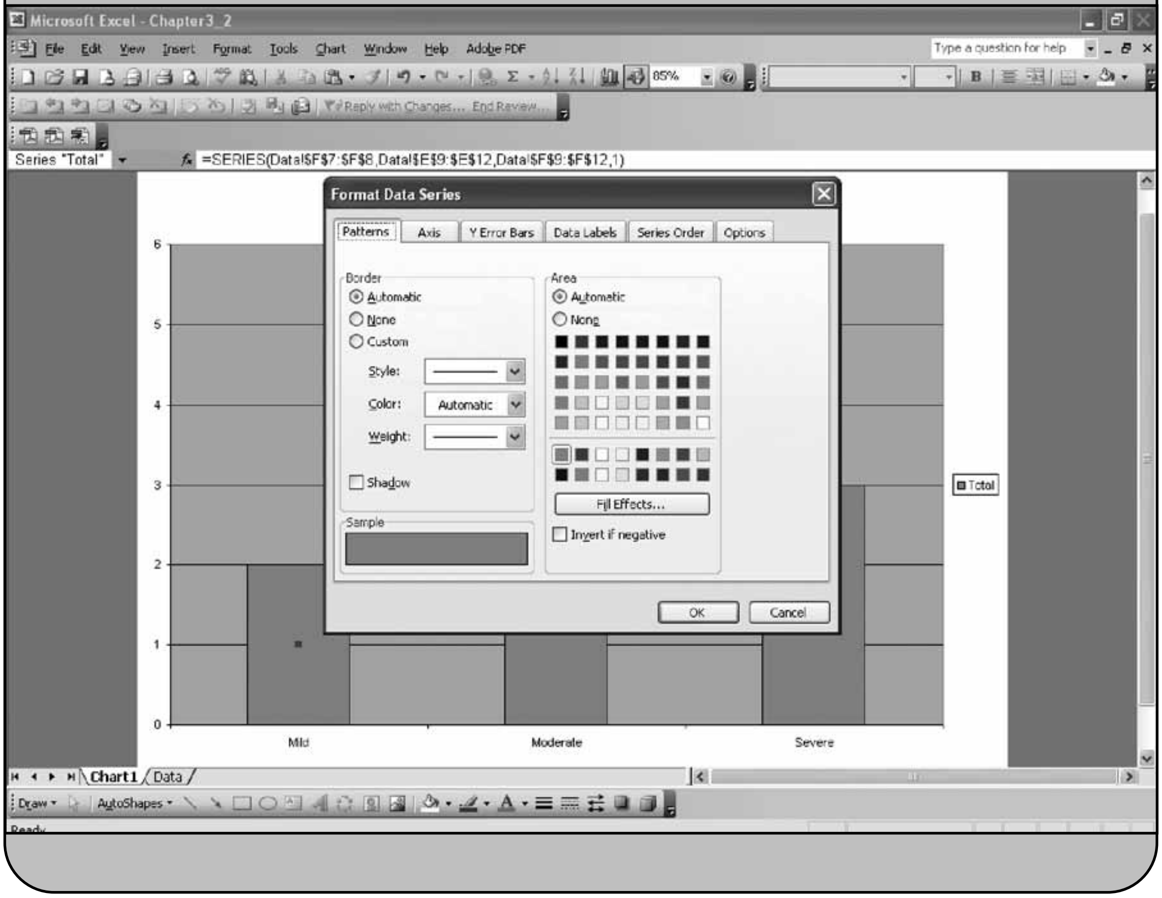
Example 3.3. In Example 4.2 in the textbook, we presented data from the seventh examination of the offspring in the Framingham Heart Study on blood pressure ($n = 3539$). Systolic and diastolic blood pressures were measured as continuous variables and organized into ordinal categories. Table 4–5 in the textbook showed a frequency distribution table for the ordinal blood pressure variable and is shown here as Table 3–3.

We want to generate a relative frequency histogram to present the blood pressure data. We first enter the information shown in Table 3–3 into Excel. Figure 3–29 contains the

TABLE 3–3 Frequency Distribution Table for Blood Pressure Categories

Blood Pressure	Frequency	Relative Frequency (%)
Normal	1206	34.1
Pre-hypertension	1452	41.1
Stage I hypertension	653	18.5
Stage II hypertension	222	6.3
Total	3533	100.0

FIGURE 3–25 Formatting the Bars



Series "Total"

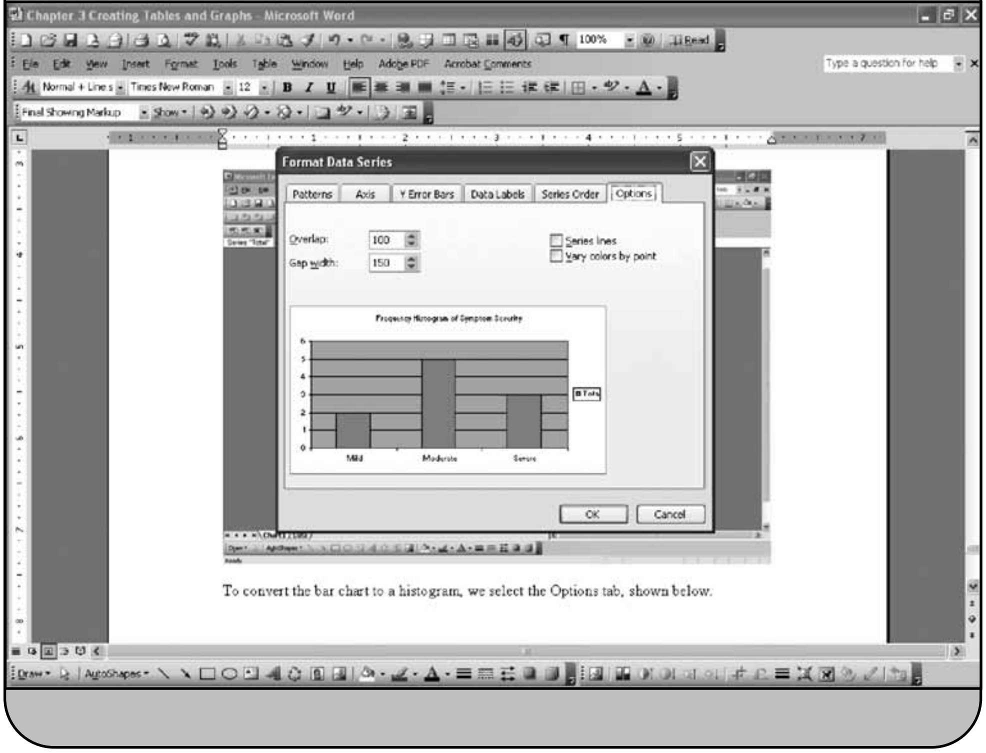
=SERIES(Data!$F$7:$F$8,Data!$E$9:$E$12,Data!$F$9:$F$12,1)

Border
● Automatic
○ None
○ Custom
Style:
Color: Automatic
Weight:
□ Shadow

Sample

Area
● Automatic
○ None
□ Invert if negative

FIGURE 3–26 Reducing the Gap Between Bars



Chapter 3 Creating Tables and Graphs - Microsoft Word

Format Data Series

Overlap: 100

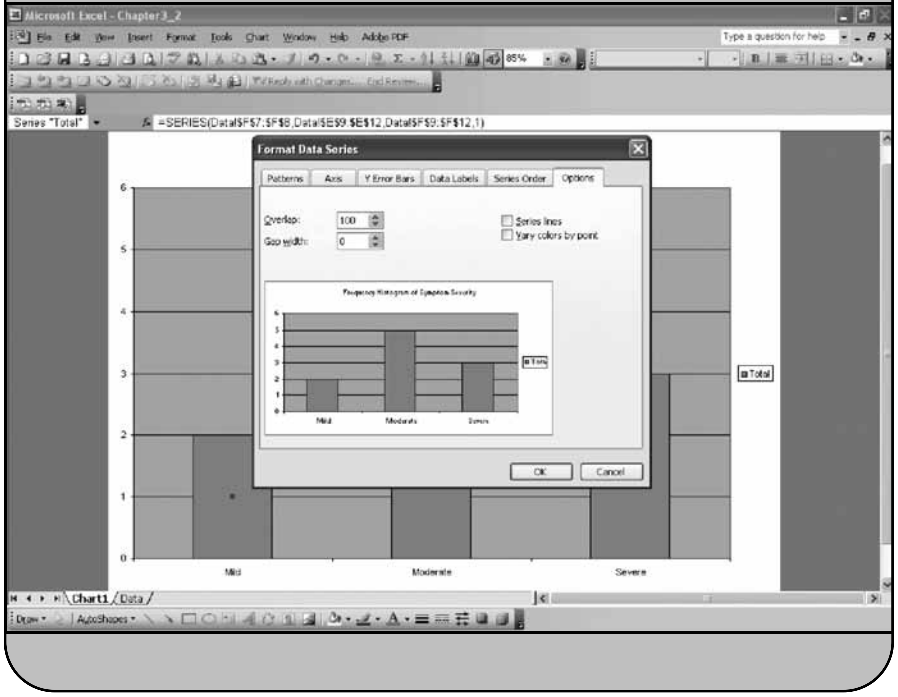
Gap width: 150

□ Series lines

□ Vary colors by point

To convert the bar chart to a histogram, we select the Options tab, shown below.

FIGURE 3–27 Setting the Gap Width to Zero



Microsoft Excel - Chapter3_2

Series "Total"

=SERIES(Data!$F$7:$F$8,Data!$E$9:$E$12,Data!$F$9:$F$12,1)

Format Data Series

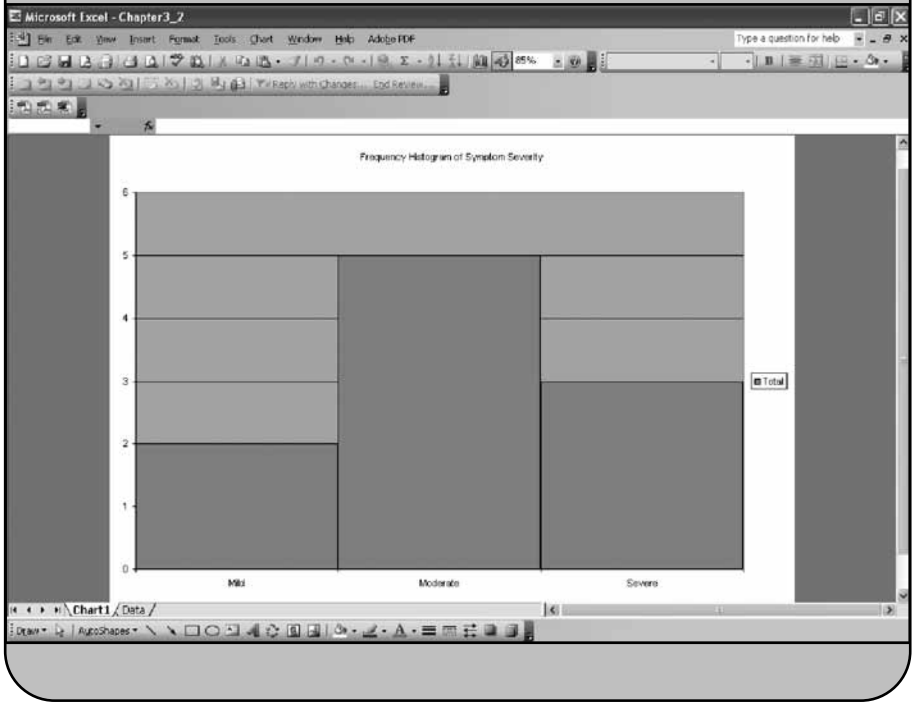
Overlap: 100

Gap width: 0

□ Series lines

□ Vary colors by point

FIGURE 3–28 Frequency Histogram for Symptom Severity

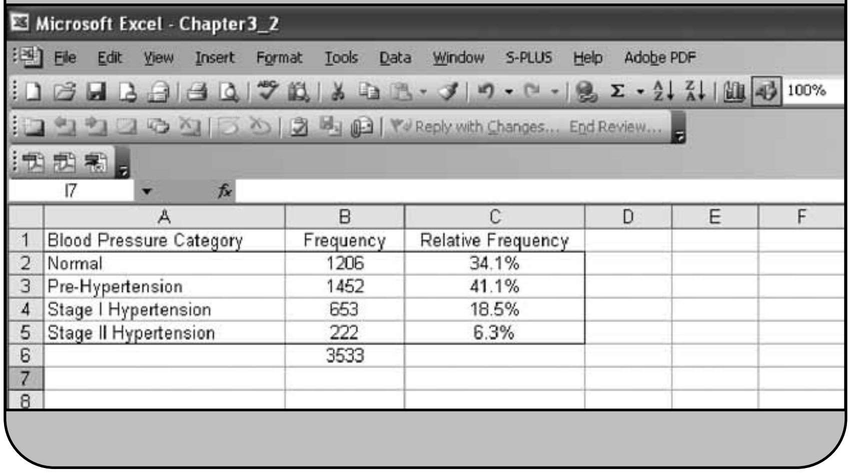

Microsoft Excel - Chapter3_2
File Edit View Insert Format Tools Chart Window Help Adobe PDF
Type a question for help
85%
Reply with Changes... End Review...
Frequency Histogram of Symptom Severity
6
5
4
3
2
1
0
Mild
Moderate
Severe
Total
Chart1
Data
Draw
AutoShapes

data. To generate the relative frequency histogram, we click the Chart Wizard icon on the menu bar. This opens the dialog box with various options shown in Figure 3–30.

The first option is a column chart, and Excel can generate various forms of this display (e.g., one-dimensional, three-dimensional). Suppose we click the top-left option (the default option shown in black in Figure 3–30) under **Chart Subtype**. Once we click **Next**, Excel then asks for the range of the data for the display (Figure 3–31).

We specify both the location of the response labels (to label the bars) and the location of the relative frequencies. The response labels are in column A and the relative frequencies are

FIGURE 3–29 Blood Pressure Data


Microsoft Excel - Chapter3_2
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...
I7

	A	B	C	D	E	F
1	Blood Pressure Category	Frequency	Relative Frequency			
2	Normal	1206	34.1%			
3	Pre-Hypertension	1452	41.1%			
4	Stage I Hypertension	653	18.5%			
5	Stage II Hypertension	222	6.3%			
6		3533				
7						
8						

FIGURE 3–30 Using the Chart Wizard

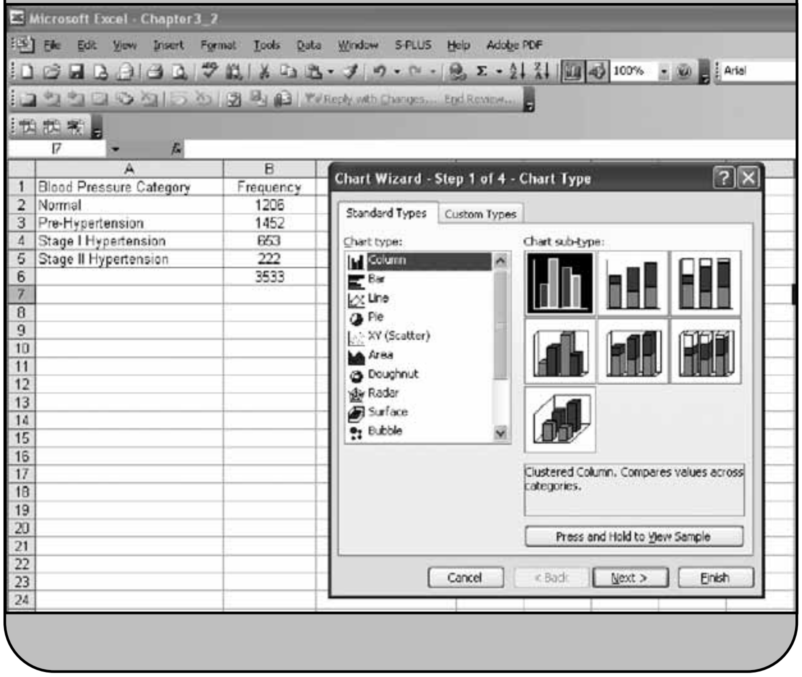

Microsoft Excel - Chapter3_2
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
I7
Standard Types
Custom Types
Chart type:
Column
Bar
Line
Pie
XY (Scatter)
Area
Doughnut
Radar
Surface
Bubble
Chart sub-type:
Press and Hold to View Sample
Cancel
< Back
Next >
Finish

Chart Wizard - Step 1 of 4 - Chart Type

	A	B
1	Blood Pressure Category	Frequency
2	Normal	1206
3	Pre-Hypertension	1452
4	Stage I Hypertension	653
5	Stage II Hypertension	222
6		3533

Clustered Column. Compares values across categories.

FIGURE 3–31 Chart Wizard Prompt for the Data Range

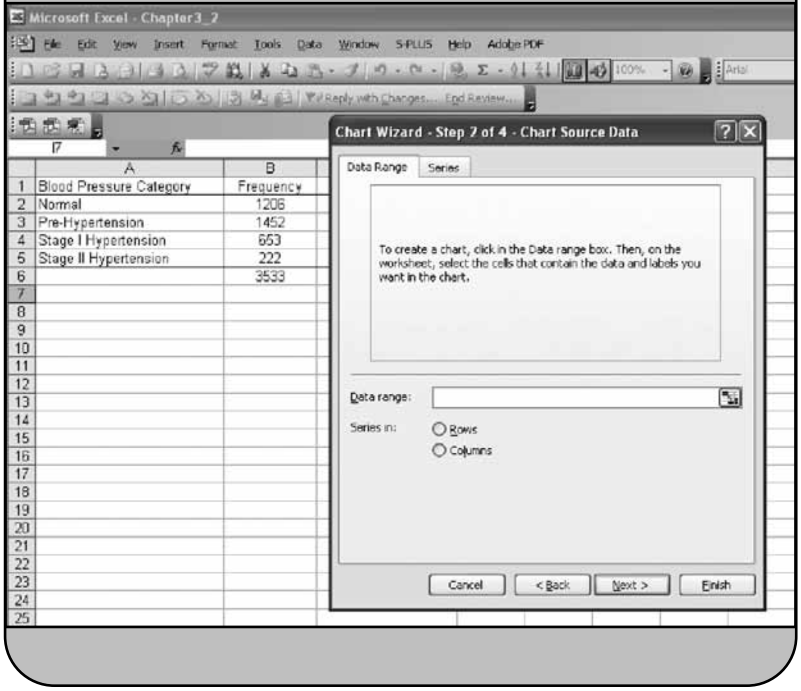

Microsoft Excel - Chapter3_2
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
I7
Data Range
Series
Cancel
< Back
Next >
Finish

Chart Wizard - Step 2 of 4 - Chart Source Data

	A	B
1	Blood Pressure Category	Frequency
2	Normal	1206
3	Pre-Hypertension	1452
4	Stage I Hypertension	653
5	Stage II Hypertension	222
6		3533

To create a chart, click in the Data range box. Then, on the worksheet, select the cells that contain the data and labels you want in the chart.

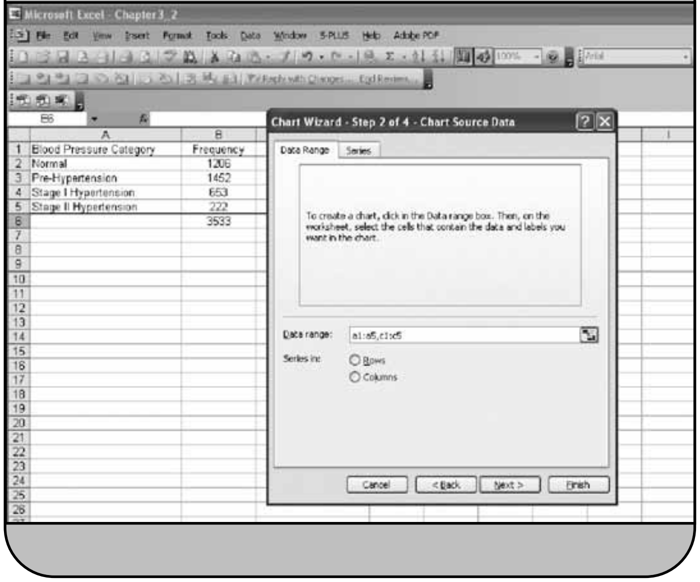
Data range:

Series in: ○ Rows ○ Columns

in column C. The data are specified as "A1:A5,C1:C5." The first range contains the response labels and the second range contains the relative frequencies (Figure 3–32).

When we click **Next**, Excel provides a snapshot of the display and opens a dialog box with a number of formatting options (Figure 3–33). The first tab contains options for labels and titles. The chart title is entered first, followed by labels for the *x*- and *y*-axis of the display (Figure 3–34). At this stage, we can also format other aspects of the display by selecting other tabs that are shown in Figure 3–34. However, if we click **Next**, Excel asks for the location of the display (Figure 3–35).

FIGURE 3–32 Specifying the Data Range


Microsoft Excel - Chapter3_2
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
B6
Chart Wizard - Step 2 of 4 - Chart Source Data
Data Range
Series
Cancel
< Back
Next >
Finish

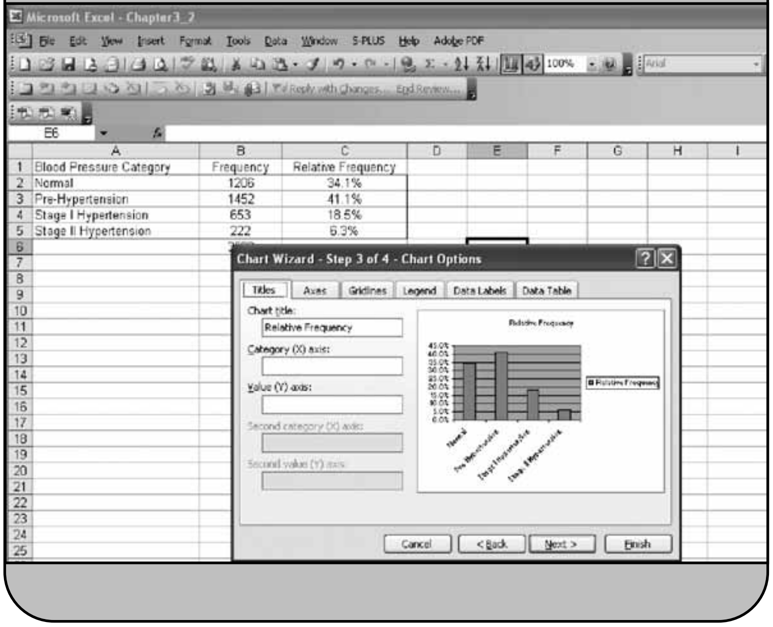
	A	B
1	Blood Pressure Category	Frequency
2	Normal	1206
3	Pre-Hypertension	1452
4	Stage I Hypertension	653
5	Stage II Hypertension	222
6		3533

To create a chart, click in the Data range box. Then, on the worksheet, select the cells that contain the data and labels you want in the chart.

Data range: a1:a5,c1:c5

Series in: ○ Rows ○ Columns

FIGURE 3–33 Formatting Options



	A	B	C
1	Blood Pressure Category	Frequency	Relative Frequency
2	Normal	1206	34.1%
3	Pre-Hypertension	1452	41.1%
4	Stage I Hypertension	653	18.5%
5	Stage II Hypertension	222	6.3%

Chart title: Relative Frequency

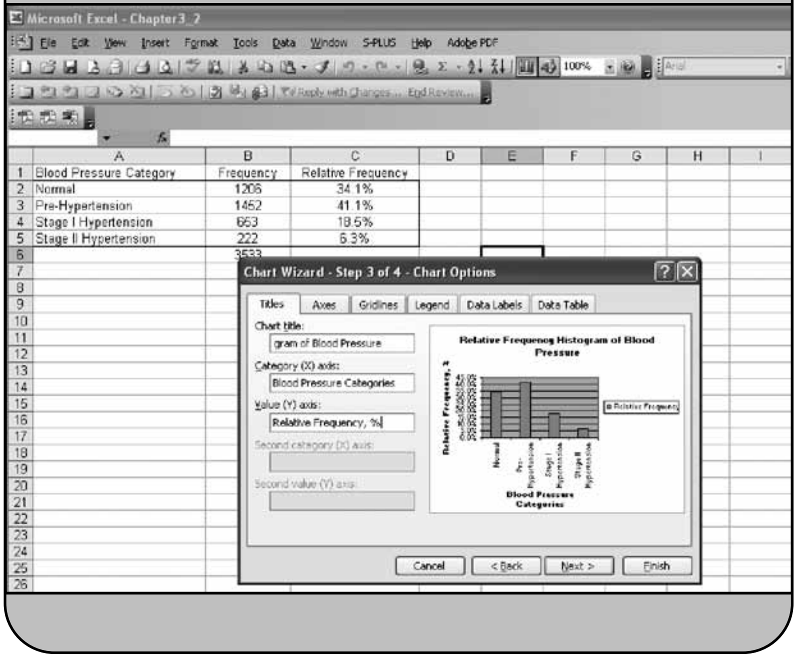
Category (X) axis:

Value (Y) axis:

Second category (X) axis:

Second value (Y) axis:

FIGURE 3–34 Specifying a Title and Labels for the Axes


Microsoft Excel - Chapter3_2
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...

	A	B	C
1	Blood Pressure Category	Frequency	Relative Frequency
2	Normal	1206	34.1%
3	Pre-Hypertension	1452	41.1%
4	Stage I Hypertension	653	18.5%
5	Stage II Hypertension	222	6.3%
6		3533	

Chart Wizard - Step 3 of 4 - Chart Options

Chart title: gram of Blood Pressure

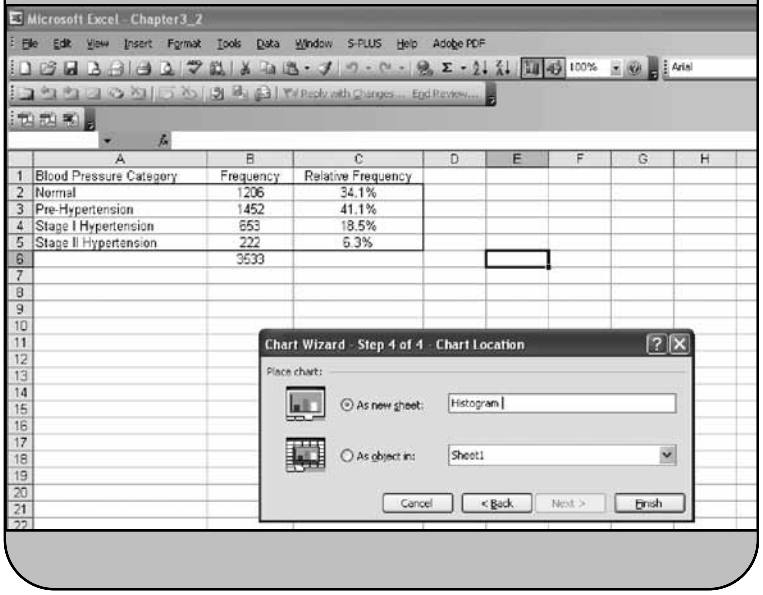
Category (X) axis: Blood Pressure Categories

Value (Y) axis: Relative Frequency, %

Second category (X) axis:

Second value (Y) axis:

FIGURE 3–35 Specifying a Location for the Display


Microsoft Excel - Chapter3_2
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...

	A	B	C
1	Blood Pressure Category	Frequency	Relative Frequency
2	Normal	1206	34.1%
3	Pre-Hypertension	1452	41.1%
4	Stage I Hypertension	653	18.5%
5	Stage II Hypertension	222	6.3%
6		3533	

Chart Wizard - Step 4 of 4 - Chart Location

Place chart:

● As new sheet: Histogram

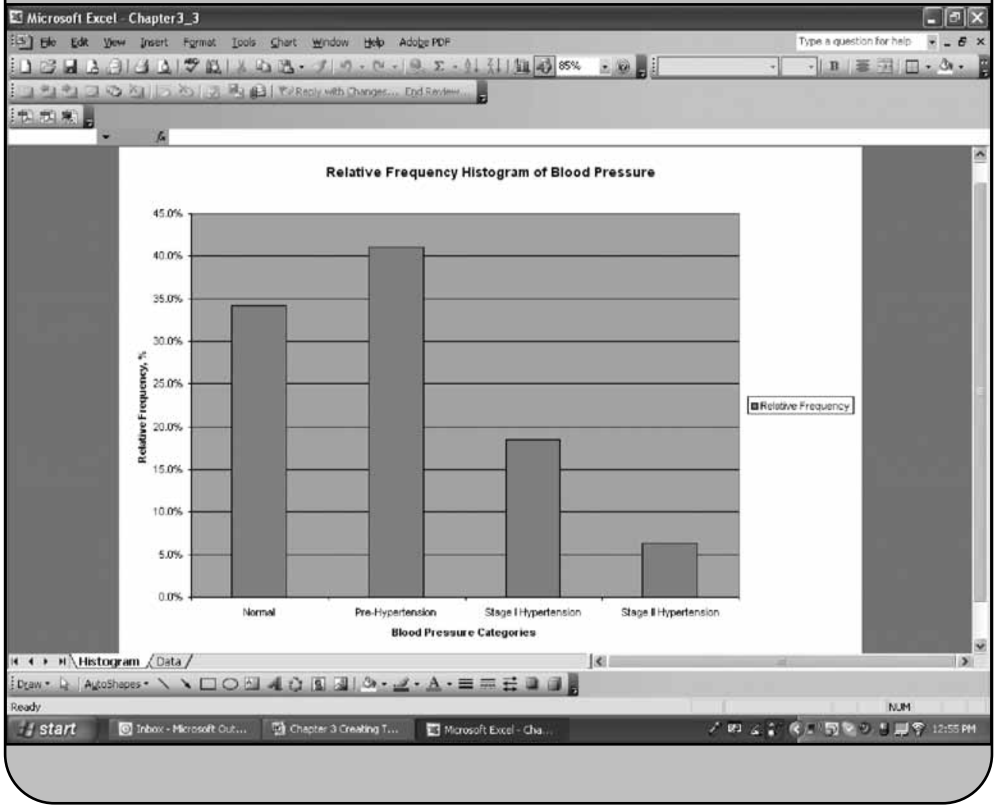
○ As object in: Sheet1

Cancel
< Back
Next >
Finish

We can place the display in a new worksheet—in which case we provide a name for the new worksheet—or we can place the display in the current worksheet. In Figure 3–35, we specify that we would like the display placed in a new worksheet called *Histogram*. We click **Finish** and Excel generates the new worksheet containing the display (Figure 3–36).

We can continue to format the display. We first reduce the gap between adjacent bars to zero by double-clicking on any of the bars, which opens the dialog box shown in Figure

FIGURE 3–36 The Display



3–27, and we specify a gap width of 0. We then remove the label on the right side ("Relative Frequency") by right-clicking on the label and clicking **Clear**. We then remove the grey background by double-clicking anywhere in the background. This opens a dialog box where we can either select another color or choose **None** under the **Area** section for no background color. Finally, we remove the horizontal lines by right-clicking on any line and clicking **Delete**. (Excel also has a **Clear** option, which can be used to remove the horizontal lines.) After implementing these steps, the histogram is as shown in Figure 3–37.

3.4 SCATTER DIAGRAMS

A popular graphical display to illustrate the relationship between two continuous variables is a scatter diagram. Scatter diagrams are useful in linear regression analysis applications to assess the relationship between a continuous independent and a continuous dependent variable (see Chapter 9 of the textbook).

Example 3.4. Suppose we wish to examine the association between body mass index (BMI) and systolic blood pressure (SBP) in a sample of $n = 12$ persons who are not taking antihypertensive medication. The data are entered into an Excel worksheet as shown in Figure 3–38.

To generate the scatter diagram, we again use the Chart Wizard. However, rather than specifying the data for the display in the dialog box that appears as part of the Chart Wizard (Figure 3–32), we illustrate an alternative approach. Here we first highlight the data for the display (in this case, cell B1 through cell C13) and then click on the Chart Wizard icon. From the list of **Standard Chart Types**, we select **XY (Scatter)**, as shown in Figure 3–39.

Recall that in regression analysis, we use *x* to denote the independent variable and *y* to denote the dependent variable (see Chapter 9 of the textbook). In this example, BMI is the independent variable and SBP is the dependent variable. Excel assumes that the first variable specified is the independent variable (x) and the second is the dependent variable (y). This is the way the data in Example 3.4 are organized. The default subtype (shown in black in Figure 3–39) is the scatter diagram. The other subtypes connect the data either with lines or curves, depending on which sub-

FIGURE 3–37 The Relative Frequency Histogram for Blood Pressure Categories

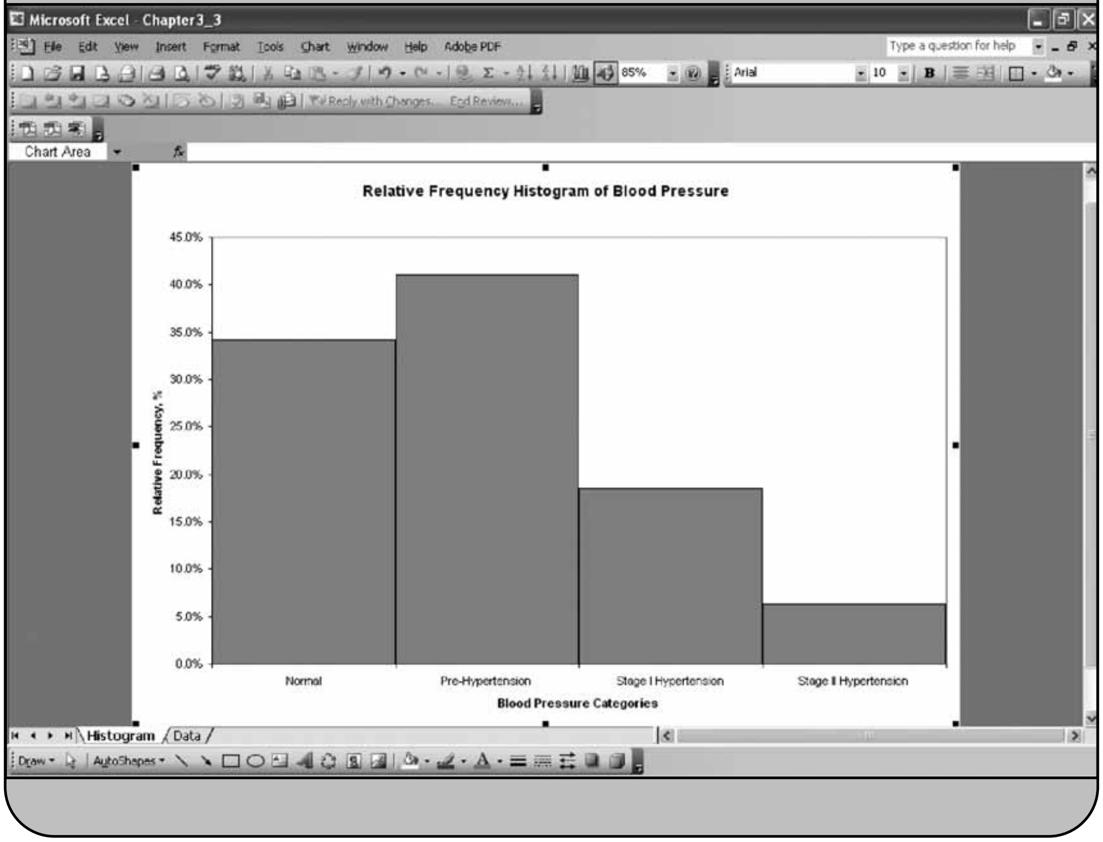
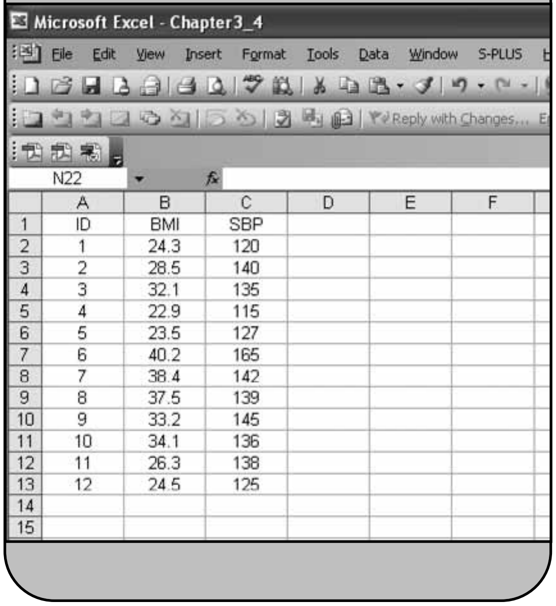


FIGURE 3–38 Data for Scatter Diagram


Microsoft Excel - Chapter3_4
File Edit View Insert Format Tools Data Window S-PLUS
Reply with Changes...
N22

	A	B	C	D	E	F
1	ID	BMI	SBP			
2	1	24.3	120			
3	2	28.5	140			
4	3	32.1	135			
5	4	22.9	115			
6	5	23.5	127			
7	6	40.2	165			
8	7	38.4	142			
9	8	37.5	139			
10	9	33.2	145			
11	10	34.1	136			
12	11	26.3	138			
13	12	24.5	125			
14						
15						

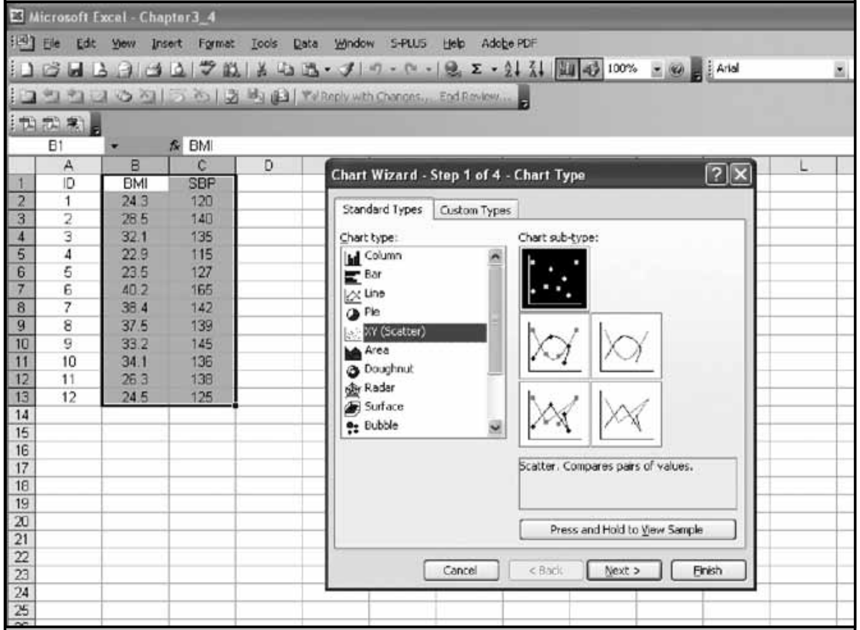
type is selected. Once we click **Next**, Excel displays a template of the scatter diagram along with a dialog box in which the range of the data are indicated (Figure 3–40). Note that the data range is automatically filled (we actually selected the data range before invoking the Chart Wizard).

We click **Next** and are presented with a template of the scatter diagram and options for formatting (Figure 3–41). We can input a title as well as labels for the *x* (horizontal) and *y* (vertical) axes. We can use the other options to remove the gridlines and the legend. When we click **Next**, Excel asks for a location for the scatter diagram (Figure 3–42).

We can place the scatter diagram in a new worksheet—in which case we provide a name for the new worksheet—or we can place the display in the current worksheet. In Figure 3–42, we specify that we would like the display placed in a new worksheet called *Scatter Diagram*. We click **Finish** and Excel generates the new worksheet containing the display (Figure 3–43).

Excel automatically scales the *x*- and *y*-axes from zero to a value larger than the maximum value in the dataset. There are many variables whose theoretical minimum is much larger

FIGURE 3–39 Using the Chart Wizard to Generate a Scatter Diagram



Microsoft Excel - Chapter3_4

B1 BMI

	A	B	C
1	ID	BMI	SBP
2	1	24.3	120
3	2	28.5	140
4	3	32.1	135
5	4	22.9	115
6	5	23.5	127
7	6	40.2	165
8	7	38.4	142
9	8	37.5	139
10	9	33.2	145
11	10	34.1	136
12	11	26.3	138
13	12	24.5	125

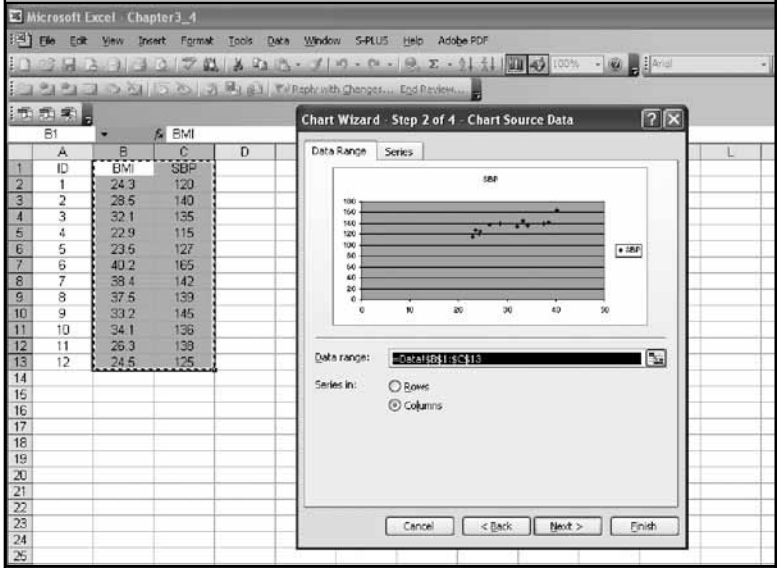
Chart Wizard - Step 1 of 4 - Chart Type

Standard Types

Chart type: XY (Scatter)

Scatter. Compares pairs of values.

FIGURE 3–40 Template for Scatter Diagram



Microsoft Excel - Chapter3_4

B1 BMI

	A	B	C
1	ID	BMI	SBP
2	1	24.3	120
3	2	28.5	140
4	3	32.1	135
5	4	22.9	115
6	5	23.5	127
7	6	40.2	165
8	7	38.4	142
9	8	37.5	139
10	9	33.2	145
11	10	34.1	136
12	11	26.3	138
13	12	24.5	125

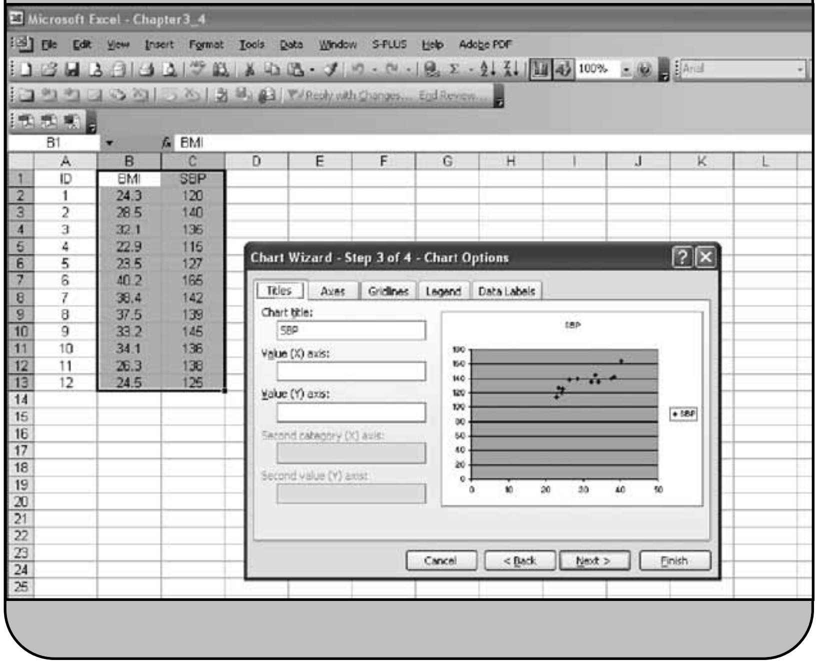
Chart Wizard - Step 2 of 4 - Chart Source Data

Data Range

Data range: =Data!$B$1:$C$13

Series in: ○ Rows ● Columns

FIGURE 3–41 Formatting Options


Microsoft Excel - Chapter3_4
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
B1
BMI
SBP
0
10
20
30
40
50
60
80
100
120
140
160
180
SBP

	A	B	C
1	ID	BMI	SBP
2	1	24.3	120
3	2	28.5	140
4	3	32.1	136
5	4	22.9	116
6	5	23.5	127
7	6	40.2	165
8	7	38.4	142
9	8	37.5	139
10	9	33.2	145
11	10	34.1	136
12	11	26.3	138
13	12	24.5	125

Chart Wizard - Step 3 of 4 - Chart Options

Titles | Axes | Gridlines | Legend | Data Labels

Chart title: SBP

Value (X) axis:

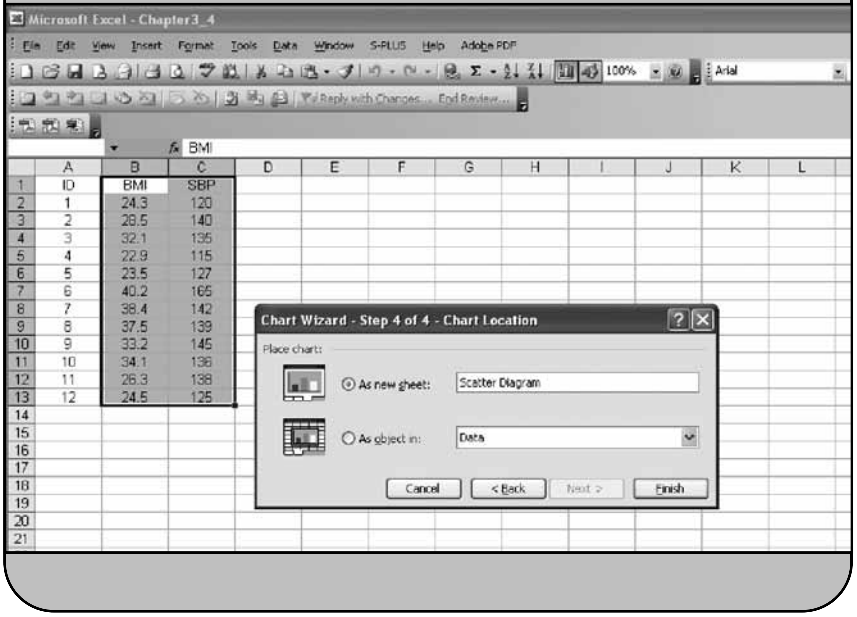
Value (Y) axis:

Second category (X) axis:

Second value (Y) axis:

Cancel | < Back | Next > | Finish

FIGURE 3–42 Specifying a Location for the Chart


Microsoft Excel - Chapter3_4
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
BMI

	A	B	C
1	ID	BMI	SBP
2	1	24.3	120
3	2	28.5	140
4	3	32.1	135
5	4	22.9	115
6	5	23.5	127
7	6	40.2	165
8	7	38.4	142
9	8	37.5	139
10	9	33.2	145
11	10	34.1	136
12	11	26.3	138
13	12	24.5	125

Chart Wizard - Step 4 of 4 - Chart Location

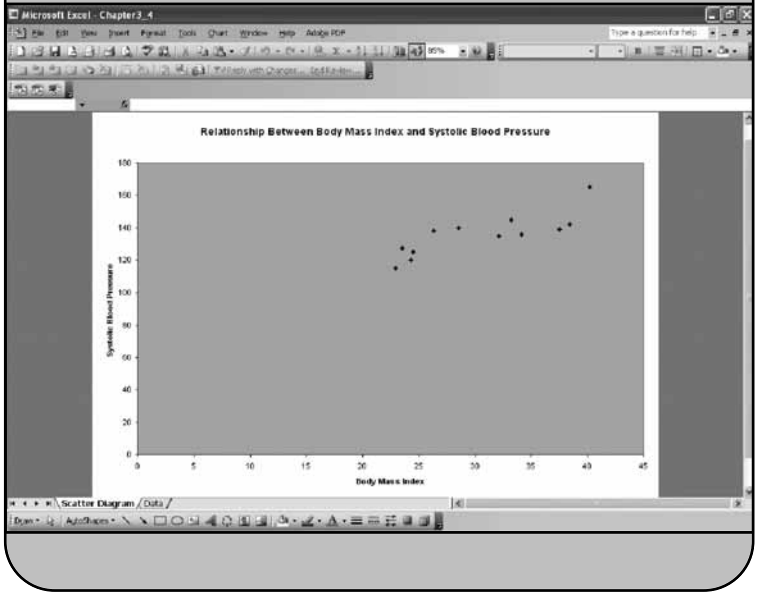
Place chart:

● As new sheet: Scatter Diagram

○ As object in: Data

Cancel | < Back | Next > | Finish

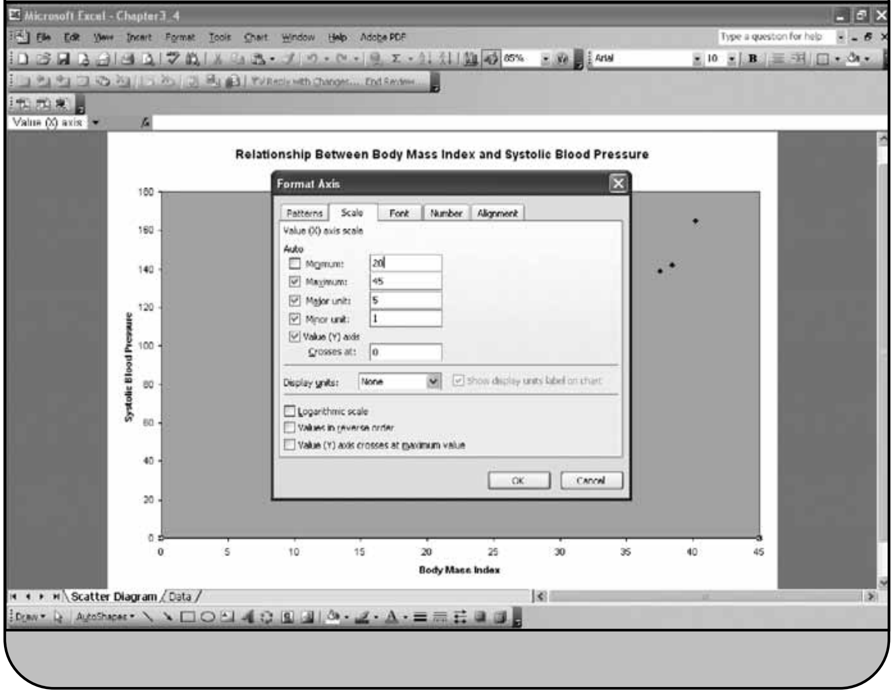
FIGURE 3–43 The Scatter Diagram


Microsoft Excel - Chapter3_4
File Edit View Insert Format Tools Chart Window Help Adobe PDF
Type a question for help
Relationship Between Body Mass Index and Systolic Blood Pressure
180
160
140
120
100
80
60
40
20
0
Systolic Blood Pressure
0
5
10
15
20
25
30
35
40
45
Body Mass Index
Scatter Diagram
Data
Draw
AutoShapes

than zero; BMI and SBP are two examples. Thus, we will next rescale the *x*- and *y*-axes to start at more reasonable values. For example, we will rescale the *x*-axis from 20 to 45 and the *y*-axis from 100 to 180. We first rescale the *x*-axis (BMI). This is done by double-clicking anywhere along the *x* axis. This brings up the dialog box shown in Figure 3–44.

In the dialog box, we choose the **Scale** tab and specify the desired minimum and maximum and the major unit. The

FIGURE 3–44 Rescaling the Axes



Value (X) axis scale

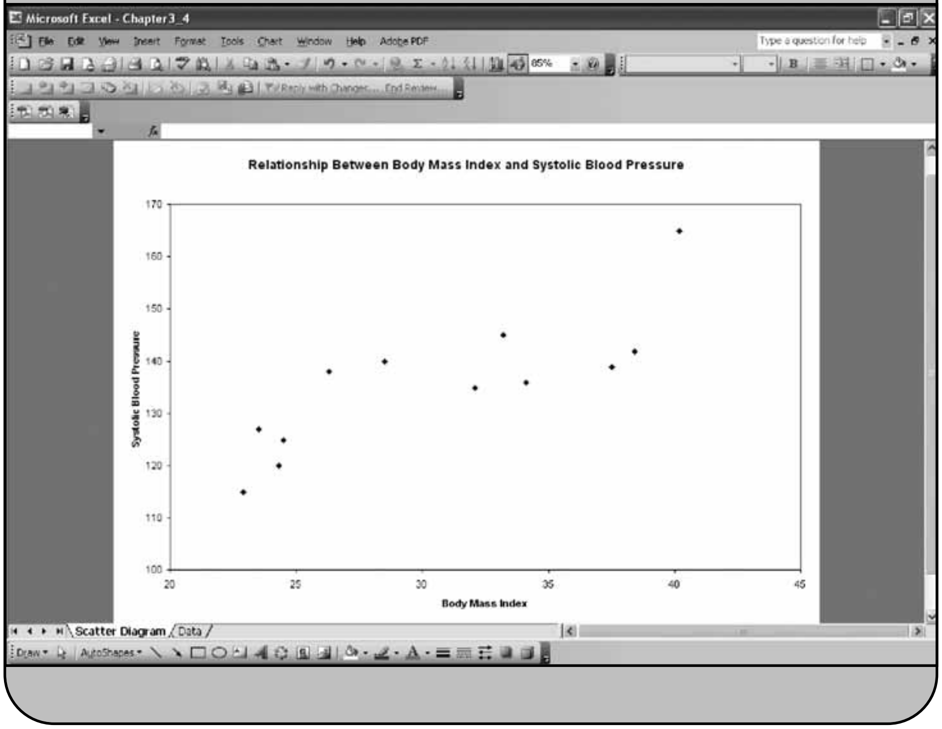
Auto

- □ Minimum: 20
- ☑ Maximum: 45
- ☑ Major unit: 5
- ☑ Minor unit: 1
- ☑ Value (Y) axis
  Crosses at: 0

Display units: None ☑ Show display units label on chart

- □ Logarithmic scale
- □ Values in reverse order
- □ Value (Y) axis crosses at maximum value

FIGURE 3–45 Scatter Diagram of BMI and SBP


Microsoft Excel - Chapter3_4
File Edit View Insert Format Tools Chart Window Help Adobe PDF
Type a question for help
85%
Relationship Between Body Mass Index and Systolic Blood Pressure
Systolic Blood Pressure
170
160
150
140
130
120
110
100
20
25
30
35
40
45
Body Mass Index
Scatter Diagram
Data
Draw
AutoShapes

major unit represents the distance between tick marks shown on that axis. We then click **OK**. After following the same sequence for the *y*-axis (SBP) and changing the background from grey to white (i.e., double-clicking anywhere in the background, which opens a dialog box where we can either select another color or choose **None** under the **Area** section for no background color). We also remove the horizontal lines by right-clicking on any line and choosing **Delete** (or **Clear**). The scatter diagram is as shown in Figure 3–45.

The scatter diagram illustrates the positive association between body mass index and systolic blood pressure. The scatter diagram can now be copied from Excel into a manuscript, report, or presentation.

3.5 PRACTICE PROBLEMS

1. The data in Table 3–4 were measured in $n = 15$ college seniors in a cross-sectional study of alcohol consumption. Each participant was asked their gender, year in school, the age at which they first consumed alcohol, and the number of alcoholic drinks they consume on a typical drinking night. Generate frequency distribution tables for gender and year in school using the "Data/Pivot Table and Pivot Chart Report" option on the menu bar.
2. For the data in Table 3–4, generate a frequency bar chart for gender using the **Pivot Table and Pivot Chart Report** option.
3. For the data in Table 3–4, generate a frequency histogram for year in school using the **Pivot Table and Pivot Chart** option.
4. For the data in Table 3–4, create a frequency distribution table for drinking status, defined by the following numbers of drinks per night:

Abstinent	0
Light	1–3
Moderate	4–5
Heavy	6 or more

5. For the data in Table 3–4, generate a frequency histogram for drinking status using the Chart Wizard.
6. For the data in Table 3–4, generate a scatter diagram to display the association between age at first drink and number of drinks per night. (Note that the sample size for analysis is $n = 13$.)

TABLE 3–4 Data for Practice Problem

ID Number	Gender	Year in School	Age at First Drink*	Number of Drinks/Night
1	M	Freshman	14	5
2	M	Senior	11	8
3	F	Junior	17	3
4	M	Junior	13	9
5	F	Freshman		0
6	F	Sophomore	15	4
7	F	Freshman	15	0
8	F	Freshman		0
9	M	Senior	15	7
10	F	Junior	21	5
11	M	Junior	18	3
12	M	Senior	14	6
13	F	Junior	19	3
14	F	Sophomore	18	2
15	M	Junior	20	4

*Students who never had a drink were excluded.

CHAPTER 4

Summarizing Continuous Variables in a Sample

In Chapter 4 of the textbook, we presented summary statistics for dichotomous, ordinal, categorical, and continuous variables. We discussed both numerical and graphical summaries. Numerical summaries for continuous variables include the sample mean, standard deviation, median, and quartiles. Numerical summaries for ordinal and categorical data use frequency distribution tables, and these were discussed in detail in Chapter 3 of the Microsoft Office® Excel® workbook. In Chapter 3 of the Excel workbook, we also presented several graphical displays for sample data. Numerical summaries for continuous variables measured in a sample are discussed here.

The most appropriate numerical summaries for continuous variables depend on whether or not there are outliers. Regardless of whether there are outliers, the summary should always include the sample size, a measure of central tendency or a typical value (e.g., the mean or median), and a measure of variability (e.g., standard deviation or interquartile range). Here we generate summaries that include numerous statistics; the investigator must choose those that are most appropriate to summarize a particular characteristic.

4.1 THE DESCRIPTIVE STATISTICS ANALYSIS TOOL

Here we generate summary statistics on a continuous variable using the Descriptive Statistics Analysis Tool available in the Excel Data Analysis ToolPak.

Example 4.1. In Example 4.3 in the textbook, we analyzed data from a subset of $n = 10$ participants attending the seventh examination of the offspring in the Framingham Heart Study. The data values were presented in Table 4–13 of the text and are shown here in Table 4–1.

The means, standard deviations, and other statistics can be computed using the Descriptive Statistics Analysis Tool in the Data Analysis ToolPak. Figure 4–1 shows the data entered into an Excel worksheet. We now use the Data Analysis ToolPak to generate descriptive statistics on each continuous variable. We begin with body mass index (BMI).

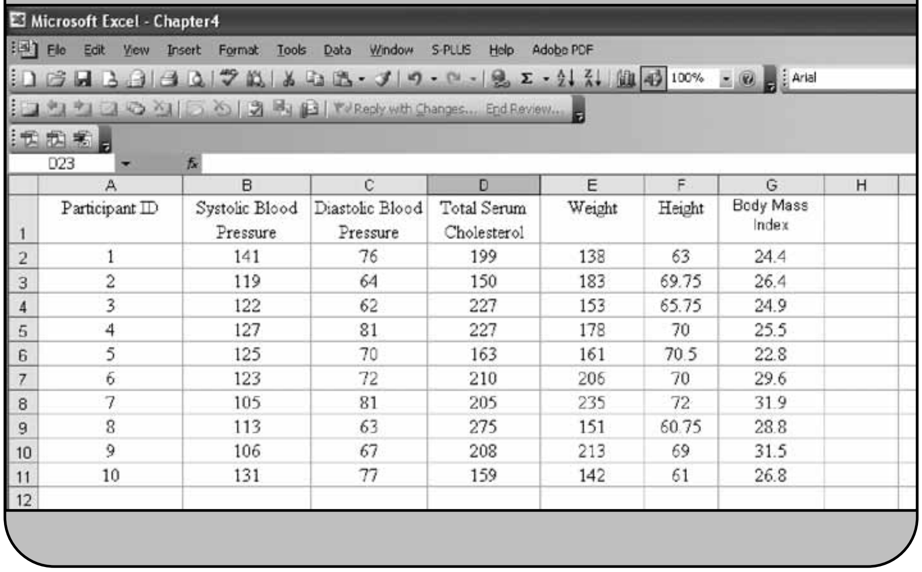
When we select the **Tools/Data Analysis** option from the menu bar, a dialog box with the various analysis tools appears (Figure 4–2). From the list of analysis tools, we choose **Descriptive Statistics**. We are then presented with a second dialog box. Figure 4–3 displays this dialog box, which requires input for the descriptive statistics.

In the dialog box, we must provide the location of the data values we wish to summarize. This is requested under **Input Range**. In the **Input Range** text field, we specify cell G1 through cell G11. The data actually reside in cell G2 through cell G11, with the variable name in cell G1. We must check the box to indicate that the variable name or label is in the first row. By doing so, Excel prints the variable name on the output. If we do not check the box to indicate that the variable label is in the first row, then we must specify the input range as "G2:G11". In the next section of the dialog box, we specify where we would like the results placed. The options are to place the output in the current worksheet, in a new worksheet within the same workbook, or in a new workbook. If we select **Output Range**, we must

TABLE 4–1 Subsample of *n* = 10 Participants Attending the Seventh Examination of the Framingham Offspring Study

Participant ID	Systolic Blood Pressure	Diastolic Blood Pressure	Total Serum Cholesterol	Weight (lb)	Height (in)	BMI
1	141	76	199	138	63.00	24.4
2	119	64	150	183	69.75	26.4
3	122	62	227	153	65.75	24.9
4	127	81	227	178	70.00	25.5
5	125	70	163	161	70.50	22.8
6	123	72	210	206	70.00	29.6
7	105	81	205	235	72.00	31.9
8	113	63	275	151	60.75	28.8
9	106	67	208	213	69.00	31.5
10	131	77	159	142	61.00	26.8

FIGURE 4–1 Data for Analysis Entered into Excel

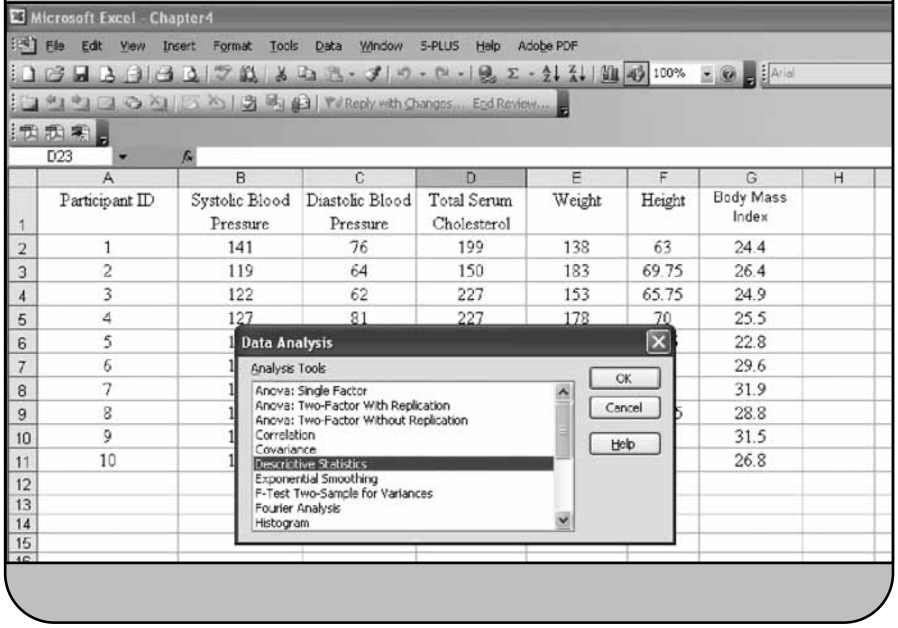

Microsoft Excel - Chapter4
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
D23
fx

	A	B	C	D	E	F	G	H
1	Participant ID	Systolic Blood Pressure	Diastolic Blood Pressure	Total Serum Cholesterol	Weight	Height	Body Mass Index	
2	1	141	76	199	138	63	24.4	
3	2	119	64	150	183	69.75	26.4	
4	3	122	62	227	153	65.75	24.9	
5	4	127	81	227	178	70	25.5	
6	5	125	70	163	161	70.5	22.8	
7	6	123	72	210	206	70	29.6	
8	7	105	81	205	235	72	31.9	
9	8	113	63	275	151	60.75	28.8	
10	9	106	67	208	213	69	31.5	
11	10	131	77	159	142	61	26.8	
12								

specify the cell address for the top-left corner of the table containing the results (e.g., I1). In Figure 4–3, we select **New Worksheet Ply** to place the results in a new worksheet in the same workbook, and we specify the name of the new worksheet as *Descriptive Statistics*. In the last section of the dialog box, we request that Excel generate summary statistics for BMI by checking the appropriate box in the dialog box. The results of the analysis are shown in Figure 4–4.

Notice that the results are contained in a new worksheet in the workbook called *Descriptive Statistics*. The data are contained in the *Data* worksheet. When the **Descriptive Statistics** option is selected, Excel generates all of the statistics shown in Figure 4–4. The default (or automatic) statistics produced

FIGURE 4–2 Invoking the Data Analysis ToolPak

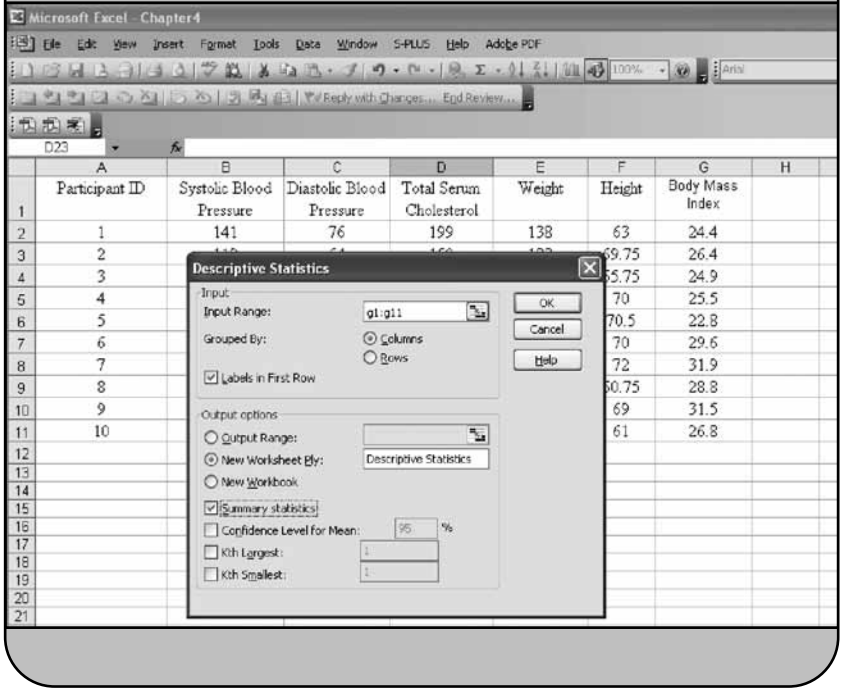

Microsoft Excel - Chapter4
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
D23
Data Analysis
OK
Cancel
Help

	A	B	C	D	E	F	G	H
1	Participant ID	Systolic Blood Pressure	Diastolic Blood Pressure	Total Serum Cholesterol	Weight	Height	Body Mass Index	
2	1	141	76	199	138	63	24.4	
3	2	119	64	150	183	69.75	26.4	
4	3	122	62	227	153	65.75	24.9	
5	4	127	81	227	178	70	25.5	
6	5	1					22.8	
7	6	1					29.6	
8	7	1					31.9	
9	8	1					28.8	
10	9	1					31.5	
11	10	1					26.8	

Analysis Tools

- Anova: Single Factor
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram

FIGURE 4–3 The Descriptive Statistics Analysis Tool


Microsoft Excel - Chapter4
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
D23
Descriptive Statistics
OK
Cancel
Help

	A	B	C	D	E	F	G	H
1	Participant ID	Systolic Blood Pressure	Diastolic Blood Pressure	Total Serum Cholesterol	Weight	Height	Body Mass Index	
2	1	141	76	199	138	63	24.4	
3	2					9.75	26.4	
4	3					5.75	24.9	
5	4					70	25.5	
6	5					70.5	22.8	
7	6					70	29.6	
8	7					72	31.9	
9	8					0.75	28.8	
10	9					69	31.5	
11	10					61	26.8	

Input

Input Range: g1:g11

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply: Descriptive Statistics

○ New Workbook

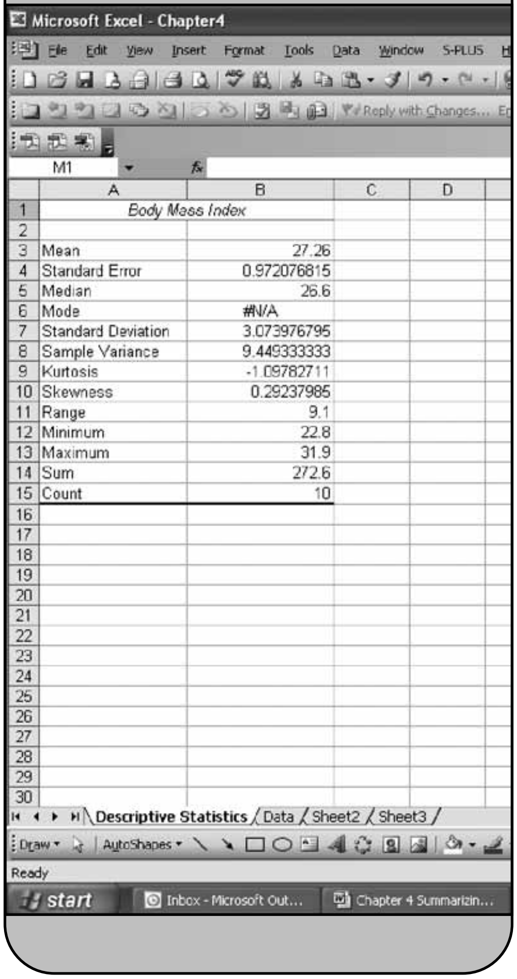
☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

FIGURE 4–4 Descriptive Statistics for BMI


Microsoft Excel - Chapter4
File Edit View Insert Format Tools Data Window S-PLUS
Reply with Changes...
M1
A B C D
Descriptive Statistics / Data / Sheet2 / Sheet3
Draw AutoShapes
Ready
start
Inbox - Microsoft Out...
Chapter 4 Summarizin...

	A	B
1	*Body Mass Index*	
2		
3	Mean	27.26
4	Standard Error	0.972076815
5	Median	26.6
6	Mode	#N/A
7	Standard Deviation	3.073976795
8	Sample Variance	9.449333333
9	Kurtosis	-1.09782711
10	Skewness	0.29237985
11	Range	9.1
12	Minimum	22.8
13	Maximum	31.9
14	Sum	272.6
15	Count	10

TABLE 4–2 Default Statistics in Descriptive Statistics Analysis Tool

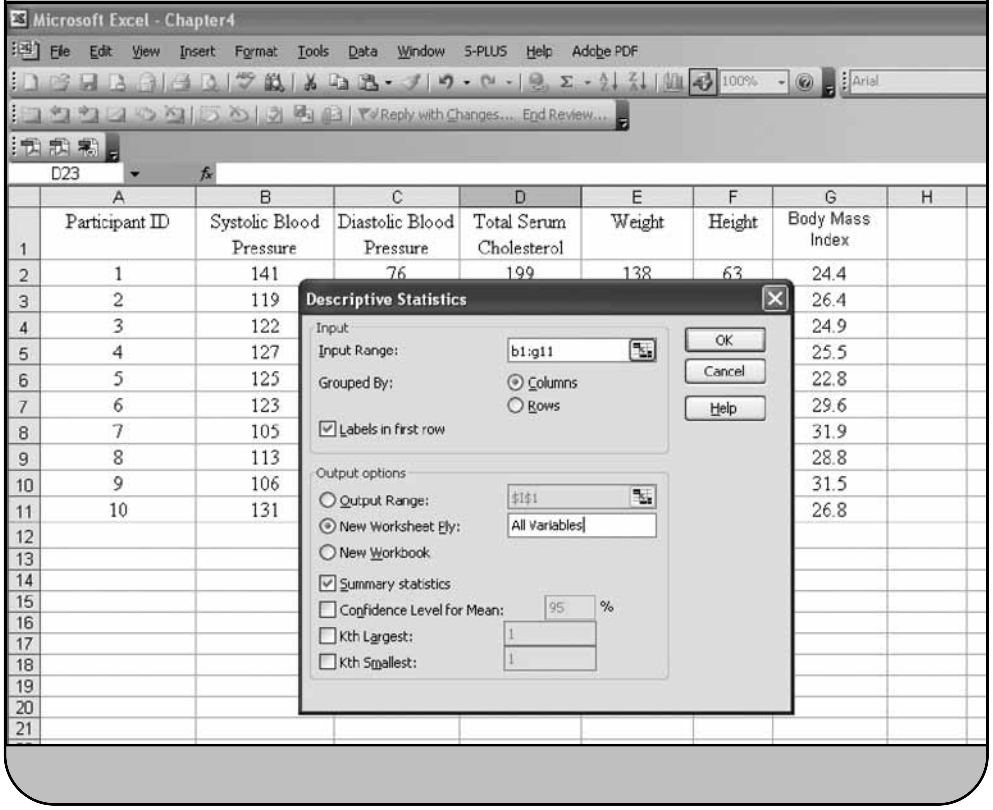
Statistic	Formula/Description
Sample mean	$\overline{X} = \dfrac{\Sigma X}{n}$
Standard error	$\text{SE} = \dfrac{s}{\sqrt{n}}$
Median	Middle value (50% above and 50% below)
Mode	The most frequent value. If there is no value that appears more than any other, Excel indicates this with "#N/A"
Sample standard deviation	$s = \sqrt{\dfrac{\Sigma(X-\overline{X})^2}{n-1}}$
Sample variance	$s^2 = \dfrac{\Sigma(X-\overline{X})^2}{n-1}$
Kurtosis	Reflects the thickness of the tails of a distribution of a continuous characteristic as compared to a normal distribution (see Chapter 5 of the textbook). The kurtosis of a normal distribution is 0.
Skewness	Reflects the symmetry of a distribution of a continuous characteristic as compared to a normal distribution (see Chapter 5 of the textbook). The skewness of a normal distribution is 0.
Range	Range = Maximum − Minimum
Minimum	The smallest value in the dataset
Maximum	The largest value in the dataset
Sum	The sum of the observations, ΣX
Count	The number of observations, n

with the Descriptive Statistics Analysis Tool are shown in Table 4–2. In Chapter 4 of the textbook, we discussed many (but not all) of these statistics. The most relevant statistics for BMI are the sample size $n = 10$ (referred to as *Count*), the sample mean $\overline{X} = 27.26$, and the sample standard deviation $s = 3.07$. Notice that Excel generates summary statistics with many more decimal places than would be reasonable to report.

It is also possible to generate descriptive statistics for several continuous variables simultaneously. Again, we select **Tools/Data Analysis**, and then **Descriptive Statistics** from the list of analysis tools. We then specify the range of all variables in the input range field in the first section of the dialog box. In Figure 4–5, we specify the input range as B1 through G11. This range includes all variables and variable names or labels (which are contained in row 1).

Again, we request that the output be placed in a new worksheet, here called *All Variables*. A screenshot of partial results is shown in Figure 4–6. Descriptive statistics are computed on each of the variables but cannot be displayed on the screen without substantially reducing the font size.

FIGURE 4–5 Descriptive Statistics for Several Variables


Microsoft Excel - Chapter4
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
D23
fx

	A	B	C	D	E	F	G	H
1	Participant ID	Systolic Blood Pressure	Diastolic Blood Pressure	Total Serum Cholesterol	Weight	Height	Body Mass Index	
2	1	141	76	199	138	63	24.4	
3	2	119					26.4	
4	3	122					24.9	
5	4	127					25.5	
6	5	125					22.8	
7	6	123					29.6	
8	7	105					31.9	
9	8	113					28.8	
10	9	106					31.5	
11	10	131					26.8	

Descriptive Statistics

Input
Input Range: b1:g11
Grouped By: ● Columns ○ Rows
☑ Labels in first row

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

4.2 DESCRIPTIVE STATISTICS USING EXCEL FUNCTIONS

There are other statistics that we discussed in Chapter 4 of the textbook that are useful for continuous variables and are not automatically generated with the Descriptive Statistics Analysis Tool. An example is the first (and third) quartile. Excel does not include the quartiles in the output of the Descriptive Statistics Analysis Tool, but they can be computed with an Excel function. The function is invoked as "= QUARTILE(*data range*, *quartile number*)".

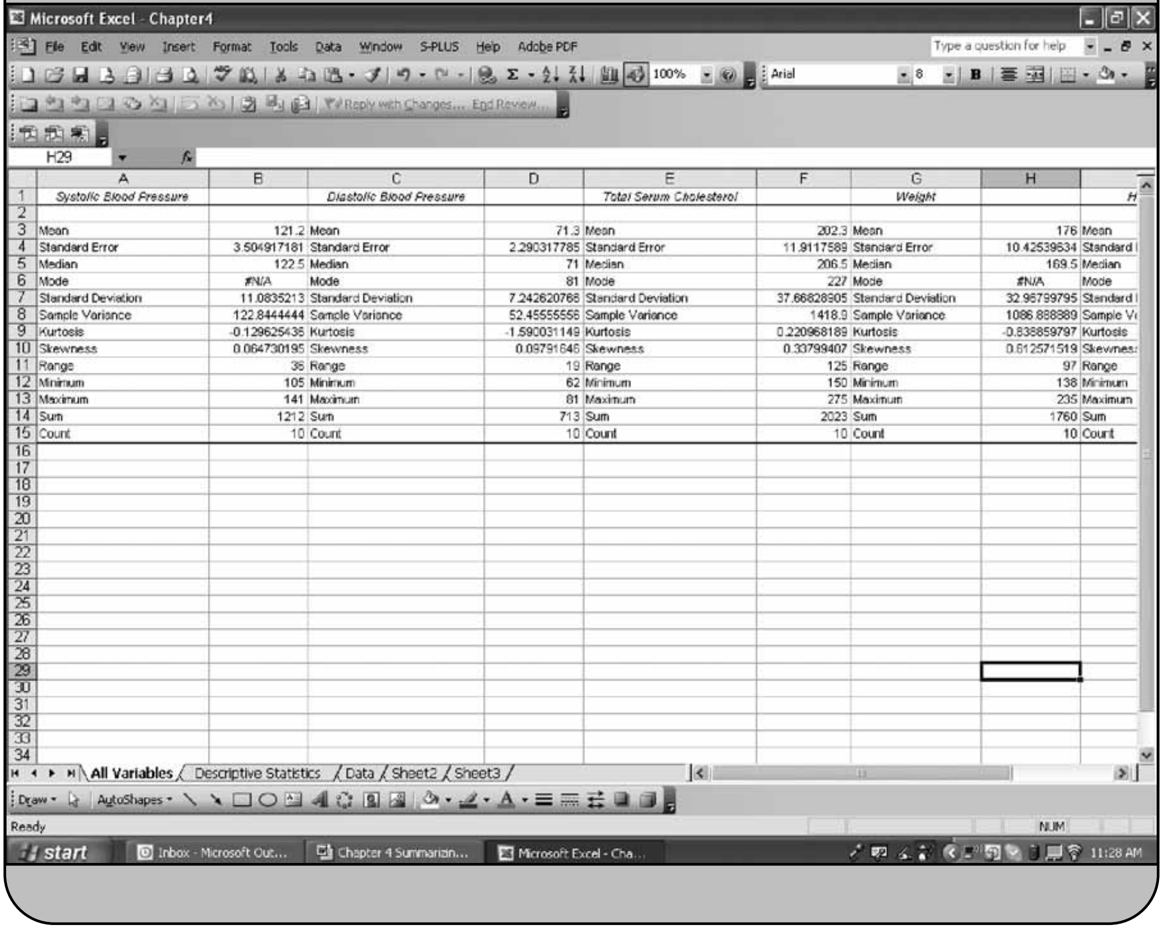
The data range is defined by the addresses of the cells containing the first and last observations in the dataset, separated by a colon. For example, using the data in Figure 4–1, systolic blood pressures occupy the range B2:B11, diastolic blood pressures occupy the range C2:C11, and total cholesterol values occupy the range D2:D11. The quartile numbers range from 0 to 4. Table 4–3 indicates which values are generated for each quartile number.

We now use the quartile function to generate quartiles for the diastolic blood pressures (DBP) shown in Figure 4–1. In Figure 4–7, we use the quartile function to compute the first quartile of DBP and place it in cell C13. Notice that we also added a label into cell B13. In Figure 4–8, we compute the second (median) and third (Q3) quartiles of DBP. These are computed using the formulas "=QUARTILE(C2:C11,2)" and "=QUARTILE(C2:C11,3)", respectively. Excel uses an

TABLE 4–3 Quartiles Produced by the QUARTILE Function

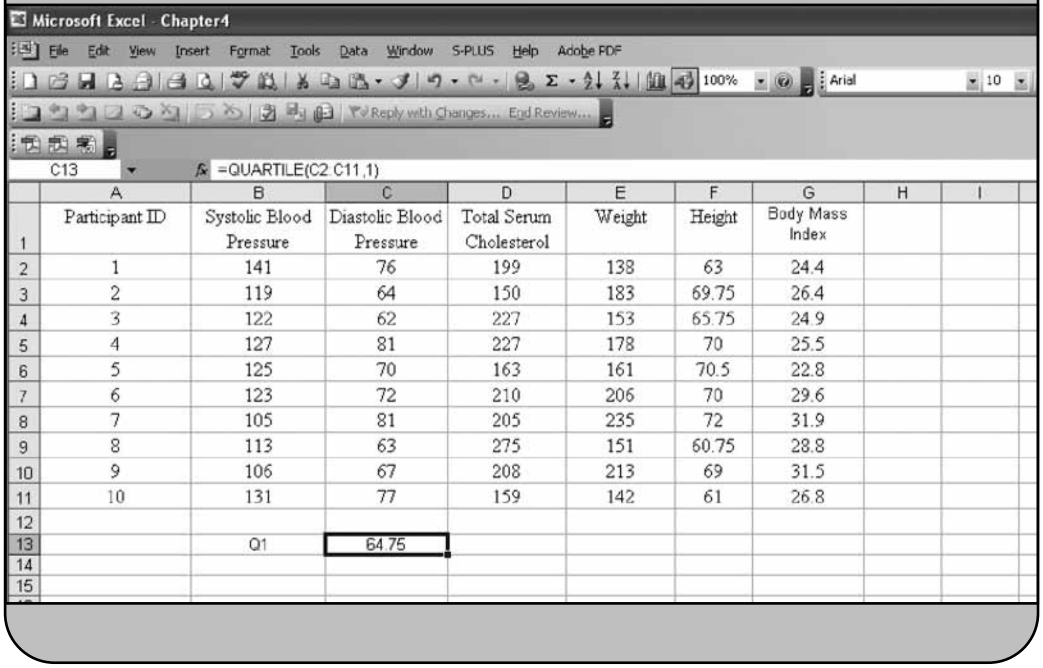
Quartile Number	Statistic
0	Minimum
1	First quartile (holds 25% of values below it)
2	Median
3	Third quartile (holds 25% of values above it)
4	Maximum

FIGURE 4–6 Descriptive Statistics for All Variables



	A	B	C	D	E	F	G	H	I
1	*Systolic Blood Pressure*		*Diastolic Blood Pressure*		*Total Serum Cholesterol*		*Weight*		*H*
2									
3	Mean	121.2	Mean	71.3	Mean	202.3	Mean	176	Mean
4	Standard Error	3.504917181	Standard Error	2.290317785	Standard Error	11.9117589	Standard Error	10.42539634	Standard I
5	Median	122.5	Median	71	Median	206.5	Median	169.5	Median
6	Mode	#N/A	Mode	81	Mode	227	Mode	#N/A	Mode
7	Standard Deviation	11.0835213	Standard Deviation	7.242620766	Standard Deviation	37.66828905	Standard Deviation	32.96799795	Standard I
8	Sample Variance	122.8444444	Sample Variance	52.45555556	Sample Variance	1418.9	Sample Variance	1086.888889	Sample V
9	Kurtosis	-0.129625435	Kurtosis	-1.590031149	Kurtosis	0.220968189	Kurtosis	-0.838859797	Kurtosis
10	Skewness	0.064730195	Skewness	0.09791646	Skewness	0.33799407	Skewness	0.612571519	Skewness
11	Range	36	Range	19	Range	125	Range	97	Range
12	Minimum	105	Minimum	62	Minimum	150	Minimum	138	Minimum
13	Maximum	141	Maximum	81	Maximum	275	Maximum	235	Maximum
14	Sum	1212	Sum	713	Sum	2023	Sum	1760	Sum
15	Count	10	Count	10	Count	10	Count	10	Count

FIGURE 4–7 First Quartile of DBP


Microsoft Excel - Chapter4
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
10
Reply with Changes... End Review...

C13 =QUARTILE(C2:C11,1)

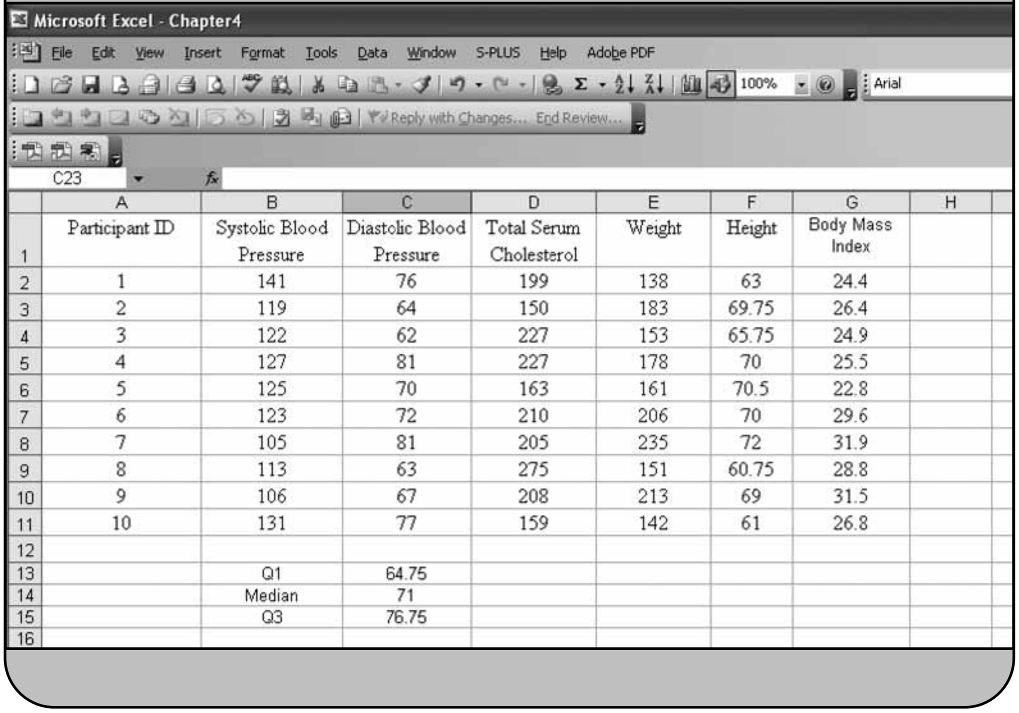
	A	B	C	D	E	F	G	H	I
1	Participant ID	Systolic Blood Pressure	Diastolic Blood Pressure	Total Serum Cholesterol	Weight	Height	Body Mass Index		
2	1	141	76	199	138	63	24.4		
3	2	119	64	150	183	69.75	26.4		
4	3	122	62	227	153	65.75	24.9		
5	4	127	81	227	178	70	25.5		
6	5	125	70	163	161	70.5	22.8		
7	6	123	72	210	206	70	29.6		
8	7	105	81	205	235	72	31.9		
9	8	113	63	275	151	60.75	28.8		
10	9	106	67	208	213	69	31.5		
11	10	131	77	159	142	61	26.8		
12									
13		Q1	64.75						
14									
15									

algorithm to determine the quartiles that involves interpolation. When we compute quartiles by hand as described in Chapter 4 of the textbook, we do not interpolate, and thus Excel may produce different values for the quartiles than those we determine by hand. Results are not likely to be different when the sample size is large.

Other Excel functions that are useful for generating summary statistics for continuous variables are shown in Table 4–4. For each function, the data (i.e., the range of cells in the worksheet containing the data to be analyzed) are specified in parentheses. The exception is the QUARTILE function, which also requires specification of the quartile number.

To compute the standard deviation of the systolic blood pressures, we enter "=STDEV(B2:B11)". Similarly, we can compute the median systolic blood pressure using the median function, "=MEDIAN(B2:B11)". Note that "=MEDIAN(B2:B11)" is equivalent to "=QUARTILE(B2:B11,2)". In Excel and in most other statistical computing packages, there are often several

FIGURE 4–8 Second and Third Quartiles of DBP



	A	B	C	D	E	F	G	H
1	Participant ID	Systolic Blood Pressure	Diastolic Blood Pressure	Total Serum Cholesterol	Weight	Height	Body Mass Index	
2	1	141	76	199	138	63	24.4	
3	2	119	64	150	183	69.75	26.4	
4	3	122	62	227	153	65.75	24.9	
5	4	127	81	227	178	70	25.5	
6	5	125	70	163	161	70.5	22.8	
7	6	123	72	210	206	70	29.6	
8	7	105	81	205	235	72	31.9	
9	8	113	63	275	151	60.75	28.8	
10	9	106	67	208	213	69	31.5	
11	10	131	77	159	142	61	26.8	
12								
13		Q1	64.75					
14		Median	71					
15		Q3	76.75					
16								

TABLE 4–4 Excel Functions for Summary Statistics

Function	Description
AVERAGE	Computes the sample mean
COUNT	Computes the sample size
MAX	Computes the maximum value
MEDIAN	Computes the sample median
MIN	Computes the minimum value
MODE	Computes the mode
QUARTILE	Computes the quartiles
STDEV	Computes the sample standard deviation
VAR	Computes the sample variance

ways to perform the same analysis. The analyst can choose the method with which they are the most comfortable or that best suits their style.

4.3 PRACTICE PROBLEMS

1. A study is conducted to estimate the mean total cholesterol level in children 2 to 6 years of age. A sample of nine participants is selected and their total cholesterol levels are measured as follows.

   185 225 240 196 175 180 194 147 223

   a. Use the Data Analysis ToolPak to compute the sample mean, standard deviation, and median.
   b. Use the QUARTILE function to compute the first and third quartiles.

2. The following data were collected as part of a study of coffee consumption among graduate students. The following reflect cups per day consumed:

   3 4 6 8 2 1 0 2

   a. Use the Data Analysis ToolPak to compute the sample mean, standard deviation, and median.
   b. Use the QUARTILE function to compute the first and third quartiles.

3. In the study of a new antihypertensive medication, systolic blood pressures are measured at baseline (or the start of the study before any treatment is administered). The data are as follows:

   120 112 138 145 135 150 145 163

   148 128 143 156 160 142 150

   Use Excel functions to compute the sample mean, standard deviation, median, and quartiles.

4. The following are height measurements (in cm) for a sample of infants participating in a study of infant health:

   28 30 41 48 29 48 62 49 51 39

   Use the Data Analysis ToolPak to compute summary statistics.

CHAPTER 5

Working with Probability Functions

Microsoft Office® Excel® has a number of probability functions that can be used to compute probabilities or to find percentiles. In Chapter 5 of the textbook, we discussed in detail two probability models, the binomial and normal distributions, which are appropriate for discrete and continuous outcomes, respectively. As we discussed in Chapter 5 of the textbook, there are many other probability distributions that describe discrete and continuous outcomes; we focused exclusively on these two. Excel has a number of functions that can be used to compute probabilities for various distributions. We focus on the binomial and normal distributions.

5.1 COMPUTING PROBABILITIES WITH THE BINOMIAL DISTRIBUTION

In Chapter 5 of the textbook, we discussed the binomial distribution model and computed probabilities using the model:

$$\mathrm{P}(x\ \text{successes}) = \frac{n!}{x!(n-x)!}\,p^{x}\,(1-p)^{n-x}$$

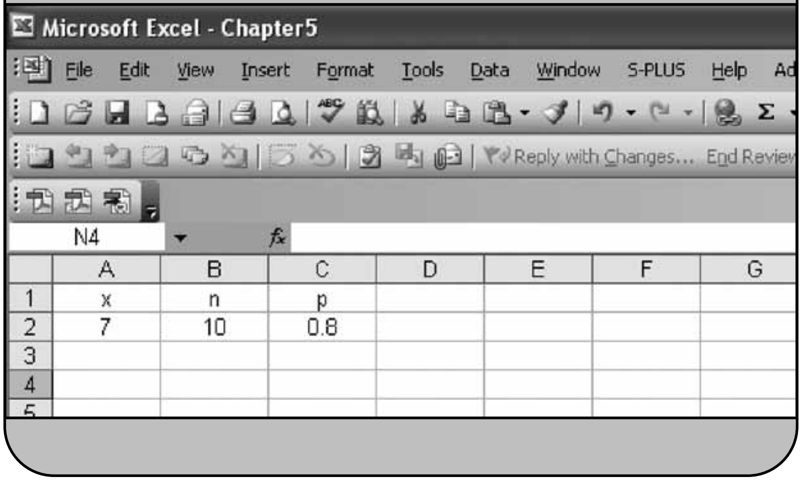
where n denotes the number of times the application or process is repeated (sometimes called the number of trials), x denotes the number of successes out of n trials of interest, and p is the probability of success for any individual.

To use the binomial distribution model, we need to specify n, p, and x. Excel has a probability function to compute probabilities from a binomial distribution. The function is invoked as "=BINOMDIST(x, n, p, *cumulative*)". The first three inputs for the function are the same as those we use in computing probabilities by hand with the binomial distribtion model (i.e., x, n, and p). Excel requires one additional input, labeled *cumulative*. This last entry in the BINOMDIST function is a logical value (i.e., one whose responses are true or false). We either use the cumulative distribution function by specifying "true" or do not use the cumulative distribution function by specifying "false." The cumulative distribution function returns the probability of observing x or fewer (rather than exactly x)successes. For example, if we specify "true" and indicate $x = 5$ in the function, then Excel computes $\mathrm{P}(X \le 5)$. In contrast, if we specify "false" and indicate $x = 5$ in the function, then Excel computes $\mathrm{P}(X = 5)$. We illustrate the use of the function in the following example.

Example 5.1. In Example 5.9 in the textbook, we presented an example assessing the extent to which adults with allergies report relief from allergic symptoms with a specific medication. We know that the medication is effective in 80% of patients with allergies. If we provide the medication to 10 patients with allergies, what is the probability that it is effective in exactly 7?

For this example, $n = 10$, $p = 0.80$, and $x = 7$. We now use Excel to compute the desired probability. In Figure 5–1, we enter n, p, and x into an Excel worksheet. We wish to compute the probability of 7 successes when $n = 10$ and the probability of success for any individual is 0.80. We use the BINOMDIST function and specify the cell locations for x (A2), n (B2), and p (C2). Because we want to compute the probability of exactly 7 successes, we set *cumulative* = false. The specification of the formula is shown in Figure 5–2.

FIGURE 5–1 Data for the Binomial Distribution Model


Microsoft Excel - Chapter5
File Edit View Insert Format Tools Data Window S-PLUS Help Ad
Reply with Changes... End Review
N4

	A	B	C	D	E	F	G
1	x	n	p				
2	7	10	0.8				
3							
4							

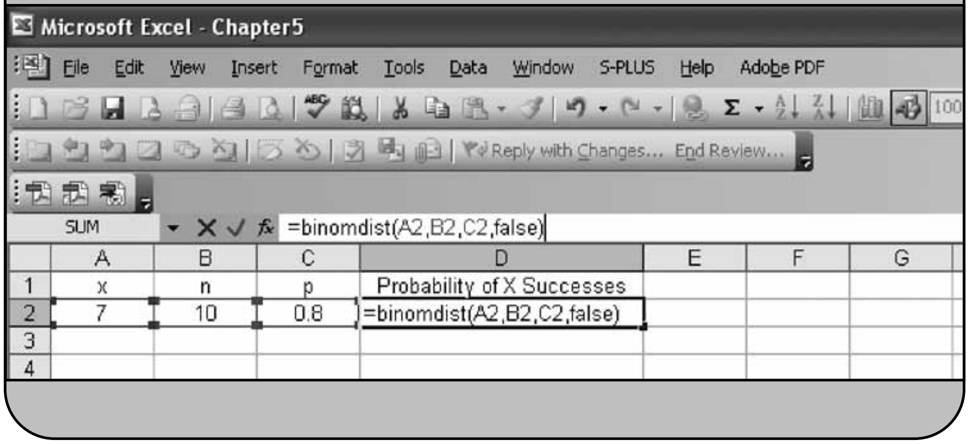
Once we enter these values, the probability is computed. The result is shown in Figure 5–3. We can also compute the probability of observing 7 or fewer successes by specifying *cumulative* = true in the function call. This is shown in Figure 5–4. The result is shown in Figure 5–5. Thus, P(X = 7) = 0.201 and P($X \leq 7$) = 0.322.

In Figure 5–6, we enter all possible values of x for n = 10 (i.e., 0 through 10) and in each case, compute the probability of exactly x successes as well as the probability of observing x successes or fewer, by using the BINOMDIST function with *cumulative* = false and *cumulative* = true, respectively. Notice that some of the probabilities are very small—e.g., P(X = 0) = 1.024E−07 = 0.0000001024—and also note the relationship between the individual and cumulative probabilities.

5.2 COMPUTING PROBABILITIES WITH THE NORMAL DISTRIBUTION

In Chapter 5 of the textbook, we presented the normal distribution and showed how to compute probabilities using the standard normal distribution (z) and Table 1 in the

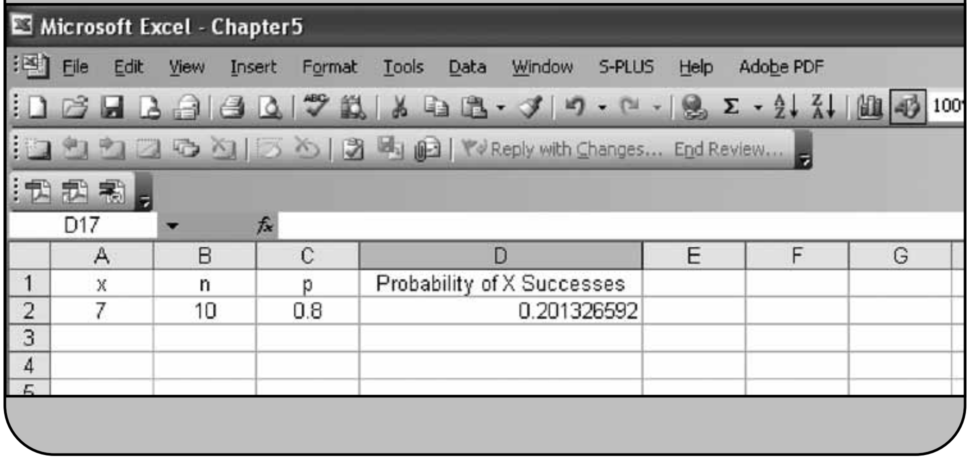
FIGURE 5–2 Requesting P(X = 7) When n = 10 and p = 0.8


Microsoft Excel - Chapter5
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...
SUM

=binomdist(A2,B2,C2,false)

	A	B	C	D	E	F	G
1	x	n	p	Probability of X Successes			
2	7	10	0.8	=binomdist(A2,B2,C2,false)			
3							
4							

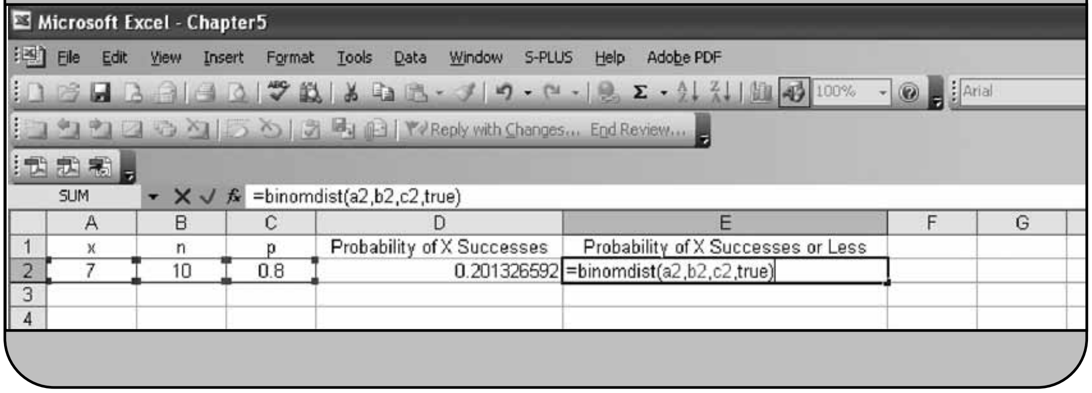
FIGURE 5–3 P(X = 7) When n = 10 and p = 0.8


Microsoft Excel - Chapter5
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

D17 f_x

	A	B	C	D	E	F	G
1	x	n	p	Probability of X Successes			
2	7	10	0.8	0.201326592			
3							
4							
5							

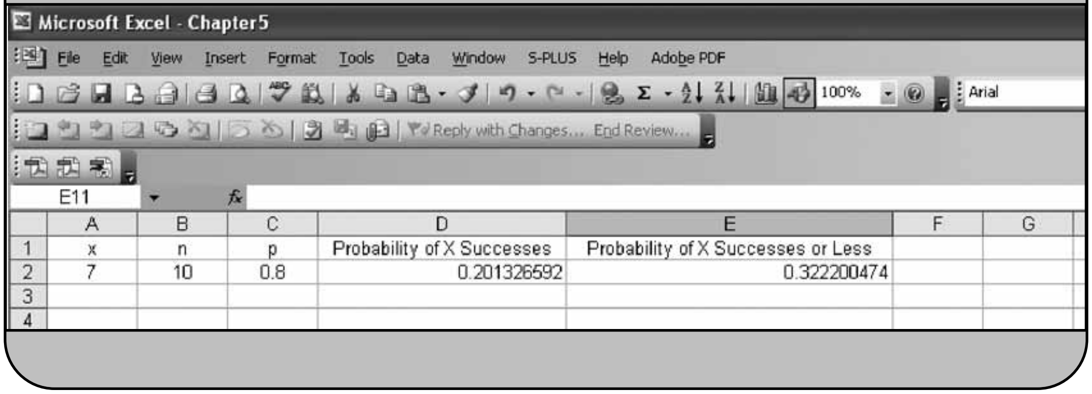
FIGURE 5–4 Requesting P($X \leq$ 7) When n = 10 and p = 0.8


Microsoft Excel - Chapter5
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...

SUM f_x =binomdist(a2,b2,c2,true)

	A	B	C	D	E	F	G
1	x	n	p	Probability of X Successes	Probability of X Successes or Less		
2	7	10	0.8	0.201326592	=binomdist(a2,b2,c2,true)		
3							
4							

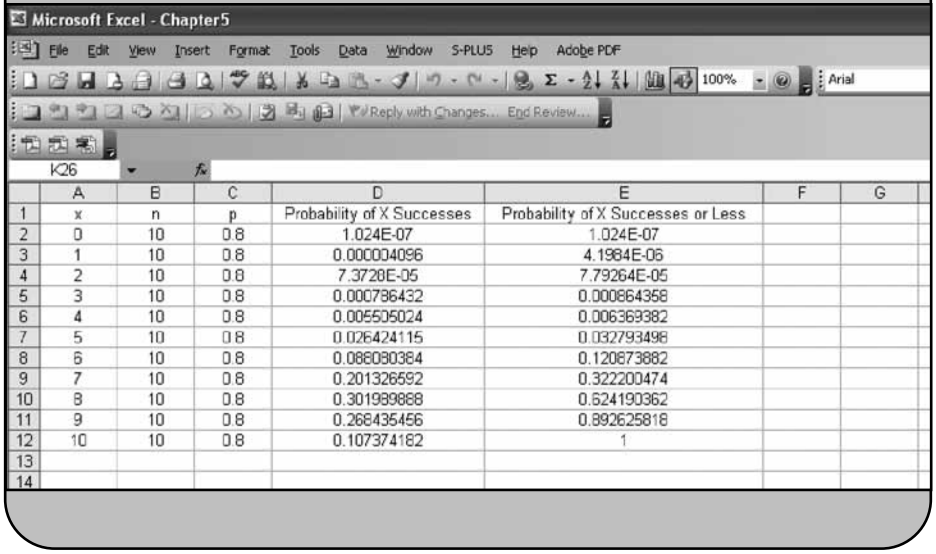
FIGURE 5–5 P($X \leq$ 7) When n = 10 and p = 0.8


Microsoft Excel - Chapter5
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...

E11 f_x

	A	B	C	D	E	F	G
1	x	n	p	Probability of X Successes	Probability of X Successes or Less		
2	7	10	0.8	0.201326592	0.322200474		
3							
4							

FIGURE 5–6 Binomial Probability Distribution for $n = 10$ and $p = 0.8$


Microsoft Excel - Chapter5
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...

K26

	A	B	C	D	E	F	G
1	x	n	p	Probability of X Successes	Probability of X Successes or Less		
2	0	10	0.8	1.024E-07	1.024E-07		
3	1	10	0.8	0.000004096	4.1984E-06		
4	2	10	0.8	7.3728E-05	7.79264E-05		
5	3	10	0.8	0.000786432	0.000864358		
6	4	10	0.8	0.005505024	0.006369382		
7	5	10	0.8	0.026424115	0.032793498		
8	6	10	0.8	0.088080384	0.120873882		
9	7	10	0.8	0.201326592	0.322200474		
10	8	10	0.8	0.301989888	0.624190362		
11	9	10	0.8	0.268435456	0.892625818		
12	10	10	0.8	0.107374182	1		
13							
14							

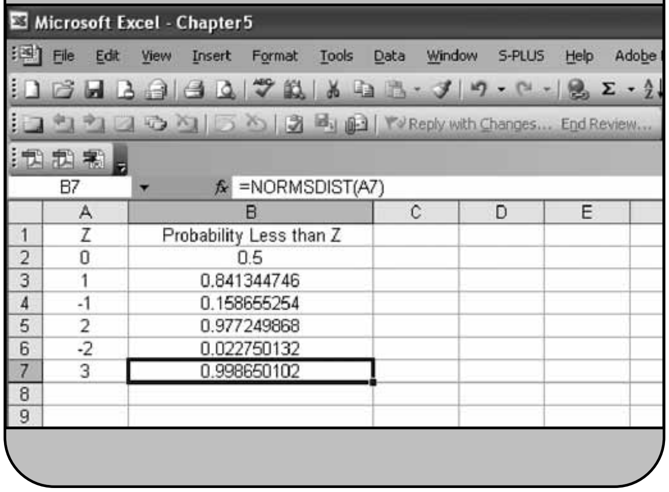
Appendix of the textbook. Here we use Excel to compute probabilities for variables that are assumed to follow a normal distribution.

Excel has two probability functions to compute probabilities from normal distributions. The first computes probabilities from the standard normal distribution. This function is "=NORMSDIST(*z*)". The input for the function is the *z* value. The function returns the probability of observing a value from the standard normal distribution less than or equal to *z* (i.e., Excel returns the area under the standard normal curve to the left of—or less than or equal to—*z*). Figure 5–7 displays a worksheet with various values of *z* entered in column A. The NORMSDIST function is used in column B to generate probabilities less than or equal to these values of *z*. For example, in cell B2 we enter "=NORMSDIST(A2)".

Recall that zero is the mean (and median) of the standard normal distribution, and thus $P(z \le 0) = 0.5$. (Recall that for the normal distribution that $P(z \le 0) = P(z < 0) = 0.5$.) In addition, the standard deviation is 1, so $P(-1 < z < 1) = P(z < 1) - P(z < -1) = 0.841 - 0.159 = 0.682$ (68% of the observations in a normal distribution fall between the mean minus one standard deviation and the mean plus one standard deviation).

In most applications, we analyze distributions that are normal with mean μ and standard deviation σ (which are not 0 and 1, respectively). Excel has a second function that computes probabilities from any normal distribution. This function is "=NORMDIST(x, μ, σ, *cumulative*)". The inputs for the function are the x value, the mean (μ) and standard deviation (σ) of the normal distribution, and a logical value labeled *cumulative*. For almost all applications, we use the cumulative distribution function by specifying *cumulative*=true. This returns the probability of observing a value from the normal distribution with specified mean and standard deviation that is less than or equal to x (i.e., Excel returns the area under the curve to the left of—or less than or equal to—x). If we specify *cumulative* = false, the function returns the probability of observing the value x. Figure 5–8 displays a worksheet

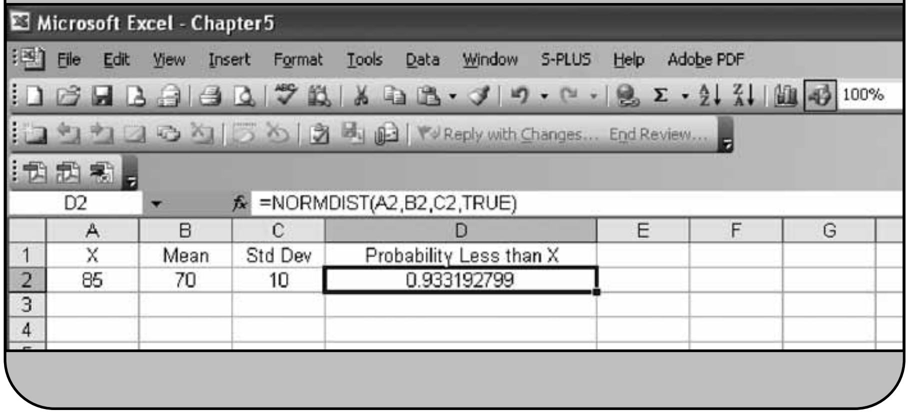
FIGURE 5–7 Probabilities From the Standard Normal Distribution


Microsoft Excel - Chapter5
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe
Reply with Changes... End Review...

B7 =NORMSDIST(A7)

	A	B	C	D	E
1	Z	Probability Less than Z			
2	0	0.5			
3	1	0.841344746			
4	-1	0.158655254			
5	2	0.977249868			
6	-2	0.022750132			
7	3	0.998650102			
8					
9					

FIGURE 5–8 P($X < 85$) When $\mu = 70$ and $\sigma = 10$


Microsoft Excel - Chapter5
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

D2 =NORMDIST(A2,B2,C2,TRUE)

	A	B	C	D	E	F	G
1	X	Mean	Std Dev	Probability Less than X			
2	85	70	10	0.933192799			
3							
4							

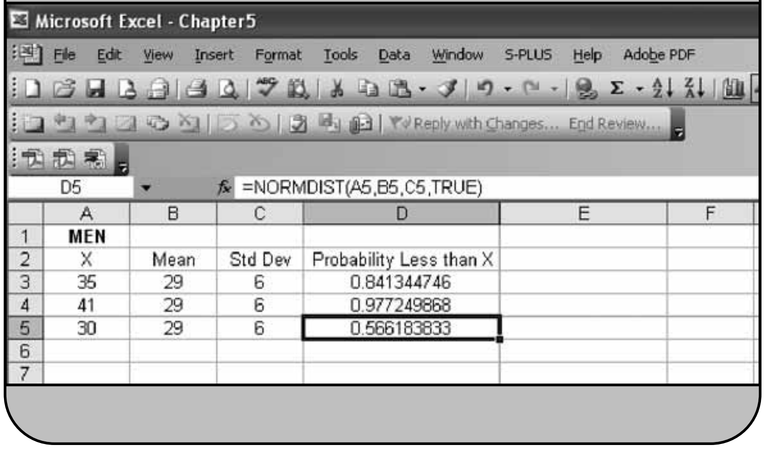
with *x* entered in column A. We specify the mean and standard deviation of the distribution in column B and column C, respectively, and use the NORMDIST function to generate probabilities less than or equal to *x*. For example, in cell D2 we enter "=NORMDIST(A2,B2,C2,true)". For a normal distribution with $\mu = 70$ and standard $\sigma = 10$, P($X < 85$) = 0.9332.

Example 5.2. In Example 5.11 in the textbook, we analyzed body mass index (BMI), which is assumed to be normally distributed for specific gender and age groups. The mean BMI for men aged 60 is 29 with a standard deviation of 6, and for women aged 60 the mean is 28 with a standard deviation of 7.

For men aged 60, we now use Excel to compute the following: P($X < 35$), P($X < 41$), and P($X < 30$). The results are shown in Figure 5–9. Suppose we now wish to compute the probability that a male has a BMI between 30 and 35—i.e., P($30 < X < 35$). This can be done using the NORMDIST function to compute the probabilities that a male has BMI less than 35 and less than 30 (as in Figure 5–9) and subtracting (Figure 5–10) to compute the desired probability: P($30 < X < 35$) = P($X < 35$) − P($X < 30$) = 0.275.

Now consider BMI in women. What are the probabilities that a female aged 60 has a BMI less than 30 and less than 35? What is the probability that a female aged 60 has a BMI between 30 and 35? We use the same approach, but recall for women aged 60, the mean is 28 and the standard deviation is 7. The results are shown in Figure 5–11.

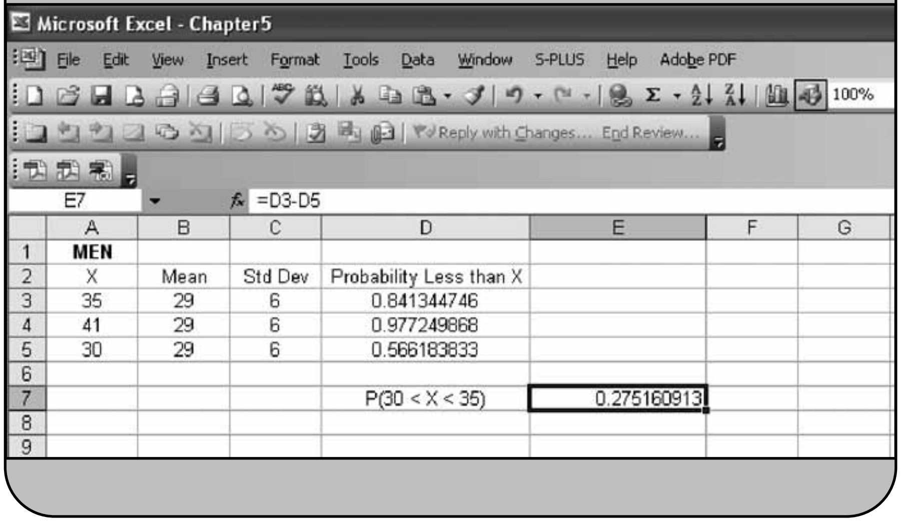
FIGURE 5–9 Probabilities from the Normal Distribution Using the NORMSDIST Function


Microsoft Excel - Chapter5
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

D5 =NORMDIST(A5,B5,C5,TRUE)

	A	B	C	D	E	F
1	**MEN**					
2	X	Mean	Std Dev	Probability Less than X		
3	35	29	6	0.841344746		
4	41	29	6	0.977249868		
5	30	29	6	0.566183833		
6						
7						

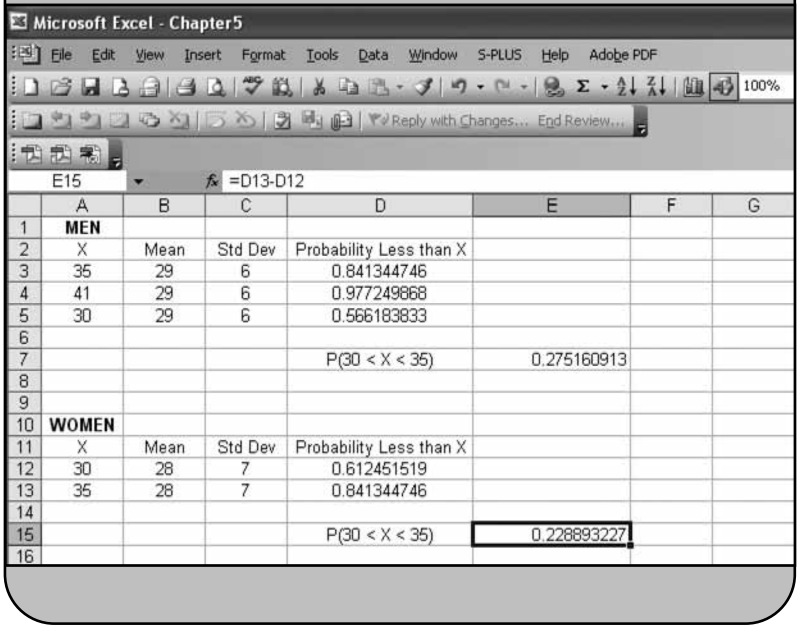
FIGURE 5–10 P(30 < *X* < 35) When $\mu = 29$ and $\sigma = 6$


Microsoft Excel - Chapter5
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

E7 =D3-D5

	A	B	C	D	E	F	G
1	MEN						
2	X	Mean	Std Dev	Probability Less than X			
3	35	29	6	0.841344746			
4	41	29	6	0.977249868			
5	30	29	6	0.566183833			
6							
7				P(30 < X < 35)	0.275160913		
8							
9							

FIGURE 5–11 P(30 < *X* < 35) When $\mu = 328$ and $\sigma = 7$


Microsoft Excel - Chapter5
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

E15 =D13-D12

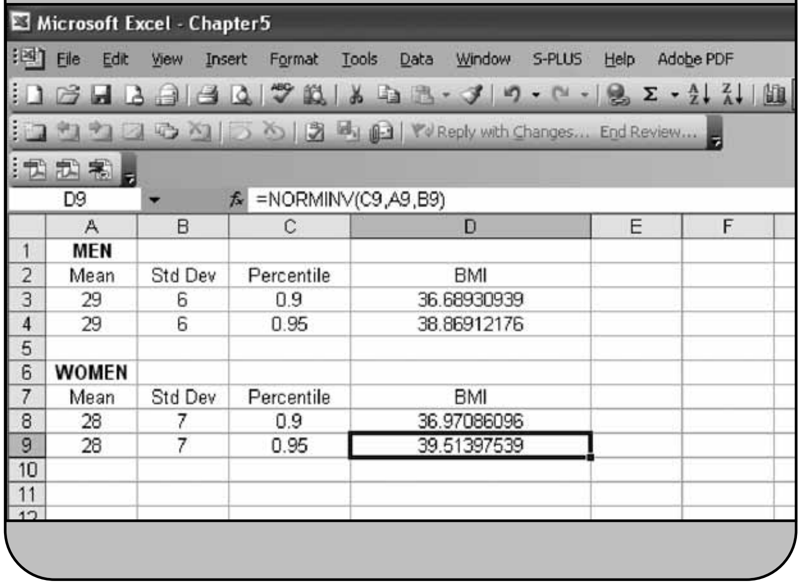
	A	B	C	D	E	F	G
1	MEN						
2	X	Mean	Std Dev	Probability Less than X			
3	35	29	6	0.841344746			
4	41	29	6	0.977249868			
5	30	29	6	0.566183833			
6							
7				P(30 < X < 35)	0.275160913		
8							
9							
10	WOMEN						
11	X	Mean	Std Dev	Probability Less than X			
12	30	28	7	0.612451519			
13	35	28	7	0.841344746			
14							
15				P(30 < X < 35)	0.228893227		
16							

5.3 FINDING PERCENTILES OF THE NORMAL DISTRIBUTION

Recall from Chapter 5 in the textbook that a *percentile* is a score that holds a specified percentage or proportion of scores below it. For example, the 80th percentile is the score that holds 80% of the scores below it. Excel can also be used to compute percentiles for the standard normal distribution and for any normal distribution. The two functions are NORMSINV and NORMINV, respectively. The inputs for the functions are shown in the following text.

The function to compute percentiles for the standard normal distribution is invoked as "=NORMSINV(*probabil-*

FIGURE 5–12 Percentiles of BMI


Microsoft Excel - Chapter5
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

D9 =NORMINV(C9,A9,B9)

	A	B	C	D	E	F
1	MEN					
2	Mean	Std Dev	Percentile	BMI		
3	29	6	0.9	36.68930939		
4	29	6	0.95	38.86912176		
5						
6	WOMEN					
7	Mean	Std Dev	Percentile	BMI		
8	28	7	0.9	36.97086096		
9	28	7	0.95	39.51397539		
10						
11						

ity)". The input for the function is the desired percentile, entered as a probability or proportion. For example, to compute the 80th or 95th percentiles, we specify 0.80 or 0.95, respectively.

The function to compute percentiles for any normal distribution is invoked as "=NORMINV(*probability*, μ, σ)". The inputs for the function are the desired percentile entered as a probability and the mean (μ) and standard deviation (σ) of the normal distribution.

Example 5.3. Using the data in Example 5.2, we now use Excel to compute the 90th and 95th percentiles of BMI for men and women. The results are shown in Figure 5–12. In men, 90% of BMIs are below 36.7 and 95% are below 38.9. In women, 90% of BMIs are below 37.0 and 95% are below 39.5.

5.4 PRACTICE PROBLEMS

1. Total cholesterol in children aged 10 to 15 years of age is assumed to follow a normal distribution with a mean of 191 and a standard deviation of 22.4.
   a. What proportion of children 10 to 15 years of age have total cholesterol between 180 and 190?
   b. What proportion of children 10 to 15 years of age would be classified as hyperlipidemic (assume that hyperlipidemia is defined as a total cholesterol level over 200)?
   c. What is the 90th percentile of total cholesterol?
2. Among coffee drinkers, men drink a mean of 3.2 cups per day with a standard deviation of 0.8 cups. Assume the number of coffee drinks per day follows a normal distribution.
   a. What proportion drink 2 cups per day or more?
   b. What proportion drink no more than 4 cups per day?
   c. If the top 5% of coffee drinkers are considered heavy coffee drinkers, what is the minimum number of cups consumed by a heavy coffee drinker? (*Hint*: Find the 95th percentile.)
3. A study is conducted to assess the impact of caffeine consumption, smoking, alcohol consumption, and physical activity on the risk of cardiovascular disease. Suppose that 40% of participants consume caffeine and smoke. If 8 participants are evaluated, what is the probability that:
   a. Exactly half of them consume caffeine and smoke?
   b. At most 6 consume caffeine and smoke?
4. A recent study of cardiovascular risk factors reports that 30% of adults meet the criteria for hypertension. If 15 adults are assessed, what is the probability that:
   a. Exactly 5 meet the criteria for hypertension?
   b. None meet the criteria for hypertension?
   c. Less than or equal to 7 meet the criteria for hypertension?

5. Diastolic blood pressures are assumed to follow a normal distribution with a mean of 85 and a standard deviation of 12.
   a. What proportion of people have diastolic blood pressures less than 90?
   b. What proportion have diastolic blood pressures between 80 and 90?
   c. If someone has a diastolic blood pressure of 100, what percentile does this represent?

CHAPTER 6

Confidence Interval Estimates

In Chapter 6 of the textbook, we presented formulas to generate confidence intervals for means (μ) and proportions (p) in one sample and for differences in means ($\mu_1 - \mu_2$) and differences in proportions ($p_1 - p_2$) in two independent samples. We also discussed confidence intervals for the mean difference (μ_d) when two samples are matched or paired. For each application, we used the same general approach. Confidence intervals for each parameter take the following form:

Point estimate ± *Margin of error*

The point estimate depends on the parameter being estimated. For example, when estimating the mean of a population, μ, the point estimate is the sample mean, $\overline{X}$. When estimating the population proportion, p, the point estimate is the sample proportion, $\hat{p}$. The margin of error includes two components. The first component is from a probability distribution (e.g., z or t) and reflects the selected confidence level (e.g., 90%, 95%) and the second component is the standard error of the point estimate. For example, the standard error of the sample mean, $\overline{X}$, is $SE = s/\sqrt{n}$. The standard error of the sample proportion, $\hat{p}$, is $SE = \sqrt{\hat{p}(1-\hat{p})/n}$.

In Chapter 6 of the textbook, we presented formulas for confidence intervals for various parameters (see Table 6–22 in Section 6.7 of the textbook for details). Here we use Microsoft Office® Excel® to generate confidence intervals for various parameters. Excel does not generate confidence intervals directly; instead, we use specific Excel functions to produce the z or t values that reflect the desired confidence level and then construct the confidence interval.

6.1 CONFIDENCE INTERVALS FOR ONE SAMPLE, CONTINUOUS OUTCOME

In Chapter 6 of the textbook, we presented the following formulas for confidence intervals for the mean of a continuous variable in one sample.

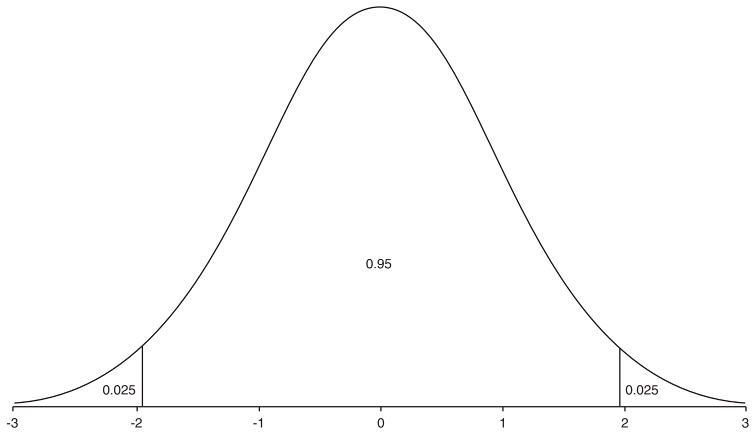
$$n \geq 30: \ \overline{X} \pm z \frac{s}{\sqrt{n}} \text{ (Find } z \text{ in Table 1B)}$$

$$n < 30: \ \overline{X} \pm t \frac{s}{\sqrt{n}} \text{ (Find } t \text{ in Table 2, } df = n - 1\text{)}$$

When computing the confidence intervals by hand, we computed the sample size, mean, and standard deviation, and then used Table 1B or Table 2 in the Appendix of the textbook to find the appropriate z or t value to reflect the desired confidence level. Here we use Excel to compute summary statistics and to determine the appropriate z or t value for the confidence interval. Once all of the requisite components are determined, we construct the confidence interval.

Excel has a function to compute z values that can be used in confidence intervals. The function is "=NORMSINV (*lower tail area*)". To use this function for confidence intervals, we specify the area under the curve in the lower tail of the standard normal distribution. For example, for a 95% confidence interval, the area in the lower tail is 0.975. Figure 6–1 shows the standard normal distribution, z, and the z values that hold the middle 95% of the distribution, $P(-1.96 < X < 1.96) = 0.95$. To produce the z value for a 95% confidence interval, we specify "=NORMSINV(0.975)", which returns 1.96.

FIGURE 6–1 Z Value for 95% Confidence



Excel has a second function that computes *t* values that can also be used in confidence intervals. The function is "=TINV(*total tail area*, *df*)." To use this function for confidence intervals, we specify the *total area* in the tail of the *t* distribution along with degrees of freedom, *df*. For a 95% confidence interval, the total tail area is 0.05, and for one sample, the degrees of freedom are $df = n - 1$. We now illustrate the use of these formulas to compute confidence intervals.

Example 6.1. In Example 6.1 in the textbook, we analyzed data on $n = 3539$ participants who attended the seventh examination of the offspring in the Framingham Heart Study. Summary statistics on variables measured in the sample are shown in Table 6–1. We use Excel to generate 95% confidence intervals for each characteristic. Because the sample size is large, we use the confidence interval formula with *z* as opposed to *t*. The data are entered into an Excel worksheet as shown in Figure 6–2.

TABLE 6–1 Summary Statistics, Framingham Heart Study Offspring

Characteristic	*n*	Mean ($\overline{X}$)	Standard Deviation (*s*)
Systolic blood pressure	3534	127.3	19.0
Diastolic blood pressure	3532	74.0	9.9
Total serum cholesterol	3310	200.3	36.8
Weight (lb)	3506	174.4	38.7
Height (in)	3326	65.957	3.749
Body mass index (BMI)	3326	28.15	5.32

First, we compute the standard errors for each characteristic as SE = $s/\sqrt{n}$ and place these values in column E. For systolic blood pressure, the following is entered into cell E2: "=D2/SQRT(B2)". We then compute the *z* scores for 95% confidence intervals and place these in column F. The computation is the same for each characteristic—for example, in cell F2 we enter "=NORMSINV(0.975)". The standard errors and *z* values are shown in Figure 6–3. Note that had we wanted 90% confidence intervals, we would have specified "=NORMSINV(0.95)" in column F.

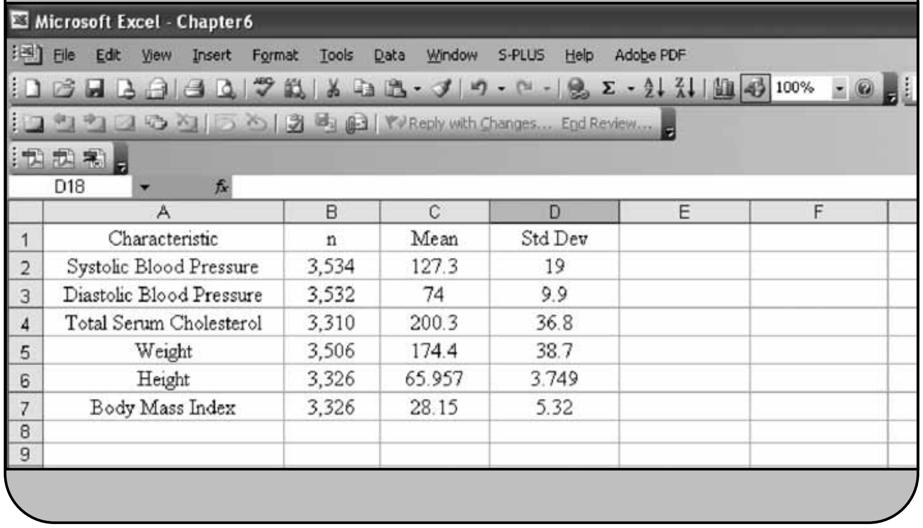
We now compute the lower and upper limits of the 95% confidence intervals using $\overline{X} \pm z \frac{s}{\sqrt{n}}$. For systolic blood pressure, the lower limit is "=C2 − (F2*E2)" and the upper limit is "=C2 + (F2*E2)". (The product of the *z* value and the standard error produce the margin of error.) The confidence intervals are shown in Figure 6–4.

Example 6.2. In Example 6.2 in the textbook, we presented data on a subsample of $n = 10$ participants who attended the seventh examination of the Framingham Offspring Study. Summary statistics on variables measured in the subsample are shown in Table 6–2.

We use Excel to generate 95% confidence intervals for each characteristic. Because the sample size is small, we use the confidence interval formula with *t* as opposed to *z*. The data are entered into an Excel worksheet as shown in Figure 6–5.

First we compute the standard errors for each characteristic as SE = $s/\sqrt{n}$ and place these values in column E—e.g.,

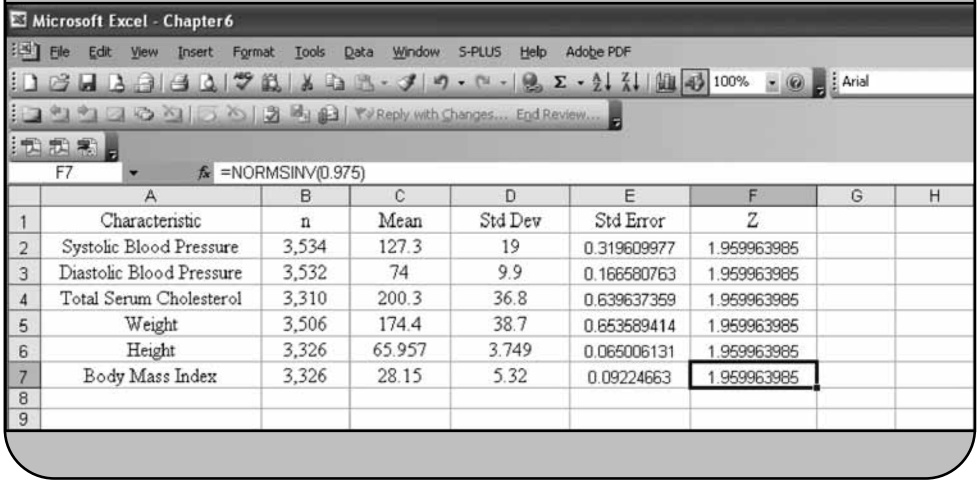
FIGURE 6–2 Data from Framingham Offspring Study


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

D18

	A	B	C	D	E	F
1	Characteristic	n	Mean	Std Dev		
2	Systolic Blood Pressure	3,534	127.3	19		
3	Diastolic Blood Pressure	3,532	74	9.9		
4	Total Serum Cholesterol	3,310	200.3	36.8		
5	Weight	3,506	174.4	38.7		
6	Height	3,326	65.957	3.749		
7	Body Mass Index	3,326	28.15	5.32		
8						
9						

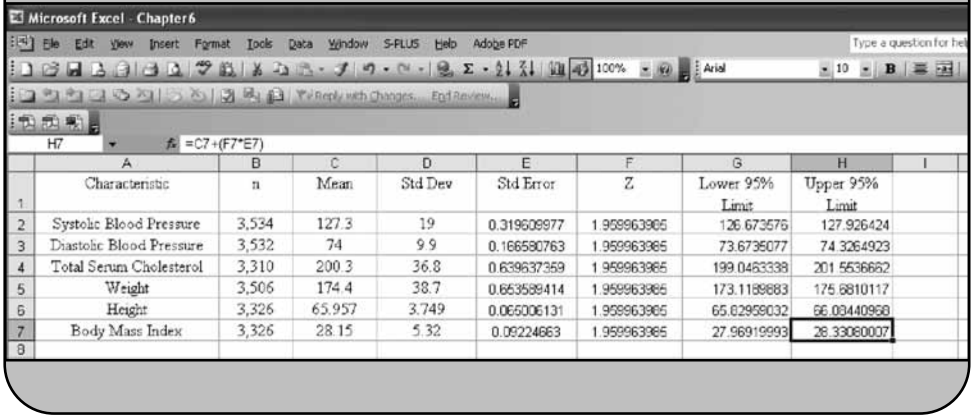
FIGURE 6–3 Standard Errors and *z* Values for Confidence Intervals


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...

F7 =NORMSINV(0.975)

	A	B	C	D	E	F	G	H
1	Characteristic	n	Mean	Std Dev	Std Error	Z		
2	Systolic Blood Pressure	3,534	127.3	19	0.319609977	1.959963985		
3	Diastolic Blood Pressure	3,532	74	9.9	0.166580763	1.959963985		
4	Total Serum Cholesterol	3,310	200.3	36.8	0.639637359	1.959963985		
5	Weight	3,506	174.4	38.7	0.653589414	1.959963985		
6	Height	3,326	65.957	3.749	0.065006131	1.959963985		
7	Body Mass Index	3,326	28.15	5.32	0.09224663	1.959963985		
8								
9								

FIGURE 6–4 Upper and Lower Limits of 95% Confidence Intervals


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a question for hel
100%
Arial
10
B
Reply with Changes... End Review...

H7 =C7+(F7*E7)

	A	B	C	D	E	F	G	H	I
1	Characteristic	n	Mean	Std Dev	Std Error	Z	Lower 95% Limit	Upper 95% Limit	
2	Systolic Blood Pressure	3,534	127.3	19	0.319609977	1.959963985	126.673576	127.926424	
3	Diastolic Blood Pressure	3,532	74	9.9	0.166580763	1.959963985	73.6735077	74.3264923	
4	Total Serum Cholesterol	3,310	200.3	36.8	0.639637359	1.959963985	199.0463338	201.5536662	
5	Weight	3,506	174.4	38.7	0.653589414	1.959963985	173.1189883	175.6810117	
6	Height	3,326	65.957	3.749	0.065006131	1.959963985	65.82959032	66.08440968	
7	Body Mass Index	3,326	28.15	5.32	0.09224663	1.959963985	27.96919993	28.33080007	
8									

TABLE 6–2 Summary Statistics, Framingham Heart Study Offspring Subsample

Characteristic	n	Mean ($\overline{X}$)	Standard Deviation (s)
Systolic blood pressure	10	121.2	11.1
Diastolic blood pressure	10	71.3	7.2
Total serum cholesterol	10	202.3	37.7
Weight (lb)	10	176.0	33.0
Height (in)	10	67.175	4.205
Body mass index (BMI)	10	27.26	3.10

for systolic blood pressure, the following is entered into cell E2: "=D2/SQRT(B2)". We then compute the t scores for 95% confidence intervals and place these in column F. The computation is the same for each characteristic—for example, in cell F2 we enter "=TINV(0.05,B2−1)". The standard errors and t values are shown in Figure 6–6.

We now compute the lower and upper limits of the 95% confidence intervals using $\overline{X} \pm t\frac{s}{\sqrt{n}}$. For systolic blood pressure, the lower limit is "=C2 − (F2*E2)" and the upper limit is "=C2 + (F2*E2)". The confidence intervals are shown in Figure 6–7.

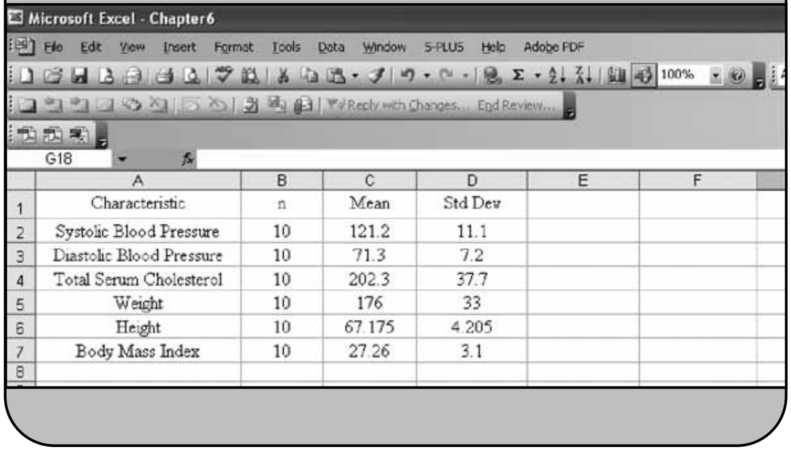
Example 6.3. In Example 4.3 in the textbook, we presented data on the subset of $n = 10$ participants, which were summarized in Example 6.2. The data values were presented in Table 4–13 of the textbook and are shown in Table 6–3. The data are entered into an Excel worksheet and shown in Figure 6–8.

In Chapter 4 of the Excel workbook, we described the Data Analysis ToolPak and generated summary statistics on continuous variables using the Descriptive Statistics Analysis Tool. This tool can be used to generate summary statistics and also to generate information that can be used to produce a confidence interval. Suppose we wish to generate descriptive statistics on systolic blood pressures (SBP). Using the Data Analysis ToolPak and selecting the Descriptive Statistics Module produces the dialog box shown in Figure 6–9.

In the dialog box, we specify the range of the data, we request that the results be placed in a new worksheet entitled *Descriptives on SBP*, and we request summary statistics and confidence interval information for a 95% confidence level. In the **Confidence Level for Mean** input field, any level from 0% to 100% can be specified (typical values are 90%, 95%, and 99%). The information in the new worksheet *Descriptives on SBP* is shown in Figure 6–10.

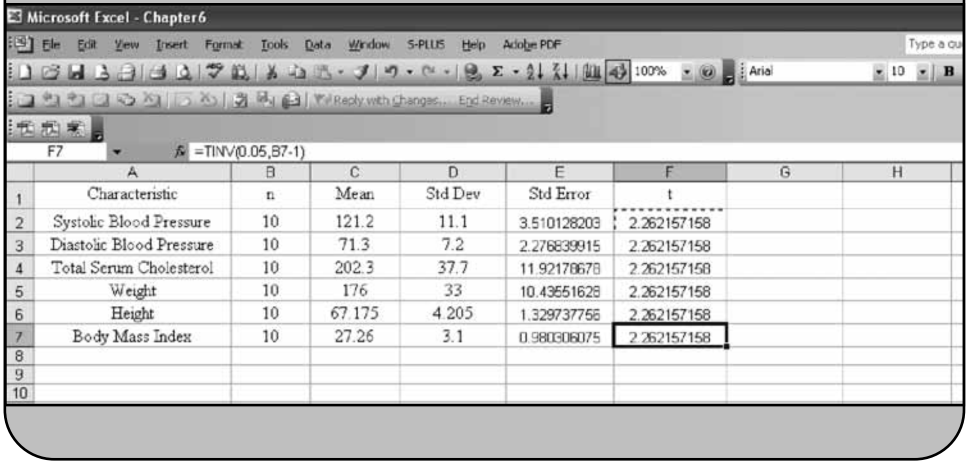
The information in cell B16 is the margin of error (i.e., the product of the t value for 95% confidence and the standard error). We must now take the margin of error and add it to and subtract it from the sample mean (point estimate) to produce the confidence limits. This is done and shown in Figure 6–11. Notice that the confidence limits shown in Figure 6–11 are identical to those shown in Figure 6–7.

FIGURE 6–5 Data From Framingham Offspring Subsample


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...
G18
fx

	A	B	C	D	E	F
1	Characteristic	n	Mean	Std Dev		
2	Systolic Blood Pressure	10	121.2	11.1		
3	Diastolic Blood Pressure	10	71.3	7.2		
4	Total Serum Cholesterol	10	202.3	37.7		
5	Weight	10	176	33		
6	Height	10	67.175	4.205		
7	Body Mass Index	10	27.26	3.1		
8						

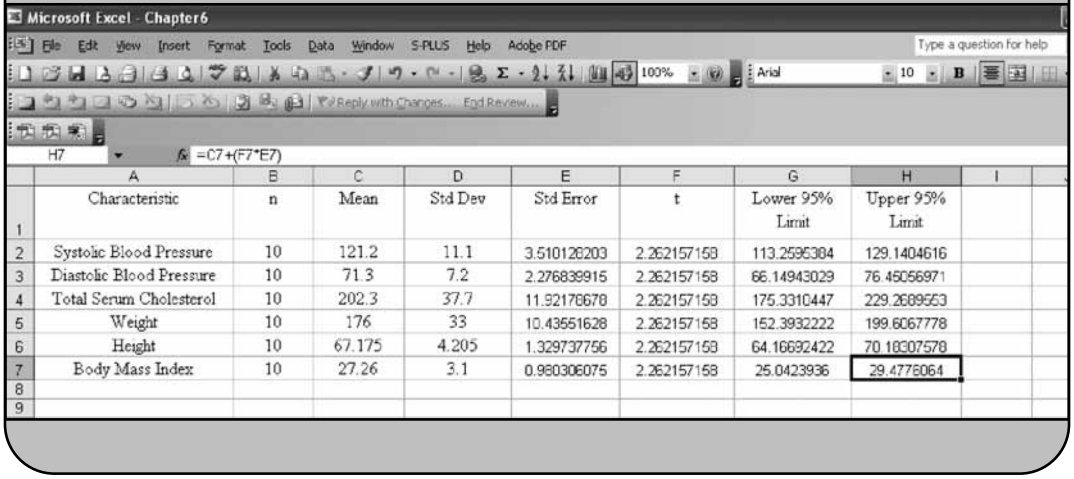
FIGURE 6–6 Standard Errors and *t* Values for Confidence Intervals


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a qu
100%
Arial
10
B
Reply with Changes... End Review...

F7 =TINV(0.05,B7-1)

	A	B	C	D	E	F	G	H
1	Characteristic	n	Mean	Std Dev	Std Error	t		
2	Systolic Blood Pressure	10	121.2	11.1	3.510128203	2.262157158		
3	Diastolic Blood Pressure	10	71.3	7.2	2.276839915	2.262157158		
4	Total Serum Cholesterol	10	202.3	37.7	11.92178678	2.262157158		
5	Weight	10	176	33	10.43551628	2.262157158		
6	Height	10	67.175	4.205	1.329737756	2.262157158		
7	Body Mass Index	10	27.26	3.1	0.980306075	2.262157158		
8								
9								
10								

FIGURE 6–7 Upper and Lower Limits of 95% Confidence Intervals


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a question for help
100%
Arial
10
B
Reply with Changes... End Review...

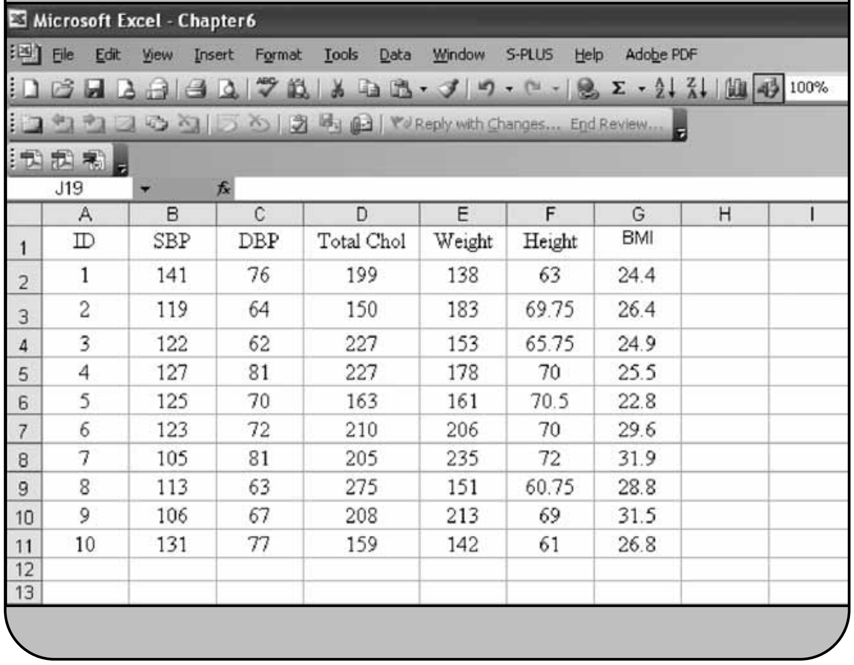
H7 =C7+(F7*E7)

	A	B	C	D	E	F	G	H	I
1	Characteristic	n	Mean	Std Dev	Std Error	t	Lower 95% Limit	Upper 95% Limit	
2	Systolic Blood Pressure	10	121.2	11.1	3.510128203	2.262157158	113.2595384	129.1404616	
3	Diastolic Blood Pressure	10	71.3	7.2	2.276839915	2.262157158	66.14943029	76.45056971	
4	Total Serum Cholesterol	10	202.3	37.7	11.92178678	2.262157158	175.3310447	229.2689553	
5	Weight	10	176	33	10.43551628	2.262157158	152.3932222	199.6067778	
6	Height	10	67.175	4.205	1.329737756	2.262157158	64.16692422	70.18307578	
7	Body Mass Index	10	27.26	3.1	0.980306075	2.262157158	25.0423936	29.4776064	
8									
9									

TABLE 6–3 Subsample of n = 10 Participants Attending the Seventh Examination of the Framingham Offspring Study

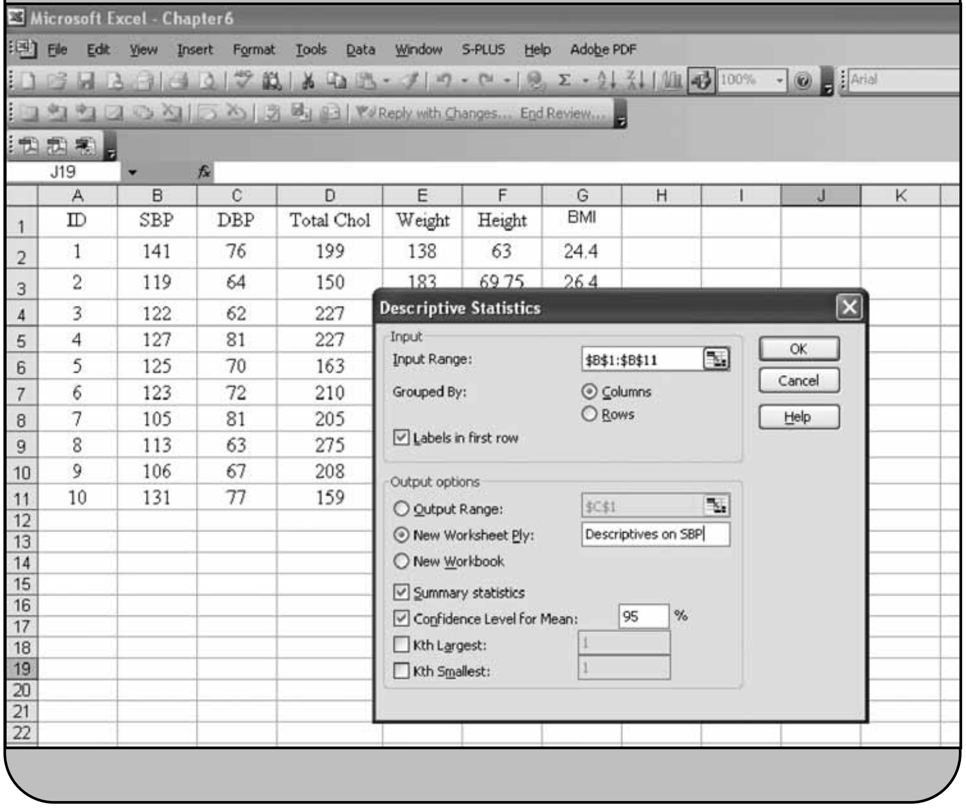
Participant ID	Systolic Blood Pressure	Diastolic Blood Pressure	Total Serum Cholesterol	Weight (lb)	Height (in)	BMI
1	141	76	199	138	63.00	24.4
2	119	64	150	183	69.75	26.4
3	122	62	227	153	65.75	24.9
4	127	81	227	178	70.00	25.5
5	125	70	163	161	70.50	22.8
6	123	72	210	206	70.00	29.6
7	105	81	205	235	72.00	31.9
8	113	63	275	151	60.75	28.8
9	106	67	208	213	69.00	31.5
10	131	77	159	142	61.00	26.8

FIGURE 6–8 Data Measured in Subsample of $n = 10$ Participants Attending Seventh Examination of the Framingham Offspring Study


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...
J19
fx

	A	B	C	D	E	F	G	H	I
1	ID	SBP	DBP	Total Chol	Weight	Height	BMI		
2	1	141	76	199	138	63	24.4		
3	2	119	64	150	183	69.75	26.4		
4	3	122	62	227	153	65.75	24.9		
5	4	127	81	227	178	70	25.5		
6	5	125	70	163	161	70.5	22.8		
7	6	123	72	210	206	70	29.6		
8	7	105	81	205	235	72	31.9		
9	8	113	63	275	151	60.75	28.8		
10	9	106	67	208	213	69	31.5		
11	10	131	77	159	142	61	26.8		
12									
13									

FIGURE 6–9 Descriptive Statistics Analysis Tool in Data Analysis ToolPak


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
J19
fx
OK
Cancel
Help

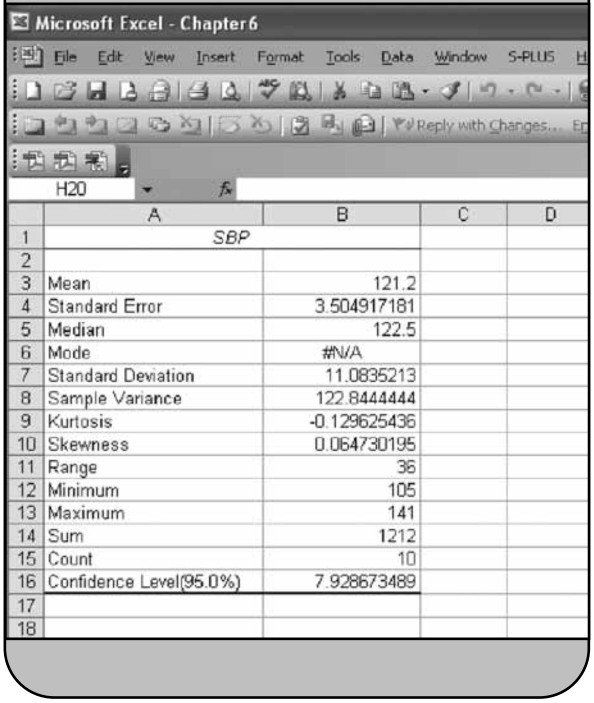
	A	B	C	D	E	F	G	H	I	J	K
1	ID	SBP	DBP	Total Chol	Weight	Height	BMI				
2	1	141	76	199	138	63	24.4				
3	2	119	64	150	183	69.75	26.4				
4	3	122	62	227							
5	4	127	81	227							
6	5	125	70	163							
7	6	123	72	210							
8	7	105	81	205							
9	8	113	63	275							
10	9	106	67	208							
11	10	131	77	159							

Descriptive Statistics

Input
Input Range: $B$1:$B$11
Grouped By: ● Columns ○ Rows
☑ Labels in first row

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

FIGURE 6–10 Descriptive Statistics on SBP


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS H
Reply with Changes... Er
H20
fx

	A	B	C	D
1	SBP			
2				
3	Mean	121.2		
4	Standard Error	3.504917181		
5	Median	122.5		
6	Mode	#N/A		
7	Standard Deviation	11.0835213		
8	Sample Variance	122.8444444		
9	Kurtosis	-0.129625436		
10	Skewness	0.064730195		
11	Range	36		
12	Minimum	105		
13	Maximum	141		
14	Sum	1212		
15	Count	10		
16	Confidence Level(95.0%)	7.928673489		
17				
18				

6.2 CONFIDENCE INTERVALS FOR ONE SAMPLE, DICHOTOMOUS OUTCOME

In Chapter 6 of the textbook, we presented the following formula for the confidence interval for a proportion (of a dichotomous variable) in one sample:

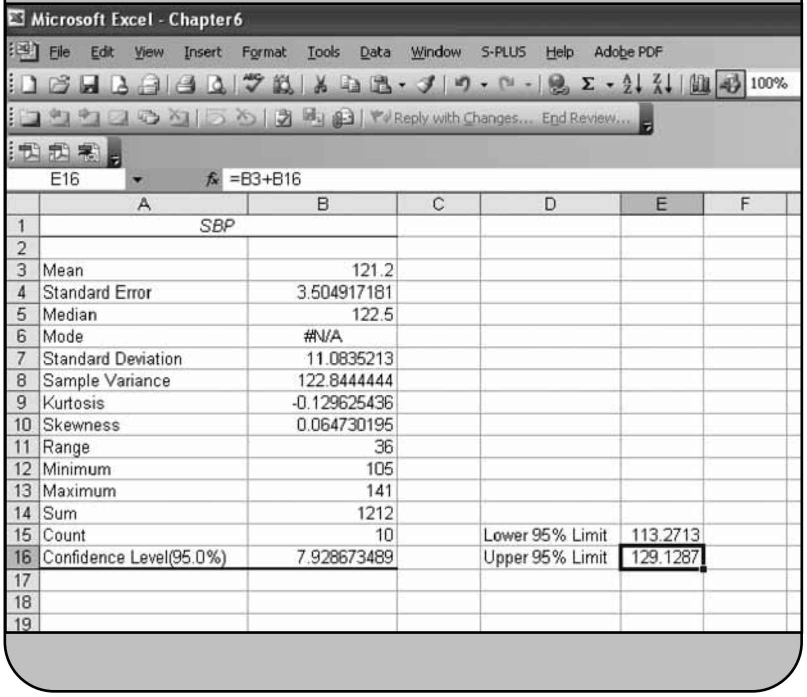
$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{(Find } z \text{ in Table 1B)}$$

When computing the confidence intervals by hand, we computed the sample size and sample proportion and then used Table 1B in the Appendix of the textbook to find the appropriate *z* value to reflect the desired confidence level. We now use Excel to compute the sample proportion and to determine the appropriate *z* value for the confidence interval.

Example 6.4. In Example 6.4 in the textbook, we analyzed data on the prevalence of cardiovascular disease (CVD) measured in men and women at the fifth examination of the Framingham Offspring Study. The data are shown in Table 6–4. The data are entered into an Excel worksheet and shown in Figure 6–12.

In Figure 6–13, we compute the sample proportions with prevalent CVD by dividing the numbers with prevalent CVD by the respective totals, and then generate the *z* values for 95%

FIGURE 6–11 Computing Confidence Limits From Point Estimate and Margin of Error


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...
E16

fx =B3+B16

	A	B	C	D	E	F
1	SBP					
2						
3	Mean	121.2				
4	Standard Error	3.504917181				
5	Median	122.5				
6	Mode	#N/A				
7	Standard Deviation	11.0835213				
8	Sample Variance	122.8444444				
9	Kurtosis	-0.129625436				
10	Skewness	0.064730195				
11	Range	36				
12	Minimum	105				
13	Maximum	141				
14	Sum	1212				
15	Count	10		Lower 95% Limit	113.2713	
16	Confidence Level(95.0%)	7.928673489		Upper 95% Limit	129.1287	
17						
18						
19						

TABLE 6–4 Prevalence of CVD in Men and Women Attending the Fifth Examination of the Framingham Offspring Study

	Free of CVD	Prevalent CVD	Total
Men	1548	244	1792
Women	1872	135	2007
Total	3420	379	3799

confidence using the NORMSINV function. The sample proportions and z values are shown in Figure 6–13.

In Figure 6–14, we compute the upper and lower limits of the 95% confidence interval using $p \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Notice how the formula is implemented. For example, the lower 95% confidence limit for the total sample is in cell G4 and is computed as "=E4−F4*SQRT(E4*(1−E4)/D4)". The upper 95% confidence limit for the total sample is in cell H4 and is computed as "=E4+F4*SQRT(E4*(1−E4)/D4)".

6.3 CONFIDENCE INTERVALS FOR TWO INDEPENDENT SAMPLES, CONTINUOUS OUTCOME

In Chapter 6 of the textbook, we presented the following formulas for confidence intervals for the difference in means of a continuous variable in two independent samples:

$n_1 \geq 30$ and $n_2 \geq 30$:

$$(\overline{X}_1 - \overline{X}_2) \pm zS_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad \text{(Find } z \text{ in Table 1B)}$$

$n_1 < 30$ or $n_2 < 30$:

$$(\overline{X}_1 - \overline{X}_2) \pm tS_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad \text{(Find } t \text{ in Table 2, } df = n_1 + n_2 - 2)$$

where

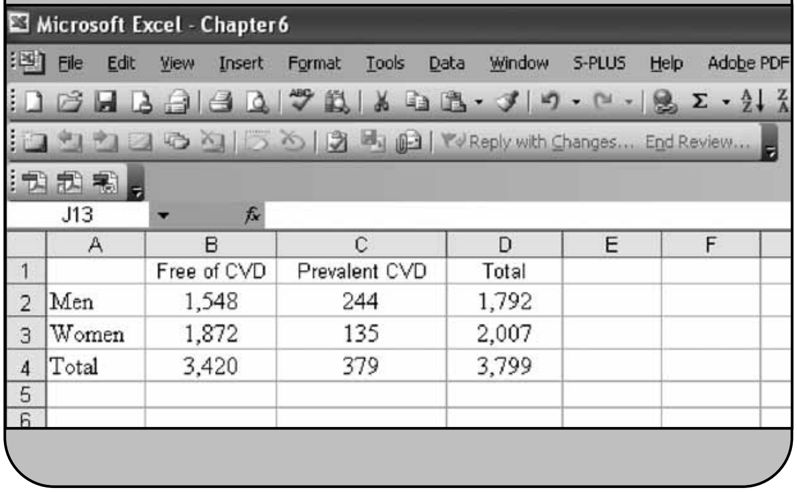
$$S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

When computing the confidence intervals by hand, we compute the sample sizes, means, and standard deviations in each sample. We then compute the pooled estimate of the common standard deviation, S_p, and use Table 1B or Table 2 in the Appendix of the textbook to find the appropriate z or t value to reflect the desired confidence level. We now use Excel to compute summary statistics and to determine the appropriate z or t value for the confidence interval. Once all of the requisite components are determined, we construct the confidence interval.

Example 6.5. In Example 6.5 in the textbook, we analyzed data on $n = 3539$ participants who attended the seventh examination of the offspring in the Framingham Heart Study and compared men and women on the characteristics shown in Table 6–5.

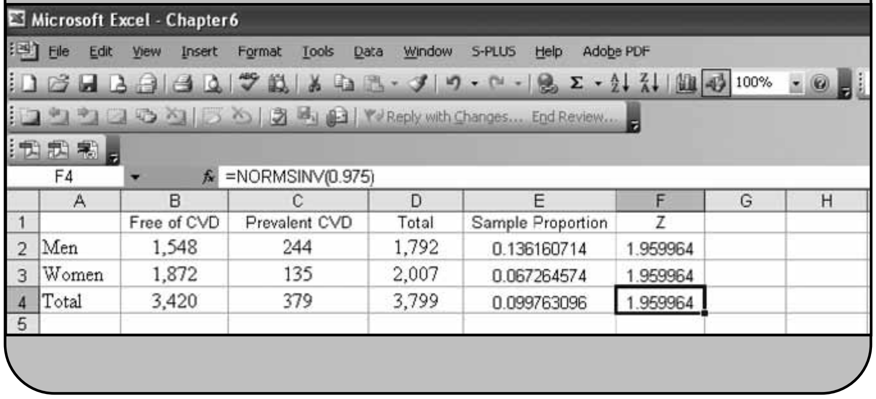
We use Excel to generate 95% confidence intervals for the difference in means between men and women. Because the sample sizes are large, we use the confidence interval formula with z as opposed to t. The data are entered into an Excel worksheet as shown in Figure 6–15.

FIGURE 6–12 Data for Confidence Intervals for Proportions


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...
J13
fx

	A	B	C	D	E	F
1		Free of CVD	Prevalent CVD	Total		
2	Men	1,548	244	1,792		
3	Women	1,872	135	2,007		
4	Total	3,420	379	3,799		
5						
6						

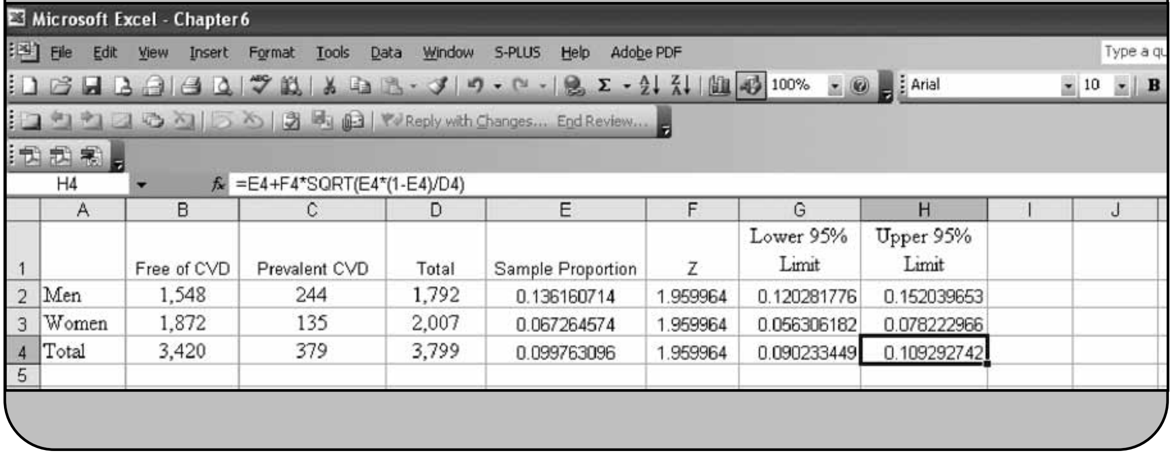
FIGURE 6–13 Sample Proportions and *z* Values for Confidence Intervals


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

F4 =NORMSINV(0.975)

	A	B	C	D	E	F	G	H
1		Free of CVD	Prevalent CVD	Total	Sample Proportion	Z		
2	Men	1,548	244	1,792	0.136160714	1.959964		
3	Women	1,872	135	2,007	0.067264574	1.959964		
4	Total	3,420	379	3,799	0.099763096	1.959964		
5								

FIGURE 6–14 Upper and Lower Limits of 95% Confidence Intervals


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a qu
100%
Arial
10
B
Reply with Changes... End Review...

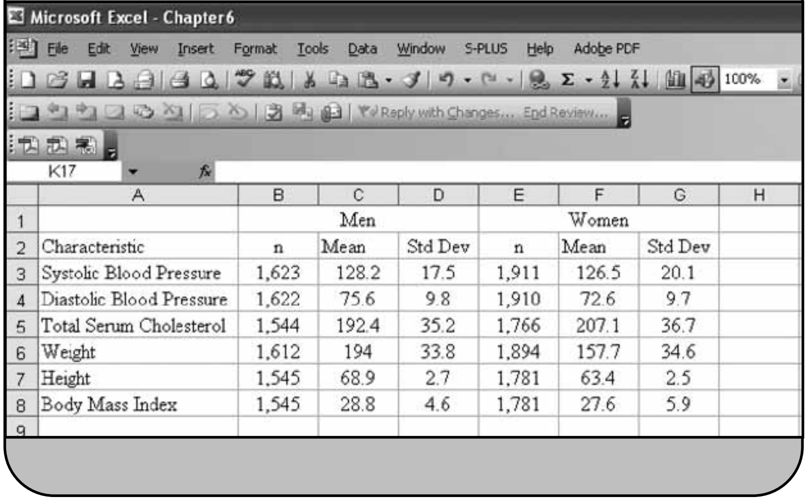
H4 =E4+F4*SQRT(E4*(1-E4)/D4)

	A	B	C	D	E	F	G	H	I	J
1		Free of CVD	Prevalent CVD	Total	Sample Proportion	Z	Lower 95% Limit	Upper 95% Limit		
2	Men	1,548	244	1,792	0.136160714	1.959964	0.120281776	0.152039653		
3	Women	1,872	135	2,007	0.067264574	1.959964	0.056306182	0.078222966		
4	Total	3,420	379	3,799	0.099763096	1.959964	0.090233449	0.109292742		
5										

TABLE 6–5 Summary Statistics on Men and Women Attending the Seventh Examination of the Framingham Heart Study

	Men			Women		
Characteristic	*N*	$\overline{X}$	*s*	*n*	$\overline{X}$	*s*
Systolic blood pressure	1623	128.2	17.5	1911	126.5	20.1
Diastolic blood pressure	1622	75.6	9.8	1910	72.6	9.7
Total serum cholesterol	1544	192.4	35.2	1766	207.1	36.7
Weight (lb)	1612	194.0	33.8	1894	157.7	34.6
Height (in)	1545	68.9	2.7	1781	63.4	2.5
Body mass index (BMI)	1545	28.8	4.6	1781	27.6	5.9

FIGURE 6–15 Data for Confidence Intervals for Differences in Means


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

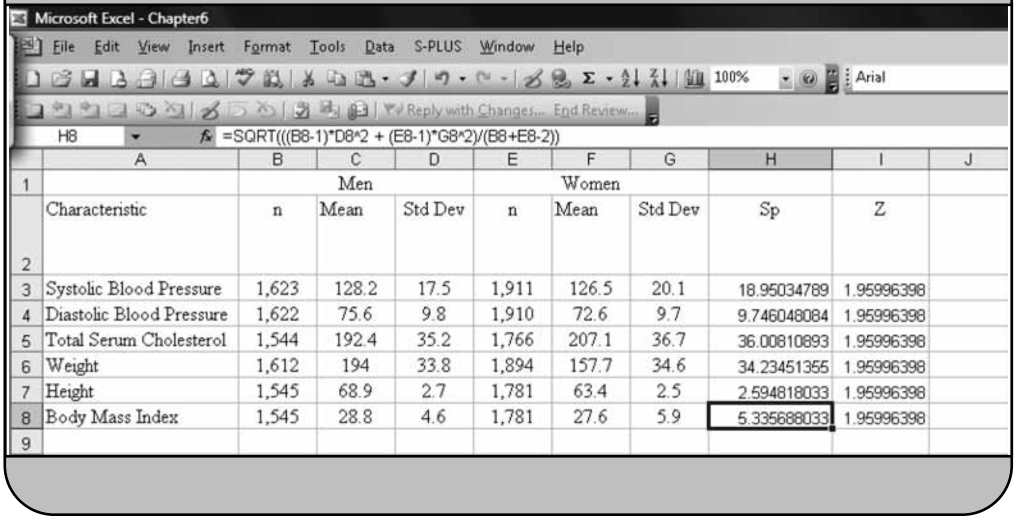
K17

	A	B	C	D	E	F	G	H
1			Men			Women		
2	Characteristic	n	Mean	Std Dev	n	Mean	Std Dev	
3	Systolic Blood Pressure	1,623	128.2	17.5	1,911	126.5	20.1	
4	Diastolic Blood Pressure	1,622	75.6	9.8	1,910	72.6	9.7	
5	Total Serum Cholesterol	1,544	192.4	35.2	1,766	207.1	36.7	
6	Weight	1,612	194	33.8	1,894	157.7	34.6	
7	Height	1,545	68.9	2.7	1,781	63.4	2.5	
8	Body Mass Index	1,545	28.8	4.6	1,781	27.6	5.9	
9								

First we compute the pooled estimates of the common standard deviations $S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$ and place these in column H. For systolic blood pressure, the following is entered into cell H3: "=SQRT(((B3−1)*D3^2 + (E3−1)*G3^2)/(B3+E3−2))". We then compute the *z* scores for 95% confidence intervals and place these in column I. The computation is the same for each characteristic—for example, in cell I3 we enter "=NORMSINV(0.975)". The pooled estimates of the common standard deviations S_p and *z* values are shown in Figure 6–16.

We now compute the point estimates for the difference in means $(\overline{X}_1 - \overline{X}_2)$ and the lower and upper limits of the 95% confidence intervals using $(\overline{X}_1 - \overline{X}_2) \pm z\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. The confidence intervals are shown in Figure 6–17. The 95% confidence limits are shown in column K and column L for each characteristic.

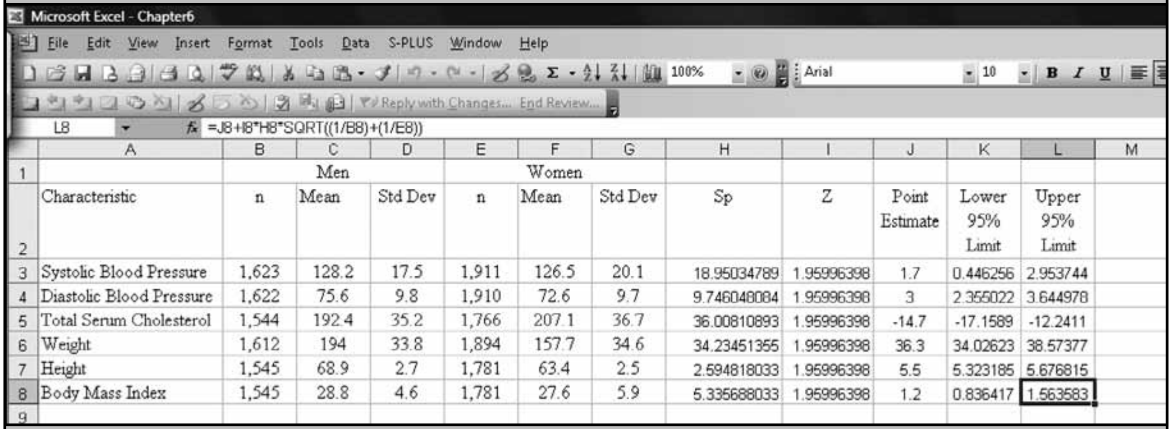
FIGURE 6–16 S_p and *z* Values for Confidence Intervals


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data S-PLUS Window Help
100%
Arial
Reply with Changes... End Review...

H8 =SQRT(((B8-1)*D8^2 + (E8-1)*G8^2)/(B8+E8-2))

	A	B	C	D	E	F	G	H	I	J
1			Men			Women				
2	Characteristic	n	Mean	Std Dev	n	Mean	Std Dev	Sp	Z	
3	Systolic Blood Pressure	1,623	128.2	17.5	1,911	126.5	20.1	18.95034789	1.95996398	
4	Diastolic Blood Pressure	1,622	75.6	9.8	1,910	72.6	9.7	9.746048084	1.95996398	
5	Total Serum Cholesterol	1,544	192.4	35.2	1,766	207.1	36.7	36.00810893	1.95996398	
6	Weight	1,612	194	33.8	1,894	157.7	34.6	34.23451355	1.95996398	
7	Height	1,545	68.9	2.7	1,781	63.4	2.5	2.594818033	1.95996398	
8	Body Mass Index	1,545	28.8	4.6	1,781	27.6	5.9	5.335688033	1.95996398	
9										

FIGURE 6–17 Upper and Lower Limits of 95% Confidence Intervals


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data S-PLUS Window Help
100%
Arial
10
B I U
Reply with Changes... End Review...

L8 =J8+I8*H8*SQRT((1/B8)+(1/E8))

	A	B	C	D	E	F	G	H	I	J	K	L	M
1			Men			Women							
2	Characteristic	n	Mean	Std Dev	n	Mean	Std Dev	Sp	Z	Point Estimate	Lower 95% Limit	Upper 95% Limit	
3	Systolic Blood Pressure	1,623	128.2	17.5	1,911	126.5	20.1	18.95034789	1.95996398	1.7	0.446256	2.953744	
4	Diastolic Blood Pressure	1,622	75.6	9.8	1,910	72.6	9.7	9.746048084	1.95996398	3	2.355022	3.644978	
5	Total Serum Cholesterol	1,544	192.4	35.2	1,766	207.1	36.7	36.00810893	1.95996398	-14.7	-17.1589	-12.2411	
6	Weight	1,612	194	33.8	1,894	157.7	34.6	34.23451355	1.95996398	36.3	34.02623	38.57377	
7	Height	1,545	68.9	2.7	1,781	63.4	2.5	2.594818033	1.95996398	5.5	5.323185	5.676815	
8	Body Mass Index	1,545	28.8	4.6	1,781	27.6	5.9	5.335688033	1.95996398	1.2	0.836417	1.563583	
9													

The same approach is used to compute confidence intervals for the difference in means when the sample sizes are small (i.e., when one or both of the sample sizes are less than 30), except that the TINV function is used to compute the appropriate value from the *t* distribution with degrees of freedom equal to $n_1 + n_2 - 2$.

6.4 CONFIDENCE INTERVALS FOR MATCHED SAMPLES, CONTINUOUS OUTCOME

In Chapter 6 of the textbook, we presented the following formulas for confidence intervals for the mean difference of a continuous variable in two dependent or matched samples.

$$n \geq 30: \quad \overline{X}_d \pm z \frac{s_d}{\sqrt{n}} \text{ (Find } z \text{ in Table 1B)}$$

$$n < 30: \quad \overline{X}_d \pm t \frac{s_d}{\sqrt{n}} \text{ (Find } t \text{ in Table 2, } df = n - 1)$$

where n is the number of participants or pairs and $\overline{X}_d$ and s_d are the mean and standard deviation of the difference scores (where differences are computed on each participant or between members of a matched pair).

We now use Excel to compute summary statistics and to determine the appropriate z or t value for the confidence interval. Once all of the requisite components are determined, we construct the confidence interval.

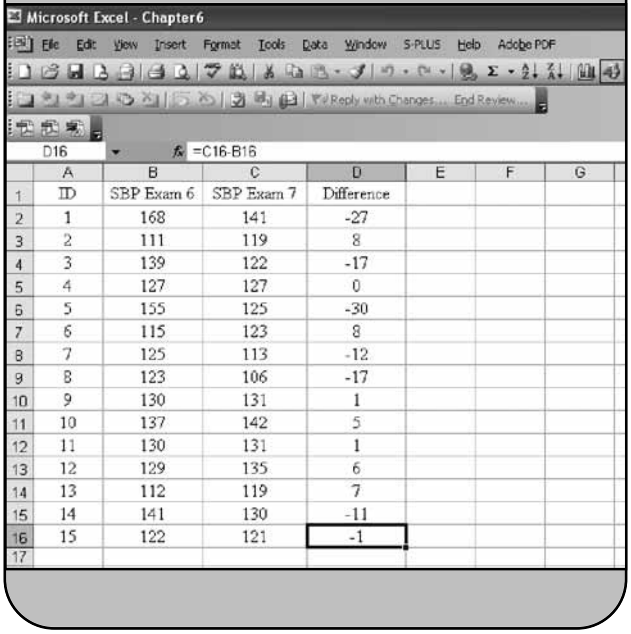
Example 6.6. In Example 6.7 in the textbook, we analyzed systolic blood pressures measured at the sixth and seventh examinations of the offspring in the Framingham Heart Study in a subsample of $n = 15$ randomly selected participants. The data are shown in Table 6–6.

We use Excel to compute difference scores, to generate summary statistics on the difference scores, and to generate a 95% confidence interval for the mean difference in systolic blood pressures over time. The data are entered into an Excel worksheet as shown in Figure 6–18. Difference scores are

TABLE 6–6 Systolic Blood Pressures Measured at Examinations 6 and 7

Subject Identification Number	Examination 6	Examination 7
1	168	141
2	111	119
3	139	122
4	127	127
5	155	125
6	115	123
7	125	113
8	123	106
9	130	131
10	137	142
11	130	131
12	129	135
13	112	119
14	141	130
15	122	121

FIGURE 6–18 Data for Confidence Interval for Mean Difference


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

D16 =C16-B16

	A	B	C	D	E	F	G
1	ID	SBP Exam 6	SBP Exam 7	Difference			
2	1	168	141	-27			
3	2	111	119	8			
4	3	139	122	-17			
5	4	127	127	0			
6	5	155	125	-30			
7	6	115	123	8			
8	7	125	113	-12			
9	8	123	106	-17			
10	9	130	131	1			
11	10	137	142	5			
12	11	130	131	1			
13	12	129	135	6			
14	13	112	119	7			
15	14	141	130	-11			
16	15	122	121	-1			
17							

computed for each participant by subtracting the systolic blood pressure measured at Exam 6 from that measured at Exam 7.

We next use the Descriptive Statistics Analysis Tool to generate descriptive statistics on the differences and we request the confidence interval information. The specifications for the analysis are shown in Figure 6–19. In the following, we request that the results are placed in the current worksheet and we specify the top-left corner of the results table as cell F1. The descriptive statistics are shown in Figure 6–20.

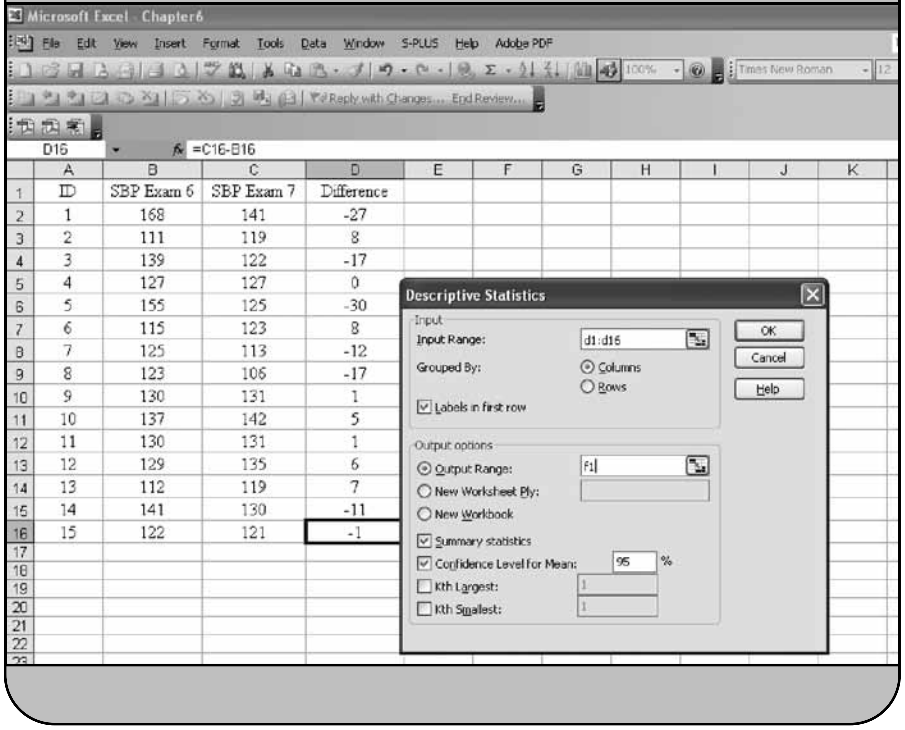
Cell G16 contains the margin of error (i.e., the product of the *t* value for 95% confidence and the standard error). We now take the margin of error and add it to and subtract it from the mean difference in the sample (the point estimate in cell G3) to produce the confidence limits. This is done and shown in Figure 6–21.

6.5 CONFIDENCE INTERVALS FOR TWO INDEPENDENT SAMPLES, DICHOTOMOUS OUTCOME

In Chapter 6 of the textbook, we presented the following formula for the confidence interval for a difference in proportions in two independent samples.

$$\hat{p}_1 - \hat{p}_2 \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \qquad \text{(Find } z \text{ in Table 1B)}$$

FIGURE 6–19 Computing a Confidence Interval on the Difference Scores


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Times New Roman
12
Reply with Changes... End Review...
OK
Cancel
Help

D16 =C16-B16

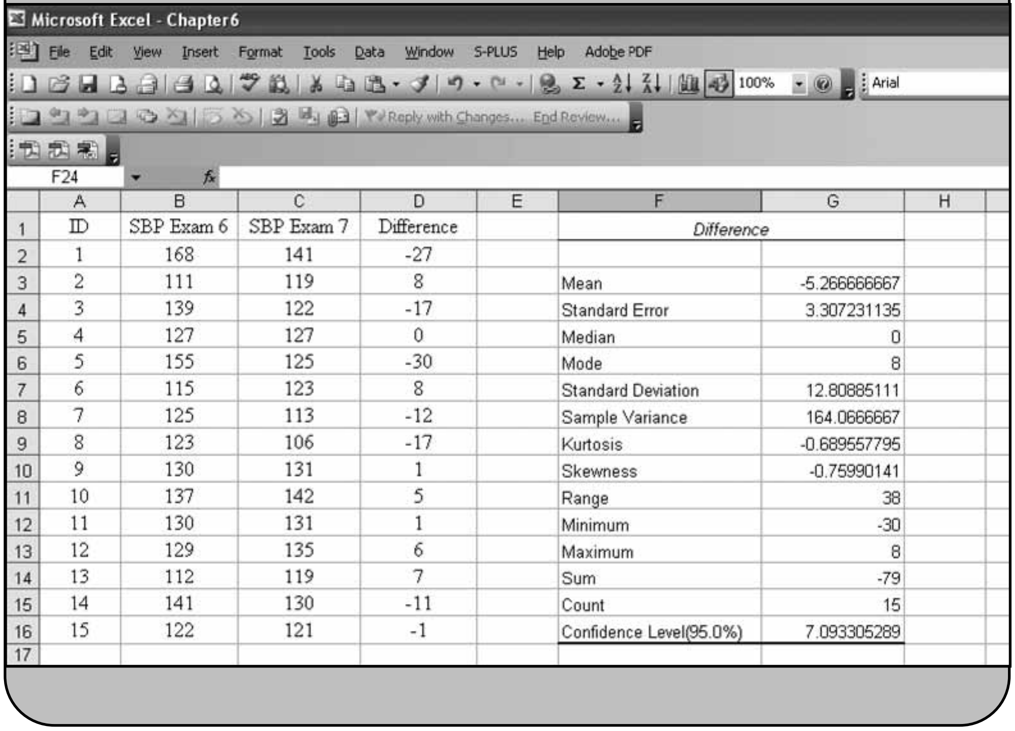
	A	B	C	D	E	F	G	H	I	J	K
1	ID	SBP Exam 6	SBP Exam 7	Difference							
2	1	168	141	-27							
3	2	111	119	8							
4	3	139	122	-17							
5	4	127	127	0							
6	5	155	125	-30							
7	6	115	123	8							
8	7	125	113	-12							
9	8	123	106	-17							
10	9	130	131	1							
11	10	137	142	5							
12	11	130	131	1							
13	12	129	135	6							
14	13	112	119	7							
15	14	141	130	-11							
16	15	122	121	-1							

Descriptive Statistics

Input
Input Range: d1:d16
Grouped By: ● Columns ○ Rows
☑ Labels in first row

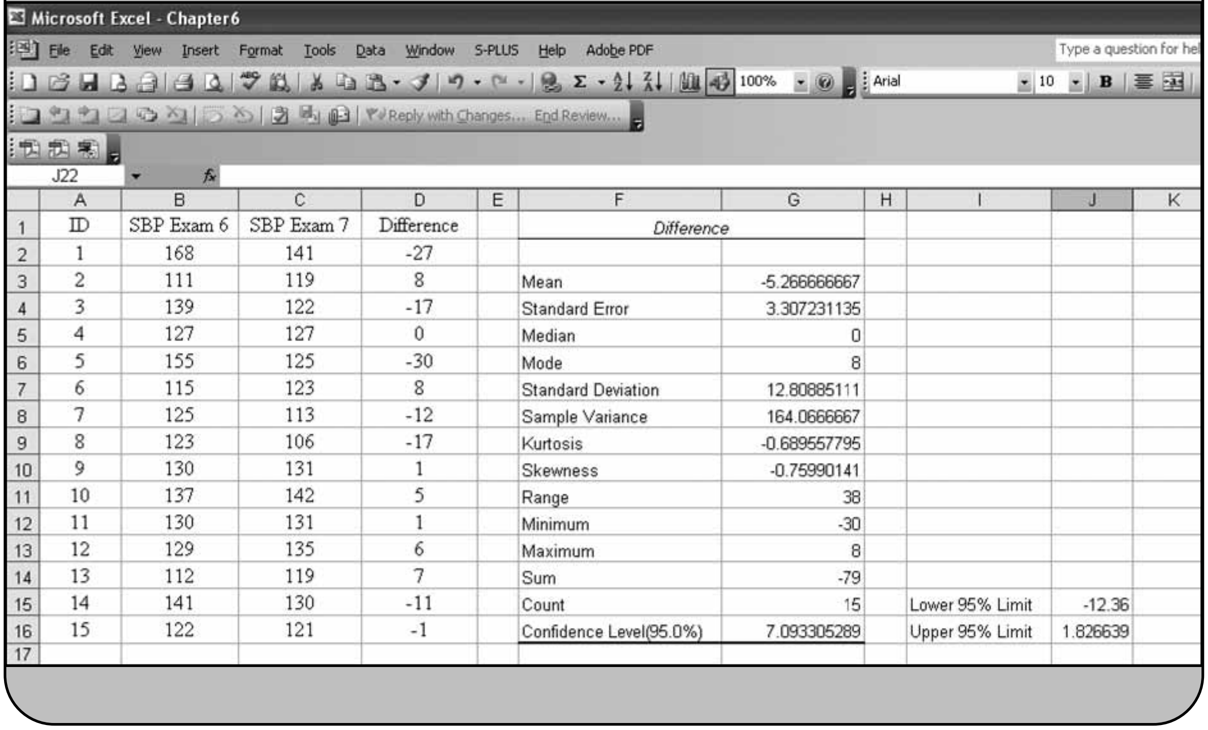
Output options
● Output Range: F1
○ New Worksheet Ply:
○ New Workbook
☑ Summary statistics
☑ Confidence Level for Mean: 95 %
□ Kth Largest: 1
□ Kth Smallest: 1

FIGURE 6–20 Descriptive Statistics on Difference Scores


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
F24
fx

ID	SBP Exam 6	SBP Exam 7	Difference		*Difference*	
1	168	141	-27			
2	111	119	8		Mean	-5.266666667
3	139	122	-17		Standard Error	3.307231135
4	127	127	0		Median	0
5	155	125	-30		Mode	8
6	115	123	8		Standard Deviation	12.80885111
7	125	113	-12		Sample Variance	164.0666667
8	123	106	-17		Kurtosis	-0.689557795
9	130	131	1		Skewness	-0.75990141
10	137	142	5		Range	38
11	130	131	1		Minimum	-30
12	129	135	6		Maximum	8
13	112	119	7		Sum	-79
14	141	130	-11		Count	15
15	122	121	-1		Confidence Level(95.0%)	7.093305289

FIGURE 6–21 Upper and Lower Limits of 95% Confidence Interval


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a question for hel
100%
Arial
10
B
Reply with Changes... End Review...
J22
fx

ID	SBP Exam 6	SBP Exam 7	Difference		*Difference*				
1	168	141	-27						
2	111	119	8		Mean	-5.266666667			
3	139	122	-17		Standard Error	3.307231135			
4	127	127	0		Median	0			
5	155	125	-30		Mode	8			
6	115	123	8		Standard Deviation	12.80885111			
7	125	113	-12		Sample Variance	164.0666667			
8	123	106	-17		Kurtosis	-0.689557795			
9	130	131	1		Skewness	-0.75990141			
10	137	142	5		Range	38			
11	130	131	1		Minimum	-30			
12	129	135	6		Maximum	8			
13	112	119	7		Sum	-79			
14	141	130	-11		Count	15		Lower 95% Limit	-12.36
15	122	121	-1		Confidence Level(95.0%)	7.093305289		Upper 95% Limit	1.826639

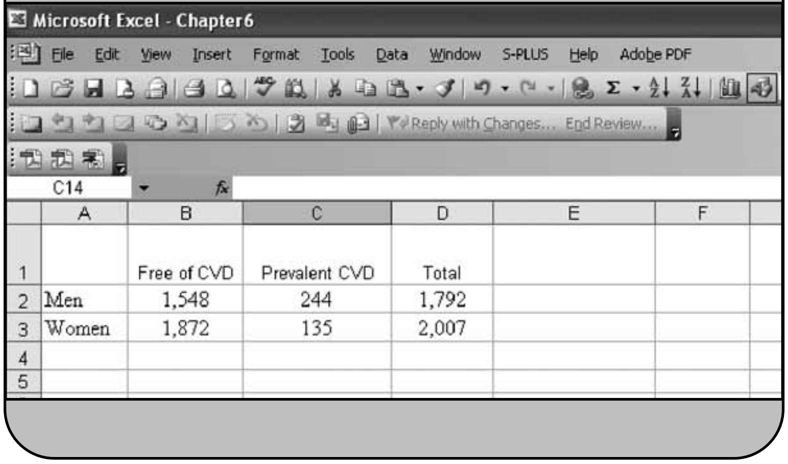
When computing the confidence intervals by hand, we computed the sample sizes and sample proportions, and then used Table 1B in the Appendix of the textbook to find the appropriate *z* value to reflect the desired confidence level. We now use Excel to compute the sample proportions and to determine the appropriate *z* value for the confidence interval.

Example 6.7. In Example 6.4 of the Excel workbook, we analyzed data on the prevalence of cardiovascular disease (CVD) measured in men and women at the fifth examination of the Framingham Offspring Study. The data are shown in Table 6–4. The data are entered into an Excel worksheet and shown in Figure 6–22.

In Figure 6–23, we compute the sample proportions with prevalent CVD by dividing the numbers with prevalent CVD by the respective totals. We then compute the point estimate as the difference in sample proportions and generate the *z* value for 95% confidence using the NORMSINV function.

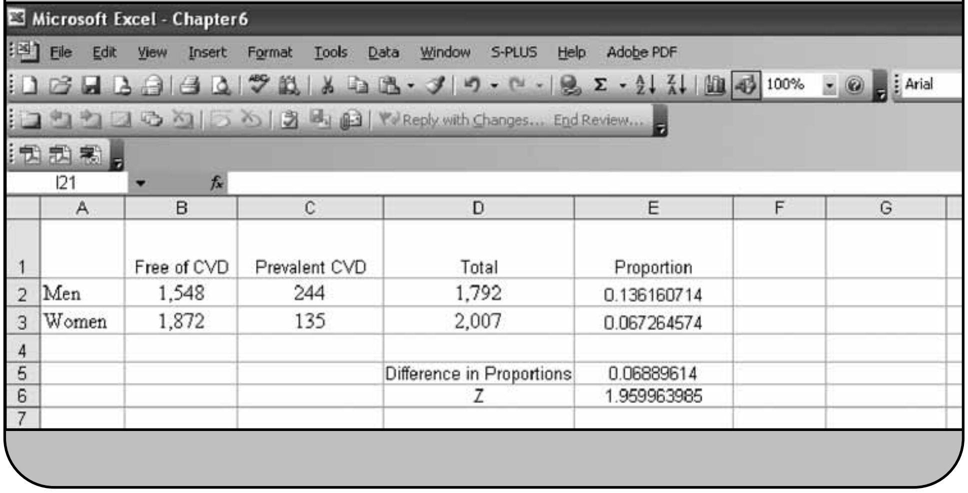
In Figure 6–24, we compute the upper and lower limits of the 95% confidence interval using $\hat{p}_1 - \hat{p}_2 \pm z \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$. Notice how the formula is implemented (see the formula for the upper limit in the formula bar). Excel can be used to generate confidence intervals for relative risks and odds ratios using a similar approach. The exact

FIGURE 6–22 Data for Confidence Interval for Difference in Proportions


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...
C14
fx

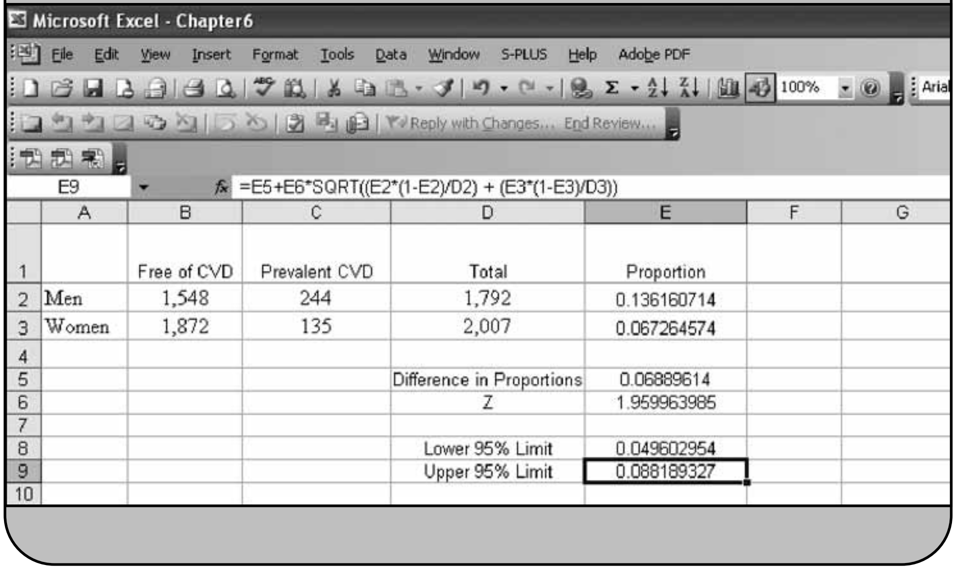
	A	B	C	D	E	F
1		Free of CVD	Prevalent CVD	Total		
2	Men	1,548	244	1,792		
3	Women	1,872	135	2,007		
4						
5						

FIGURE 6–23 Sample Proportions, Difference in Proportions, and *z* Value for Confidence Interval


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
I21
fx

	A	B	C	D	E	F	G
1		Free of CVD	Prevalent CVD	Total	Proportion		
2	Men	1,548	244	1,792	0.136160714		
3	Women	1,872	135	2,007	0.067264574		
4							
5				Difference in Proportions	0.06889614		
6				Z	1.959963985		
7							

FIGURE 6–24 Upper and Lower Limits of 95% Confidence Interval


Microsoft Excel - Chapter6
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
E9

=E5+E6*SQRT((E2*(1-E2)/D2) + (E3*(1-E3)/D3))

	A	B	C	D	E	F	G
1		Free of CVD	Prevalent CVD	Total	Proportion		
2	Men	1,548	244	1,792	0.136160714		
3	Women	1,872	135	2,007	0.067264574		
4							
5				Difference in Proportions	0.06889614		
6				Z	1.959963985		
7							
8				Lower 95% Limit	0.049602954		
9				Upper 95% Limit	0.088189327		
10							

formulas for these confidence intervals can be found in Chapter 6 of the textbook.

6.6 PRACTICE PROBLEMS

1. A study is run to estimate the mean total cholesterol level in children 2 to 6 years of age. A sample of 9 participants is selected and their total cholesterol levels are measured as follows:

   185 225 240 196 175 180 194 147 223

   Generate a 95% confidence interval for the true mean total cholesterol levels in children.

2. A clinical trial is planned to compare an experimental medication designed to lower blood pressure to a placebo. Before starting the trial, a pilot study is conducted involving 10 participants. The objective of the study is to assess how systolic blood pressure changes untreated over time. Systolic blood pressures are measured at baseline and again 4 weeks later. Compute a 95% confidence interval for the mean difference in blood pressures over 4 weeks.

   Baseline: 120 145 130 160 152 143 126 121 115 135

   4 Weeks: 122 142 135 158 155 140 130 120 124 130

3. After the pilot study described in Problem 2, the main trial is conducted and involves a total of 200 patients. Patients are enrolled and randomized to receive either the experimental medication or the placebo. The data shown in Table 6–7 are data collected at the end of the study after 6 weeks on the assigned treatment. Generate a 95% confidence interval for the difference in proportions of patients with hypertension between groups.

4. The following data were collected as part of a study of coffee consumption among male and female undergraduate students. The following reflect cups per day consumed:

Male:	3	4	6	8	2	1	0	2
Female:	5	3	1	2	0	4	3	1

TABLE 6–7 Data for Problem 3

	Experimental **(n = 100)**	**Placebo** **(n = 100)**
% Hypertensive	14%	22%

Generate a 95% confidence interval for the difference in mean numbers of cups of coffee consumed between men and women.

5. A clinical trial is conducted comparing a new pain reliever for arthritis to a placebo. Participants are randomly assigned to receive the new medication or a placebo. The outcome is pain relief within 30 minutes. The data are shown in Table 6–8.
   a. Generate a 95% confidence interval for the proportion of patients on the new medication who report pain relief.
   b. Generate a 95% confidence interval for the difference in proportions of patients who report pain relief.

TABLE 6–8 Data for Problem 5

	Pain Relief	**No Pain Relief**
New medication	44	76
Placebo	21	99

CHAPTER 7

Hypothesis Testing Procedures

In Chapter 7 of the textbook, we presented the approach for hypothesis testing for means (μ) and proportions (p) in one sample, for differences in means ($\mu_1 - \mu_2$) and differences in proportions ($p_1 - p_2$) in two independent samples, for the mean difference in two dependent samples (μ_d), and for differences in means and proportions in more than two independent samples. For each test we used the same general five-step approach, which is outlined below:

- ***Step 1:*** Set up hypotheses (H_0 and H_1) and select a level of significance, α.
- ***Step 2:*** Choose the appropriate test statistic (e.g., z, t, F, χ^2).
- ***Step 3:*** Determine critical values and set up the decision rule (which depends on α, the test statistic, and whether the test is upper-, lower-, or two-tailed).
- ***Step 4:*** Compute the test statistic based on observed sample data.
- ***Step 5:*** Draw a conclusion by comparing the test statistic to the critical value.

The test statistic (Step 2) varies depending on the specific test. When we conducted tests of hypothesis by hand in Chapter 7 of the textbook, we ultimately drew a conclusion by comparing the test statistic to the critical value, which was derived from an appropriate probability distribution. There is an alternative means of drawing a conclusion, and it involves comparing the p-value of a test (defined as the exact significance level) to the selected level of significance, α. The p-value is the probability of observing a test statistic as or more extreme than that observed, and it can be one-sided or two-sided. For example, suppose we conduct an upper-tailed test for the population mean (i.e., H_1: $\mu > \mu_0$) and observe a test statistic $z = 2.04$. The p-value is $P(z \geq 2.04)$. If we conduct a two-sided test for the population mean (i.e., H_1: $\mu \neq \mu_0$) and observe a test statistic $z = 2.04$, the p-value is $P(z \geq 2.04) + P(z \leq -2.04) = 2 \times P(z \geq 2.04)$. We use Microsoft Office® Excel® to compute the test statistics for each test and to compute p-values for each test to draw conclusions based on the following:

$$\text{Reject } H_0 \text{ if } p \leq \alpha.$$

p-values for tests involving z statistics are computed with the NORMSDIST function, and p-values for tests involving t statistics are computed with the TDIST function.

To compute p-values for tests involving a z statistic, we use the NORMSDIST function: "=NORMSDIST(z)". To use the NORMSDIST function, we specify the value of the test statistic, z. The function returns the area under the standard normal curve below z. To use the NORMSDIST function to compute p-values, we make the following modifications:

One-sided z test = "1−NORMSDIST(ABS(z))"

Two-sided z test = "2*(1−NORMSDIST(ABS(z)))"

The ABS function takes the absolute value of the test statistic. By using the ABS function, we can use the preceding one-sided formula for both upper- and lower-tailed tests.

To compute p-values for tests involving t statistics, we use the TDIST function: "=TDIST(t, df, *test type*)". To use the TDIST function, we specify the test statistic t, the degrees of freedom (e.g., for a one-sample test of means, $df = n - 1$), and then the *test type*. The *test type* indicates whether the test is one- or two-tailed (i.e., *test type* = 1 for upper- or lower-tailed tests and *test type* = 2 for two-tailed tests). The function returns the area in the t distribution in one or two tails (depending on the test type). The TDIST function is used as follows to compute p-values:

One-sided t test = "TDIST(ABS(t), df, 1)"

Two-sided t test = "TDIST(ABS(t), df, 2)"

Again, the ABS function is used to take the absolute value of the test statistic.

7.1 TESTS WITH ONE SAMPLE, CONTINUOUS OUTCOME

For a one-sample test of a hypothesis with a continuous outcome, the hypotheses are as follows:

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0,\ H_1: \mu < \mu_0,\ \text{or } H_1: \mu \neq \mu_0$$

where μ is the mean of the population of interest and μ_0 is a known mean (e.g., an historical control).

In Chapter 7 of the textbook, we presented the following formulas for test statistics:

$$n \geq 30: \quad z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \quad \text{(Find critical value in Table 1C)}$$

$$n < 30: \quad t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \quad \text{(Find critical value in Table 2, } df = n - 1\text{)}$$

When performing the test of hypothesis by hand, we compute the sample size, the mean and standard deviation, and then the test statistic. We use Table 1C or Table 2 in the Appendix of the textbook to find the appropriate critical values of z or t and compare the test statistic to the critical value to draw a conclusion. Excel does not have a specific analysis tool for a one-sample test of means. However, Excel can be used to compute the test statistic and the p-value to draw a conclusion.

Example 7.1. In Example 7.1 in the textbook, we analyzed data on expenditures on health care and prescription drugs. We specifically analyzed whether there was significant evidence of a reduction in expenditures from the reported value of \$3302 per year. To test the hypothesis, a sample of 100 Americans were selected and their expenditures on health care and prescription drugs in 2005 were summarized as follows: $\bar{X}$ = \$3190 and s = \$890. We now run the test using Excel.

The hypotheses are:

$$H_0: \mu = 3302$$

$$H_1: \mu < 3302$$

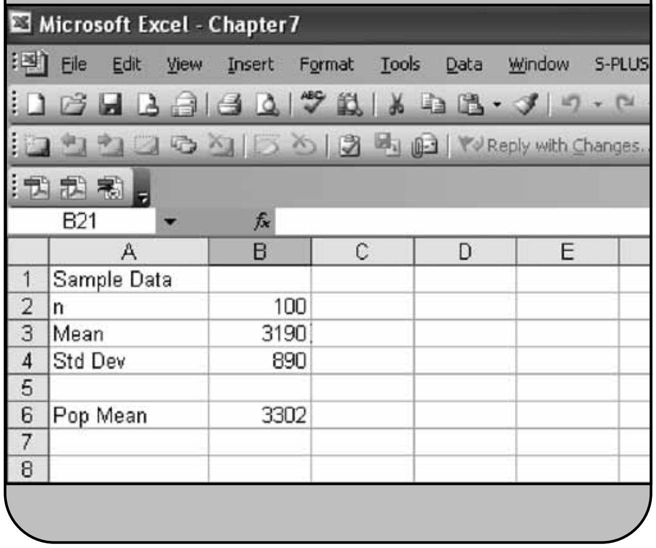
$$\alpha = 0.05$$

Because the sample size is large ($n > 30$), the appropriate test statistic is $z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$.

We use Excel to compute the test statistic and the p-value. The data are entered into an Excel worksheet, as shown in Figure 7–1. We now compute the test statistic, z, and place it in cell B8. The formula is "=(B3−B6)/(B4/SQRT(B2))". The p-value is computed with the NORMSDIST function. Because this is a lower-tailed test, we use "=1−NORMSDIST(ABS(z))". The test statistic and p-value are shown in Figure 7–2.

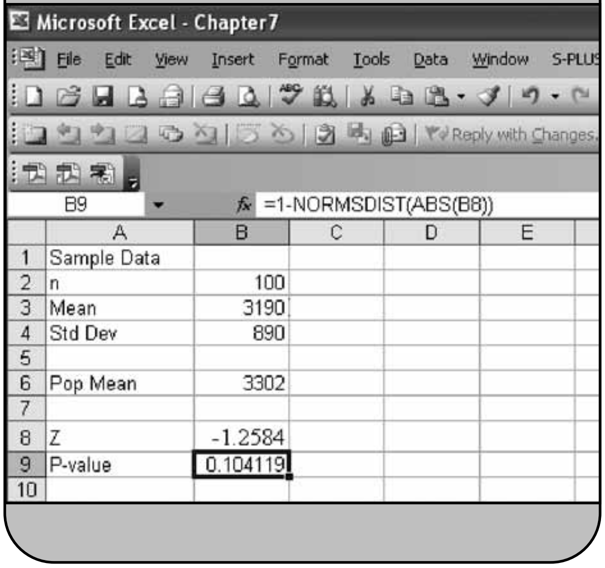
We do not reject H_0 because $p = 0.1041 > \alpha = 0.05$. We do not have significant evidence to show that there is a reduction in expenditures from the reported value of \$3302 per year. (Recall that when this test was done by hand, we compared the test statistic $z = -1.26$ to the critical value from the standard normal distribution and did not reject H_0 because $-1.26 > -1.645$).

FIGURE 7–1 Data for One-Sample Test of Means


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS
Reply with Changes...
B21
fx

	A	B	C	D	E
1	Sample Data				
2	n	100			
3	Mean	3190			
4	Std Dev	890			
5					
6	Pop Mean	3302			
7					
8					

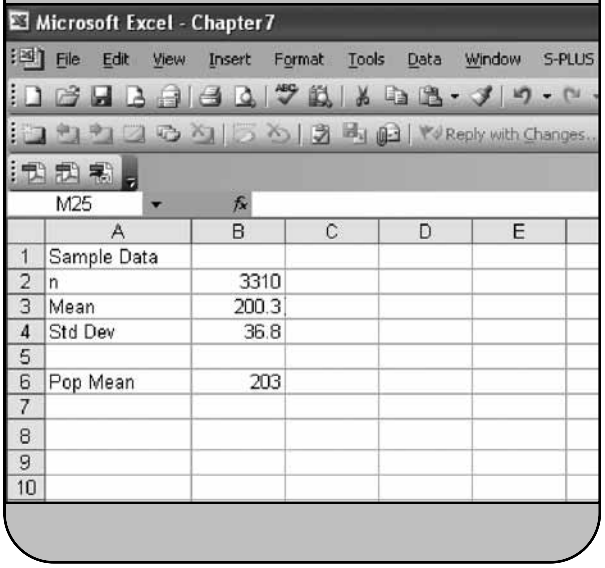
FIGURE 7–2 Test Statistic and *p*-Value


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS
Reply with Changes...

B9 fx =1-NORMSDIST(ABS(B8))

	A	B	C	D	E
1	Sample Data				
2	n	100			
3	Mean	3190			
4	Std Dev	890			
5					
6	Pop Mean	3302			
7					
8	Z	-1.2584			
9	P-value	0.104119			
10					

FIGURE 7–3 Data for One-Sample Test of Means


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS
Reply with Changes...

M25 fx

	A	B	C	D	E
1	Sample Data				
2	n	3310			
3	Mean	200.3			
4	Std Dev	36.8			
5					
6	Pop Mean	203			
7					
8					
9					
10					

Example 7.2. In Example 7.2 in the textbook, we tested whether the mean total cholesterol level in the Framingham Offspring Study was different from the national mean value of 203. The following statistics on total cholesterol levels of participants in the Framingham Offspring Study were available: $n = 3310$, $\overline{X} = 200.3$, and $s = 36.8$. Here we use Excel to test if there is statistical evidence of a difference in mean cholesterol level in the Framingham offspring as compared to the national mean of 203.

The hypotheses are as follows:

$$H_0: \mu = 203$$

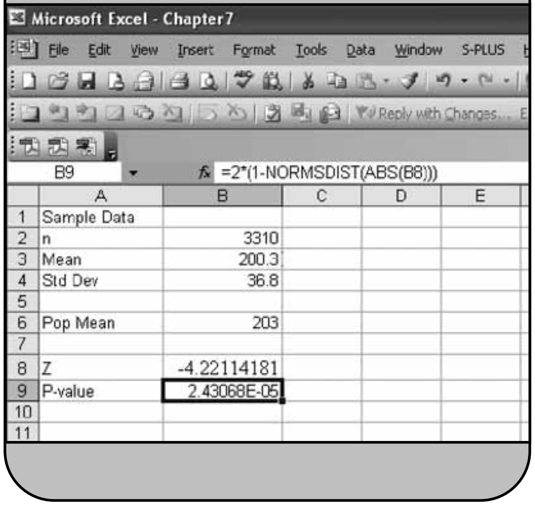
$$H_1: \mu \neq 203$$

$$\alpha = 0.05.$$

Because the sample size is large ($n > 30$), the appropriate test statistic is $z = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$.

We use Excel to compute the test statistic and the *p*-value. The data are entered into an Excel worksheet, as shown in Figure 7–3. We now compute the test statistic, *z*, and place it in cell B8. The formula is "=(B3−B6)/(B4/SQRT(B2))". The two-sided *p*-value is computed with the NORMSDIST function using "=2*(1−NORMSDIST(ABS(*z*)))". The test statistic and *p*-value are shown in Figure 7–4.

FIGURE 7–4 Test Statistic and *p*-Value


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS
Reply with Changes...

B9 fx =2*(1-NORMSDIST(ABS(B8)))

	A	B	C	D	E
1	Sample Data				
2	n	3310			
3	Mean	200.3			
4	Std Dev	36.8			
5					
6	Pop Mean	203			
7					
8	Z	-4.22114181			
9	P-value	2.43068E-05			
10					
11					

We reject H_0 because $p = 0.00002 < \alpha = 0.05$. (Note that the *p*-value is given as 2.43E−05, which is equivalent to $2.43 \times 10^{-5} = 0.0000243$.) We have significant evidence to show that the mean total cholesterol level in the Framingham offspring is different from the national value of 203. (Recall that when this test was done by hand, we compared the test statistic $z = -4.22$ to the critical value from the standard normal distribution and rejected H_0 because $-4.22 < -1.96$.)

7.2 TESTS WITH ONE SAMPLE, DICHOTOMOUS OUTCOME

For a one-sample test of hypothesis with a dichotomous outcome, the hypotheses are as follows:

$$H_0: p = p_0$$

$$H_1: p > p_0, \quad H_1: p < p_0, \quad \text{or } H_1: p \neq p_0$$

where p is the proportion of successes in the population of interest and p_0 is a known proportion (e.g., an historical control).

In Chapter 7 of the textbook, we presented the following test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \text{ (Find critical value in Table 1C)}$$

When performing this test of hypothesis by hand, we compute the sample size and sample proportion, and then the test statistic. We use Table 1C in the Appendix of the textbook to find the appropriate critical value of z and compare the test statistic to the critical value to draw a conclusion. We now use Excel to conduct the test of hypothesis.

Example 7.3. In Example 7.4 in the textbook, we tested whether the prevalence of smoking in the Framingham Offspring Study was lower than the prevalence of smoking among American adults, reported as 21.1%. In the Framingham Offspring Study, 482 of 3536 (13.6%) of the respondents were currently smoking at the time of the exam.

The hypotheses are as follows:

$$H_0: p = 0.211$$

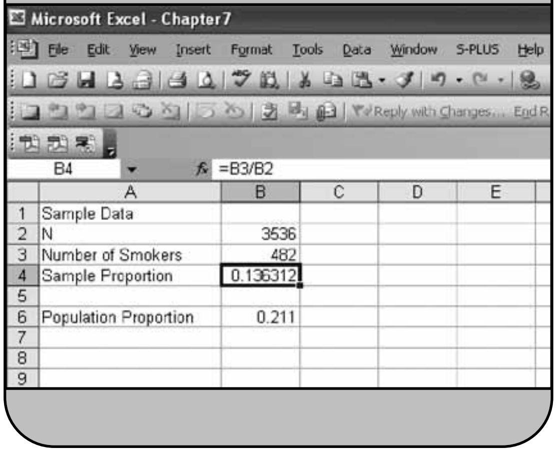
$$H_1: p < 0.211$$

$$\alpha = 0.05.$$

The appropriate test statistic is $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$.

We use Excel to compute the test statistic and the p-value. The data are entered into an Excel worksheet, as shown in Figure 7–5. The sample proportion is shown in cell B4 and is computed by dividing the number of smokers in the sample by the sample size (i.e., B3 / B2).

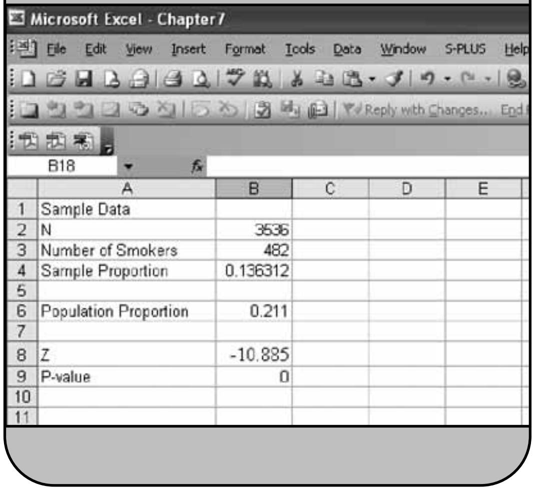
FIGURE 7–5 Data for One-Sample Test of Proportions


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help
Reply with Changes... End R

B4 =B3/B2

	A	B	C	D	E
1	Sample Data				
2	N	3536			
3	Number of Smokers	482			
4	Sample Proportion	0.136312			
5					
6	Population Proportion	0.211			
7					
8					
9					

FIGURE 7–6 Test Statistic and p-Value


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help
Reply with Changes... End

B18

	A	B	C	D	E
1	Sample Data				
2	N	3536			
3	Number of Smokers	482			
4	Sample Proportion	0.136312			
5					
6	Population Proportion	0.211			
7					
8	Z	-10.885			
9	P-value	0			
10					
11					

We now compute the test statistic, z, and place it in cell B8. The formula is "=(B4−B6)/(SQRT(B6*(1−B6)/B2))". The one-sided p-value is computed with the NORMSDIST function using "=1−NORMSDIST(ABS(z))". The test statistic and p-value are shown in Figure 7–6.

In this test, the test statistic is $z = -10.89$ and we reject H_0 because $p = 0 < \alpha = 0.05$. We have statistically significant evidence to show that the prevalence of smoking in the Framingham offspring is lower than the prevalence of smoking among American adults reported at 21.1%.

7.3 TESTS WITH ONE SAMPLE, CATEGORICAL AND ORDINAL OUTCOMES: THE CHI-SQUARE GOODNESS-OF-FIT TEST

For a χ^2 goodness-of-fit test, the hypotheses are as follows:

$$H_0: p_1 = p_{10},\ p_2 = p_{20},\ \ldots,\ p_k = p_{k0}$$

$$H_1: H_0 \text{ is false}$$

where the p_i are in the population proportions of successes in each response category in the population and the p_{i0} are the known proportions in each response category.

In Chapter 7 of the textbook, we presented the following formula for the test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad \text{(Find critical value in Table 3, } df = k - 1)$$

where O = observed frequency and E = expected frequency in each of the response categories and k = the number of response options.

When performing the goodness-of-fit test by hand, we compute the expected frequencies for each category and then compute the test statistic. We then use Table 3 in the Appendix of the textbook to find the appropriate critical value from the χ^2 distribution and compare the test statistic to the critical value to draw a conclusion.

Excel does not have a specific analysis tool to perform the χ^2 goodness-of-fit test. However, it does have a CHIDIST function, which can be used to produce p-values. The CHIDIST function is used as "=CHIDIST(χ^2, *df*)". To use the CHIDIST function, we specify the test statistic, χ^2, and the degrees of freedom, *df*. For the χ^2 goodness-of-fit test, $df = k - 1$, where k represents the number of response categories. The CHIDIST function returns the area in the right tail of the distribution, which is the p-value for the χ^2 goodness-of-fit test. We now use Excel to conduct a goodness-of-fit test.

Example 7.4. In Example 7.6 of the textbook, we analyzed a university's survey of its graduates, in which demographic and health information were collected for future planning purposes. In response to a question on regular exercise, 60% of all graduates reported getting no regular exercise, 25% reported exercising sporadically, and 15% reported exercising regularly as undergraduates. The next year, the university launched a health-promotion campaign on campus in an attempt to increase healthy behaviors among undergraduates and conducted another survey that was completed by 470 graduates. The data shown in Table 7–1 were collected. Based on the data, is there evidence of a shift in the distribution of responses to the exercise question following the implementation of the health-promotion campaign on campus?

TABLE 7–1 Data from University Survey

	No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Number of students	255	125	90	470

The hypotheses are as follows:

$$H_0: p_1 = 0.60,\ p_2 = 0.25,\ p_3 = 0.15, \text{ or equivalently}$$

$$H_0: \text{Distribution of responses is } 0.60, 0.25, 0.15$$

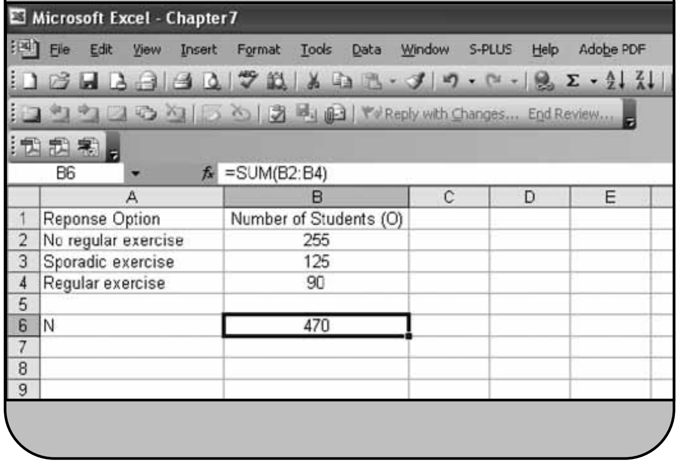
$$H_1: H_0 \text{ is false}$$

$$\alpha = 0.05$$

The appropriate test statistic is $\chi^2 = \sum \frac{(O-E)^2}{E}$.

Recall that the expected frequencies (E) are computed based on the assumption that H_0 is true. The data for the test are entered into an Excel worksheet, as shown in Figure 7–7. The sample data (i.e., the numbers of students in each response category) are the observed frequencies. The total sample size is computed using the SUM function and is shown in cell B6.

FIGURE 7–7 Data for Chi-Square Goodness-of-Fit Test



B6 =SUM(B2:B4)

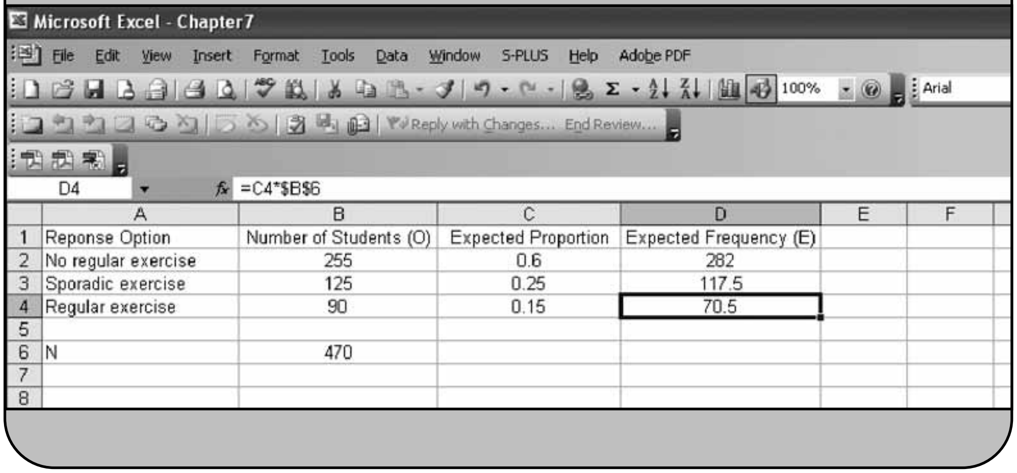
	A	B	C	D	E
1	Reponse Option	Number of Students (O)			
2	No regular exercise	255			
3	Sporadic exercise	125			
4	Regular exercise	90			
5					
6	N	470			
7					
8					
9					

We compute the expected frequencies by multiplying the hypothesized or expected proportions in each response category (from H_0) by the total sample size. The expected proportions are first entered into the Excel worksheet in column C. We then multiply the expected proportions in column C by the sample size to produce the expected frequencies. For example, the expected frequency in cell D2 is computed using "=C2*$B$6". (Notice that we use the absolute cell address for the total sample size so that the same value is used to compute each expected frequency.) The expected proportions and expected frequencies are shown in Figure 7–8.

We now compute $(O - E)^2$ / E in each response category and sum to produce the χ^2 statistic. The χ^2 test statistic is shown in cell E6. The p-value is computed with the CHIDIST function using "=CHIDIST(E6,2)", where "2" reflects the degrees of freedom ($df = k - 1 = 3 - 1 = 2$). The test statistic and p-value are shown in Figure 7–9.

In this test, the test statistic is $\chi^2 = 8.46$ and we reject H_0 because $p = 0.0146 < \alpha = 0.05$. We have statistically significant

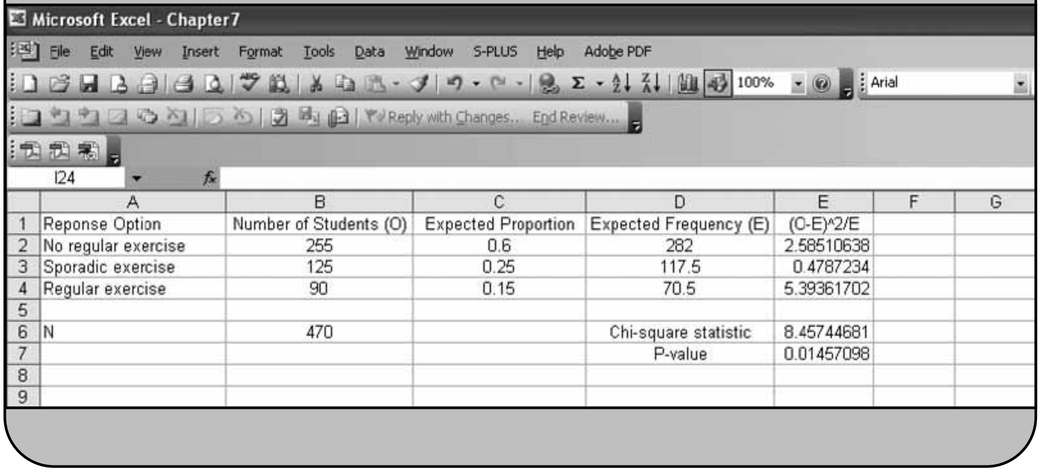
FIGURE 7–8 Expected Frequencies


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...

D4 =C4*$B$6

	A	B	C	D	E	F
1	Reponse Option	Number of Students (O)	Expected Proportion	Expected Frequency (E)		
2	No regular exercise	255	0.6	282		
3	Sporadic exercise	125	0.25	117.5		
4	Regular exercise	90	0.15	70.5		
5						
6	N	470				
7						
8						

FIGURE 7–9 Test Statistic and *p*-Value


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...

I24

	A	B	C	D	E	F	G
1	Reponse Option	Number of Students (O)	Expected Proportion	Expected Frequency (E)	(O-E)^2/E		
2	No regular exercise	255	0.6	282	2.58510638		
3	Sporadic exercise	125	0.25	117.5	0.4787234		
4	Regular exercise	90	0.15	70.5	5.39361702		
5							
6	N	470		Chi-square statistic	8.45744681		
7				P-value	0.01457098		
8							
9							

evidence to show that the distribution of responses is not 0.60, 0.25, 0.15.

7.4 TESTS WITH TWO INDEPENDENT SAMPLES, CONTINUOUS OUTCOME

For a two-independent-samples test of hypothesis with a continuous outcome, the hypotheses are as follows:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 > \mu_2,\ H_1: \mu_1 < \mu_2,\ \text{or } H_1: \mu_1 \neq \mu_2$$

where μ_1 and μ_2 are the means of the two independent populations of interest.

In Chapter 7 of the textbook, we presented the following formulas for test statistics:

$$n_1 \geq 30 \text{ and } n_2 \geq 30: \quad z = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{1/n_1 + 1/n_2}}$$

(Find critical value of z in Table 1C)

$$n_1 < 30 \text{ or } n_2 < 30: \quad t = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{1/n_1 + 1/n_2}}$$

(Find critical value of t in Table 2, $df = n_1 + n_2 - 2$)

$$\text{where } S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

When performing tests of hypothesis by hand, we compute the sample sizes, means, and standard deviations in each sample. We then compute the pooled estimate of the common standard deviation, S_p, and the test statistic. We then use Table 1C or Table 2 in the Appendix of the textbook to find the appropriate critical values of z or t and compare the test statistic to the critical value to draw a conclusion. We now use Excel to conduct the test of hypothesis.

Example 7.5. A clinical trial is run to compare an experimental drug to a placebo for its effectiveness in lowering systolic blood pressure. A total of 18 participants are enrolled in the study and randomly assigned to receive either the experimental drug or placebo. After 6 weeks on the assigned treatment, each patient's systolic blood pressure is measured and the data are shown here:

Experimental drug	125	130	135	121	140	137	129	145	115
Placebo	145	140	132	129	145	150	160	140	120

Is there statistical evidence of a difference in mean systolic blood pressures between treatments? We now run the test using Excel.

The hypotheses are:

$$H_0: \mu_1 = \mu_2$$

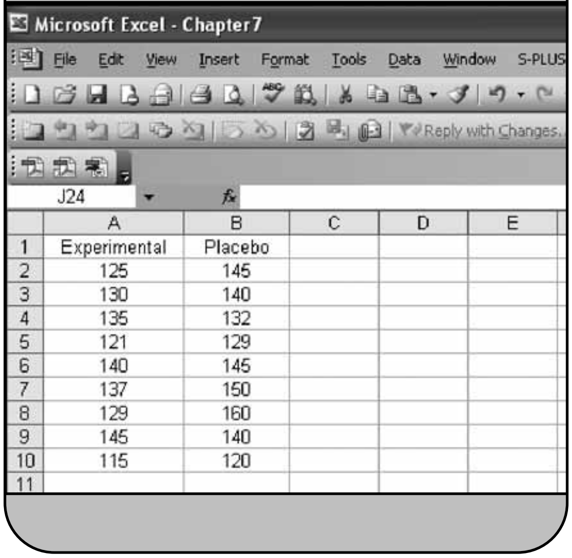
$$H_1: \mu_1 \neq \mu_2$$

$$\alpha = 0.05$$

Because the sample sizes are small (both $n_1 < 30$ and $n_2 < 30$), the appropriate test statistic is $t = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{1/n_1 + 1/n_2}}$. Excel has an analysis tool to perform a two-independent-samples test of means in its Data Analysis ToolPak. We first enter the data into an Excel worksheet, as shown in Figure 7–10.

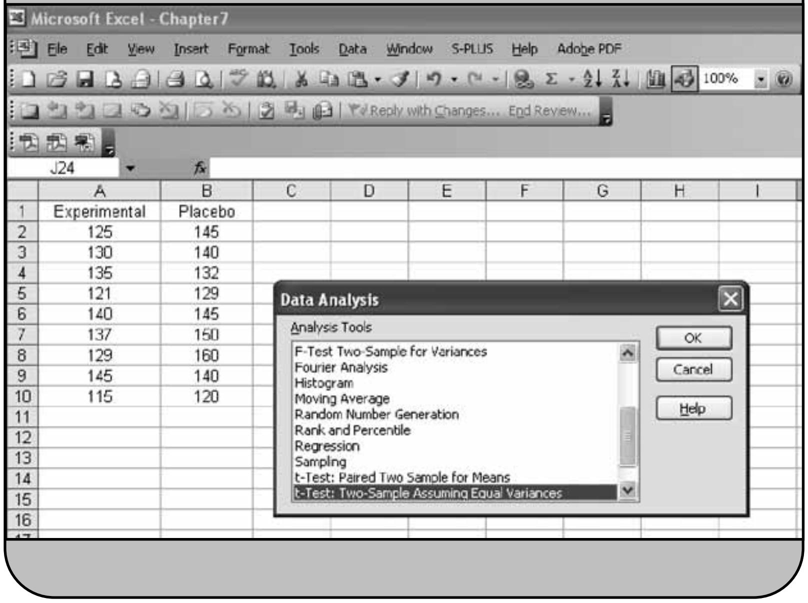
Under the **Tools/Data Analysis** option, we choose the ***t* Test: Two-Sample Assuming Equal Variances** analysis tool, as shown in Figure 7–11. Excel offers other options to perform a two-independent-samples test for the equality of means that do not assume that the variances are equal and thus would not involve S_p. In Chapter 7 of the textbook, we presented guidelines for using the formulas that assume equal variances. Once we click **OK**, Excel presents the dialog box shown in Figure 7–12.

FIGURE 7–10 Data for Two-Independent-Samples Test for Difference in Means


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS
Reply with Changes...
J24
fx

	A	B	C	D	E
1	Experimental	Placebo			
2	125	145			
3	130	140			
4	135	132			
5	121	129			
6	140	145			
7	137	150			
8	129	160			
9	145	140			
10	115	120			
11					

FIGURE 7–11 Two-Independent-Samples Test Using Data Analysis Tool


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...
J24
OK
Cancel
Help

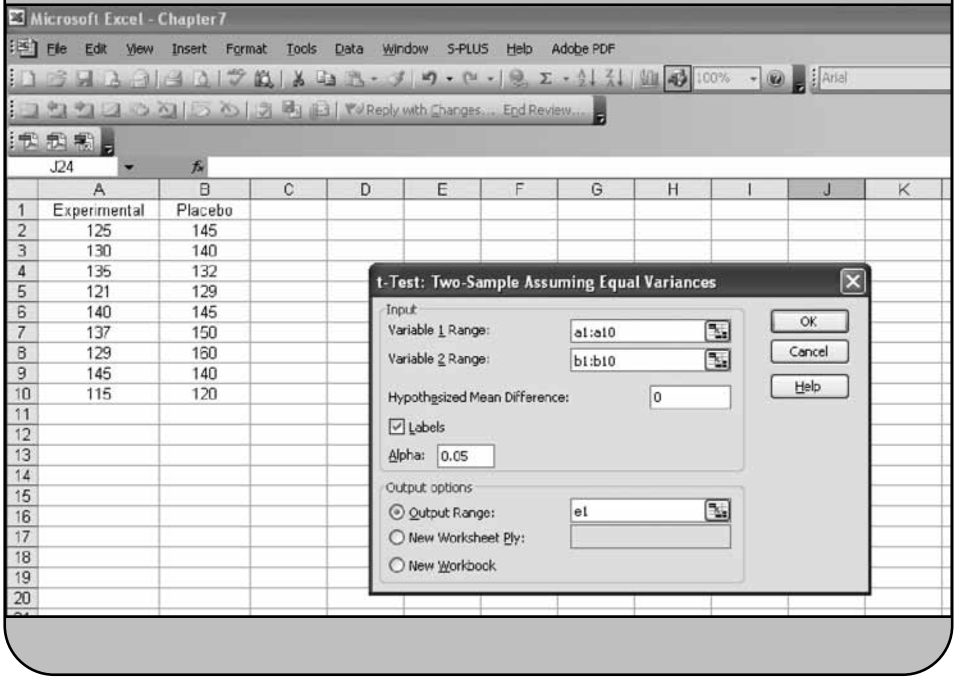
	A	B
1	Experimental	Placebo
2	125	145
3	130	140
4	135	132
5	121	129
6	140	145
7	137	150
8	129	160
9	145	140
10	115	120

Data Analysis

Analysis Tools

- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances

FIGURE 7–12 Specifications for Test


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
J24
OK
Cancel
Help

	A	B
1	Experimental	Placebo
2	125	145
3	130	140
4	135	132
5	121	129
6	140	145
7	137	150
8	129	160
9	145	140
10	115	120

t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range: a1:a10

Variable 2 Range: b1:b10

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

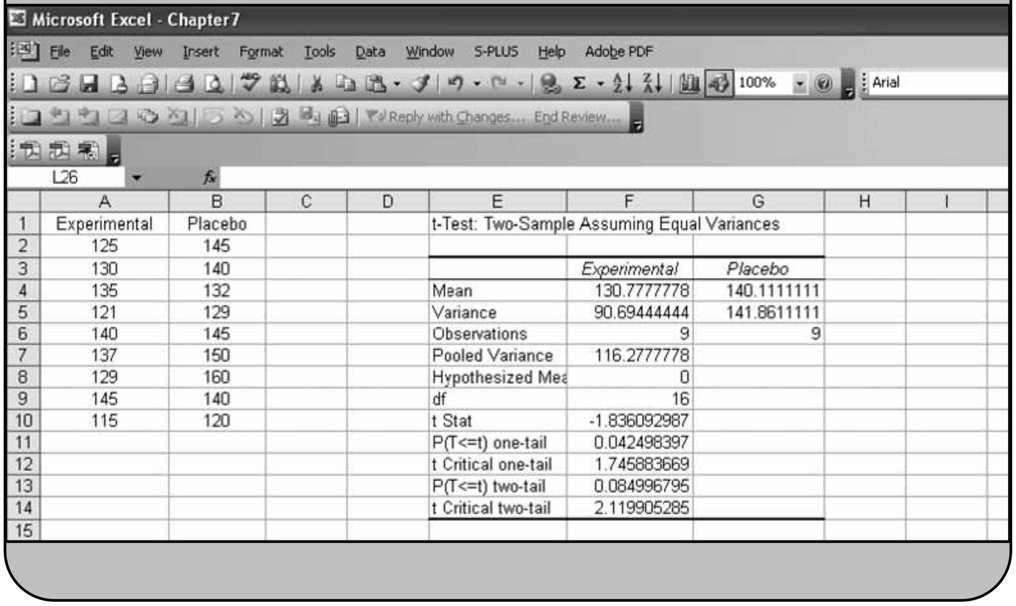
Output options

● Output Range: e1

○ New Worksheet Ply:

○ New Workbook

FIGURE 7–13 Results of Two-Independent-Samples Test for Difference in Means


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
L26
fx

	A	B	C	D	E	F	G	H	I
1	Experimental	Placebo			t-Test: Two-Sample Assuming Equal Variances				
2	125	145							
3	130	140				*Experimental*	*Placebo*		
4	135	132			Mean	130.7777778	140.1111111		
5	121	129			Variance	90.69444444	141.8611111		
6	140	145			Observations	9	9		
7	137	150			Pooled Variance	116.2777778			
8	129	160			Hypothesized Mea	0			
9	145	140			df	16			
10	115	120			t Stat	-1.836092987			
11					P(T<=t) one-tail	0.042498397			
12					t Critical one-tail	1.745883669			
13					P(T<=t) two-tail	0.084996795			
14					t Critical two-tail	2.119905285			
15									

In the dialog box, we specify the range of the data for each group. The data for variable 1 (experimental drug) is in cell A1 through cell A10 and the data for variable 2 (placebo) is in cell B1 through cell B10. Because we included the first row of labels (A1 and B1), we click the **Labels** box. We then specify the difference in means in the **Hypothesized Mean Differences** input field. For most situations, the difference is zero. We then specify the level of significance, $\alpha = 0.05$, and specify a location for the results. Finally, we request that Excel place the results in the current worksheet and we specify the top-left corner of the results table as E1. The results are shown in Figure 7–13.

The mean systolic blood pressure for patients on the experimental drug is 130.8 as compared to 140.1 for patients on the placebo. The estimate of the pooled variance is $S_p^2 = 116.3$, and the test statistic is $t = -1.84$. The two-sided p-value is 0.085, and thus we do not have significant evidence to show that there is a difference in mean blood pressures between treatments because $p = 0.085 > \alpha = 0.05$.

For two-independent-samples tests of means, as long as the variances are assumed to be equal, the ***t* Test: Two Sample Assuming Equal Variances** analysis tool can be used, regardless of the sample sizes (n_1 and n_2). If the sample sizes are large, Excel makes the appropriate adjustments for larger samples (essentially, uses a z statistic) and produces an appropriate test statistic and p-value. The test statistic is always labeled t, even for large samples.

7.5 TESTS WITH MATCHED SAMPLES, CONTINUOUS OUTCOME

For a two-dependent-samples test of hypothesis with a continuous outcome, the hypotheses are as follows:

$$H_0\text{: } \mu_d = 0$$

$$H_1\text{: } \mu_d > 0,\ \ H_1\text{: } \mu_d < 0,\ \text{ or } H_1\text{: } \mu_d \neq 0$$

where μ_d is the mean difference of the two dependent, matched, or paired populations.

In Chapter 7 of the textbook, we presented the following formulas for test statistics:

$$n \geq 30\text{:} \quad z = \frac{\bar{X}_d - \mu_d}{s_d / \sqrt{n}} \quad \text{(Find critical value of } z \text{ in Table 1C)}$$

$$n < 30\text{:} \quad t = \frac{\bar{X}_d - \mu_d}{s_d / \sqrt{n}} \quad \text{(Find critical value of } t \text{ in Table 2, } df = n - 1\text{)}$$

To perform the test of hypothesis by hand, we first compute difference scores, and then the sample size and mean and standard deviation of the difference scores in the sample. We then compute the test statistic and determine the appropriate critical values of z or t from Table 1C or Table 2 in the Appendix of the

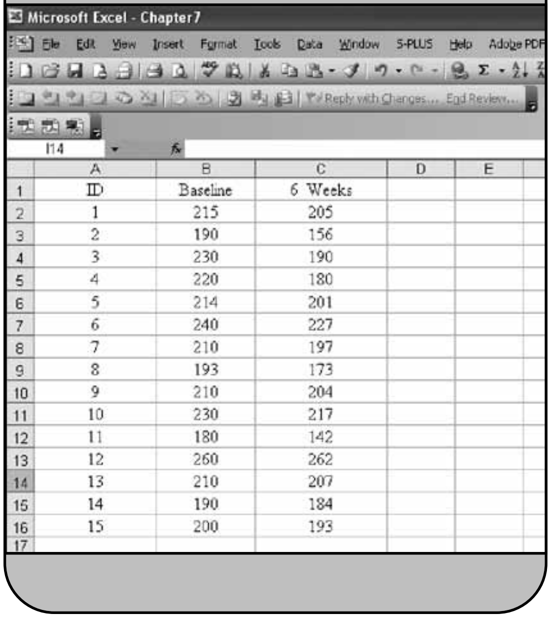
TABLE 7–2 Data from Cholesterol Study

Subject Identification Number	Baseline	6 Weeks
1	215	205
2	190	156
3	230	190
4	220	180
5	214	201
6	240	227
7	210	197
8	193	173
9	210	204
10	230	217
11	180	142
12	260	262
13	210	207
14	190	184
15	200	193

textbook, and compare the test statistic to the critical value to draw a conclusion.

Excel has a procedure to perform a two-dependent-samples test of means in its **Data Analysis ToolPak** called the ***t* Test: Paired Two Sample for Means**. We use this procedure to perform the test.

FIGURE 7–14 Data for Test of Mean Difference


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...
I14

	A	B	C	D	E
1	ID	Baseline	6 Weeks		
2	1	215	205		
3	2	190	156		
4	3	230	190		
5	4	220	180		
6	5	214	201		
7	6	240	227		
8	7	210	197		
9	8	193	173		
10	9	210	204		
11	10	230	217		
12	11	180	142		
13	12	260	262		
14	13	210	207		
15	14	190	184		
16	15	200	193		
17					

Example 7.6. In Example 7.11 of the textbook, we evaluated the efficacy of a new drug for lowering cholesterol. Fifteen patients had a pretreatment or baseline total cholesterol level measured, and then after taking the drug for 6 weeks, each patient's total cholesterol level was measured again. The data are shown in Table 7–2.

The hypotheses are as follows:

$$H_0: \mu_d = 0$$

$$H_1: \mu_d > 0$$

$$\alpha = 0.05$$

Because the sample size is small, the appropriate test statistic is $t = \dfrac{\bar{X}_d - \mu_d}{s_d/\sqrt{n}}$.

We first enter the data into an Excel worksheet, as shown in Figure 7–14. Under the **Tools/Data Analysis** option, we choose the ***t* Test: Paired Two-Sample For Means** analysis tool as shown in Figure 7–15. Once we click **OK**, Excel presents the dialog box shown in Figure 7–16.

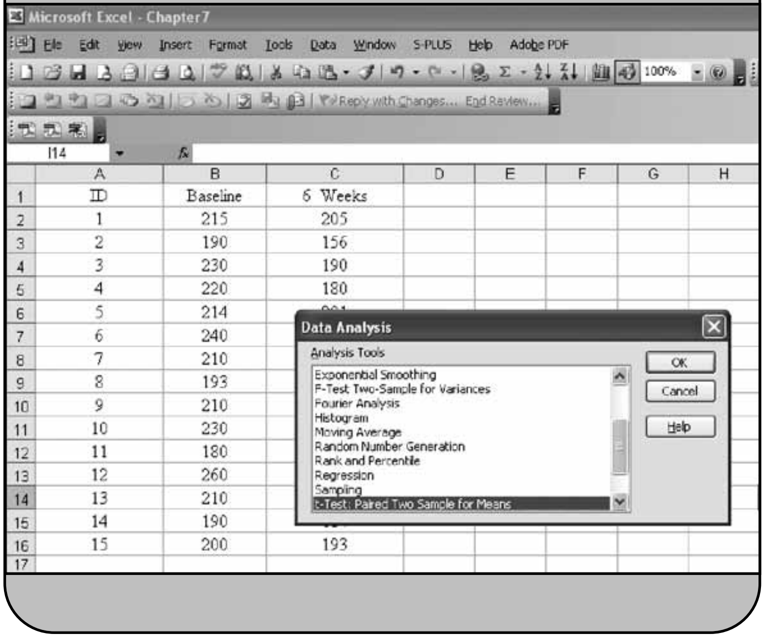
In the dialog box, we specify the range of the data for each measurement. The first measurement on each participant (variable 1) is the baseline measurement, and these values are in cell B1 through cell B16. The second measurement on each participant (variable 2) is the measurement taken at 6 weeks and these values are in cell C1 through cell C16. Because we included the first row of labels (B1 and C1), we click the **Labels** box. We then specify the difference in means in the **Hypothesized Mean Differences** input field. For most situations, the mean difference is zero. We then specify the level of significance, $\alpha = 0.05$, and specify a location for the results. In Figure 7–16, we specify the top-left corner of the results table as E1. The results are shown in Figure 7–17.

The mean cholesterol level at baseline is 212.8 and the mean cholesterol level at 6 weeks is 195.9. The two-dependent-samples test is based on difference scores (see Chapter 7 of the textbook for details). The test statistic is $t = 4.63$ and the one-sided p-value is 0.0002. We reject H_0 because $p = 0.0002 < \alpha = 0.05$. We have statistically significant evidence at $\alpha = 0.05$ to show that there is a reduction in cholesterol levels over 6 weeks. Notice that Excel does not produce summary statistics on the difference scores (i.e., $\bar{X}_d$, s_d). However, these values are used in the computation of the test statistic.

7.6 TESTS WITH TWO INDEPENDENT SAMPLES, DICHOTOMOUS OUTCOME

For a two-independent-samples test of a hypothesis with a dichotomous outcome, the hypotheses are as follows:

FIGURE 7–15 Two-Matched-Samples Test Data Analysis Tool


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...
I14
fx

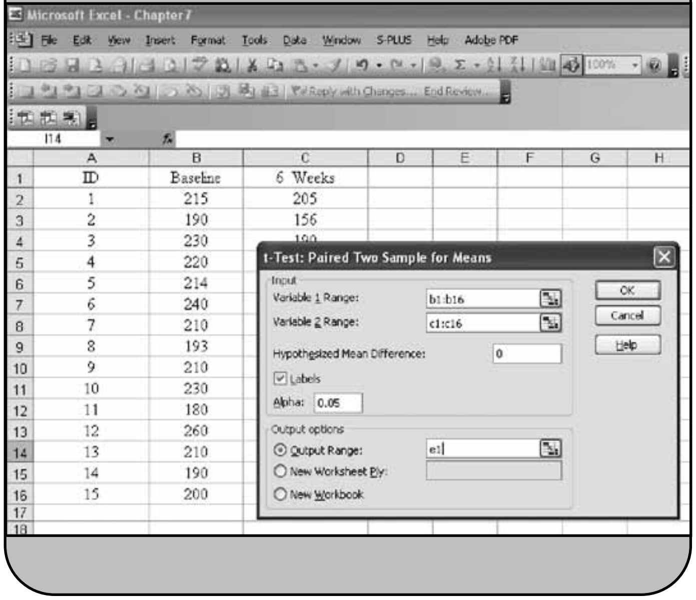
	A	B	C	D	E	F	G	H
1	ID	Baseline	6 Weeks					
2	1	215	205					
3	2	190	156					
4	3	230	190					
5	4	220	180					
6	5	214						
7	6	240						
8	7	210						
9	8	193						
10	9	210						
11	10	230						
12	11	180						
13	12	260						
14	13	210						
15	14	190						
16	15	200	193					
17								

Data Analysis

Analysis Tools

- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means

FIGURE 7–16 Specifications for Test


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...
I14
fx

	A	B	C	D	E	F	G	H
1	ID	Baseline	6 Weeks					
2	1	215	205					
3	2	190	156					
4	3	230						
5	4	220						
6	5	214						
7	6	240						
8	7	210						
9	8	193						
10	9	210						
11	10	230						
12	11	180						
13	12	260						
14	13	210						
15	14	190						
16	15	200						
17								
18								

t-Test: Paired Two Sample for Means

Input

Variable 1 Range: b1:b16

Variable 2 Range: c1:c16

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

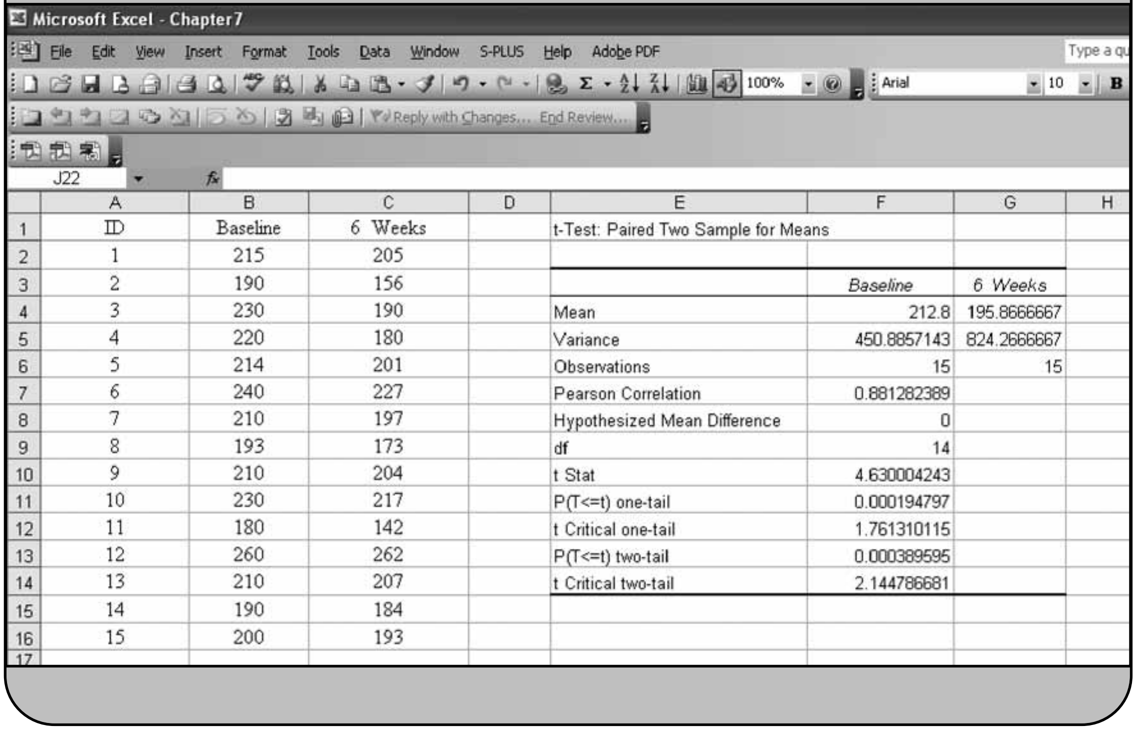
Output options

● Output Range: e1

○ New Worksheet Ply:

○ New Workbook

FIGURE 7–17 Results of Test for Mean Difference



Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a qu
100%
Arial
10
B
Reply with Changes... End Review...
J22
fx

	A	B	C	D	E	F	G	H
1	ID	Baseline	6 Weeks		t-Test: Paired Two Sample for Means			
2	1	215	205					
3	2	190	156			Baseline	6 Weeks	
4	3	230	190		Mean	212.8	195.8666667	
5	4	220	180		Variance	450.8857143	824.2666667	
6	5	214	201		Observations	15	15	
7	6	240	227		Pearson Correlation	0.881282389		
8	7	210	197		Hypothesized Mean Difference	0		
9	8	193	173		df	14		
10	9	210	204		t Stat	4.630004243		
11	10	230	217		P(T<=t) one-tail	0.000194797		
12	11	180	142		t Critical one-tail	1.761310115		
13	12	260	262		P(T<=t) two-tail	0.000389595		
14	13	210	207		t Critical two-tail	2.144786681		
15	14	190	184					
16	15	200	193					
17								

$$H_0: p_1 = p_2$$

$$H_1: p_1 > p_2,\ H_1: p_1 < p_2,\ \text{or } H_1: p_1 \neq p_2$$

where p_1 and p_2 are the proportions of successes in the two populations of interest.

In Chapter 7 of the textbook, we presented the following test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1 - \hat{p})(1/n_1 + 1/n_2)}} \qquad \text{(Find critical value of } z \text{ in Table 1C)}$$

where $\hat{p}_1$ is the proportion of successes in sample 1, $\hat{p}_2$ is the proportion of successes in sample 2, and $\hat{p}$ is the proportion of successes in the pooled sample, $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$. Excel does not have a specific analysis tool to perform this test. We instead use Excel to compute the z statistic and the p-value.

Example 7.7. In Example 7.13 of the textbook, we analyzed data from a randomized trial designed to evaluate the effectiveness of a newly developed pain reliever as compared to a standard pain relievers in reducing pain in patients following joint replacement surgery. A total of 100 patients undergoing joint replacement surgery agreed to participate in the trial and were randomly assigned to receive either the new pain reliever or the standard pain reliever following surgery, and were blind to the treatment assignment. Before receiving the assigned treatment, patients were asked to rate their pain on a scale of 0 to 10, with higher scores indicative of more pain. Each patient was then given the assigned treatment and after 30 minutes was again asked to rate their pain on the same scale. The primary outcome was a reduction in pain of 3 or more scale points (defined by clinicians as a clinically meaningful reduction). The data in Table 7–3 were observed in the trial.

We use Excel to test whether there is a statistically significant difference in the proportions of patients reporting a meaningful reduction (i.e., a reduction of 3 or more scale points). The hypotheses are as follows:

TABLE 7–3 Data from Clinical Trial

		Reduction of 3 + Points	
Treatment Group	***n***	**Number**	**Proportion**
New pain reliever	50	23	0.46
Standard pain reliever	50	11	0.22

$$H_0\text{: } p_1 = p_2$$

$$H_1\text{: } p_1 \neq p_2$$

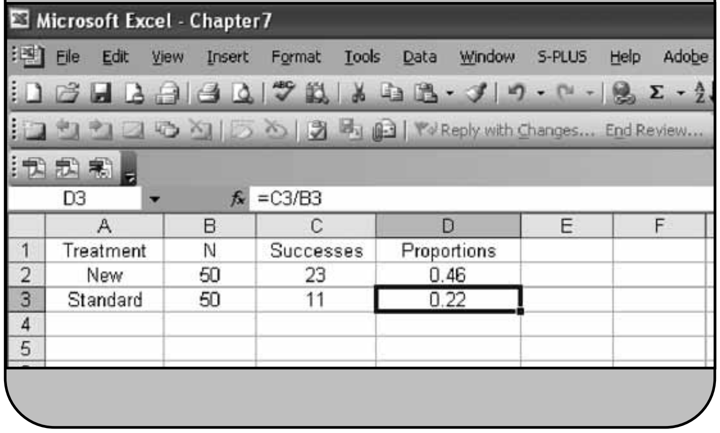
$$\alpha = 0.05$$

The appropriate test statistic is $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1-\hat{p})(1/n_1 + 1/n_2)}}$.

We now use Excel to compute the test statistic and the p-value. The data are entered into an Excel worksheet, as shown in Figure 7–18. The sample proportions are computed by dividing the numbers of successes (column C) by the sample sizes (column B) in each group.

Before computing the test statistic, z, we need to compute the overall proportion. This is placed in cell D5 and is computed as "=(C2+C3)/(B2+B3)". We next compute the test statistic, z, and place it in cell D7. The formula is "=(D2−D3)/SQRT(D5*(1−D5)*(1/B2+1/B3))". The overall proportion and z statistic are shown in Figure 7–19. The last step involves computing the two-sided p-value using the NORMSDIST function as "=2*(1− NORMSDIST(ABS(z)))". The p-value is shown in Figure 7–20.

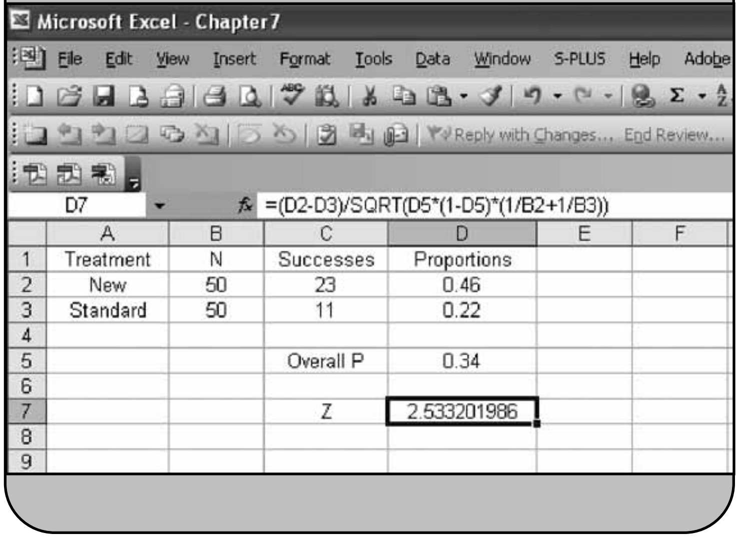
FIGURE 7–18 Data for Two-Independent-Samples Test for Difference in Proportions


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe
Reply with Changes... End Review...

D3 =C3/B3

	A	B	C	D	E	F
1	Treatment	N	Successes	Proportions		
2	New	50	23	0.46		
3	Standard	50	11	0.22		
4						
5						

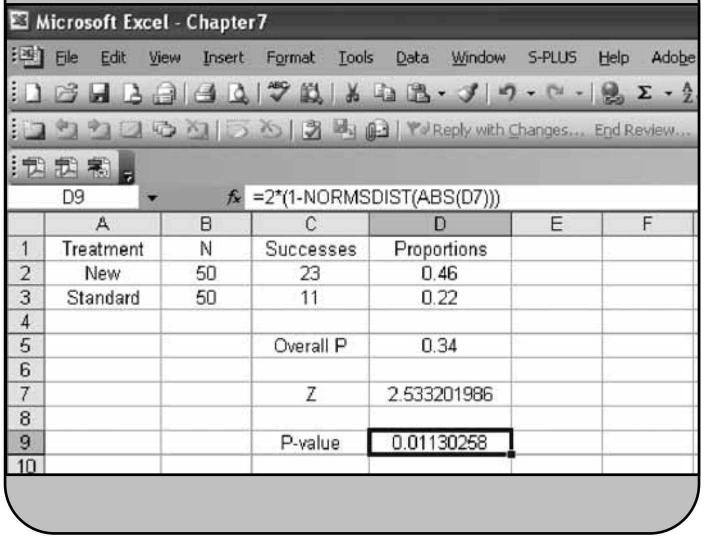
FIGURE 7–19 Test Statistic


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe
Reply with Changes... End Review...

D7 =(D2-D3)/SQRT(D5*(1-D5)*(1/B2+1/B3))

	A	B	C	D	E	F
1	Treatment	N	Successes	Proportions		
2	New	50	23	0.46		
3	Standard	50	11	0.22		
4						
5			Overall P	0.34		
6						
7			Z	2.533201986		
8						
9						

FIGURE 7–20 *p*-Value


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe
Reply with Changes... End Review...

D9 =2*(1-NORMSDIST(ABS(D7)))

	A	B	C	D	E	F
1	Treatment	N	Successes	Proportions		
2	New	50	23	0.46		
3	Standard	50	11	0.22		
4						
5			Overall P	0.34		
6						
7			Z	2.533201986		
8						
9			P-value	0.01130258		
10						

In this test, the test statistic is $z = 2.53$ and we reject H_0 because $p = 0.011 < \alpha = 0.05$. We have statistically significant evidence at $\alpha = 0.05$ to show that there is a difference in the proportions of patients on the new pain reliever reporting a meaningful reduction (i.e., a reduction of 3 or more scale points) as compared to patients on the standard pain reliever.

TABLE 7–4 Data from Clinical Trial

Low-Calorie	Low-Fat	Low-Carbohydrate	Control
8	2	3	2
9	4	5	2
6	3	4	−1
7	5	2	0
3	1	3	3

7.7 TESTS WITH MORE THAN TWO INDEPENDENT SAMPLES, CONTINUOUS OUTCOME: ANALYSIS OF VARIANCE

In analysis of variance, the hypotheses are as follows:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_1\text{: Means are not all equal}$$

where μ_j is the mean in the jth group and k is the number of independent comparison groups.

In Chapter 7 of the textbook, we presented the test statistic for analysis of variance as:

$$F = \frac{\sum n_j(\overline{X}_j - \overline{X})^2/(k-1)}{\sum\sum(X - \overline{X}_j)^2/(N-k)}$$

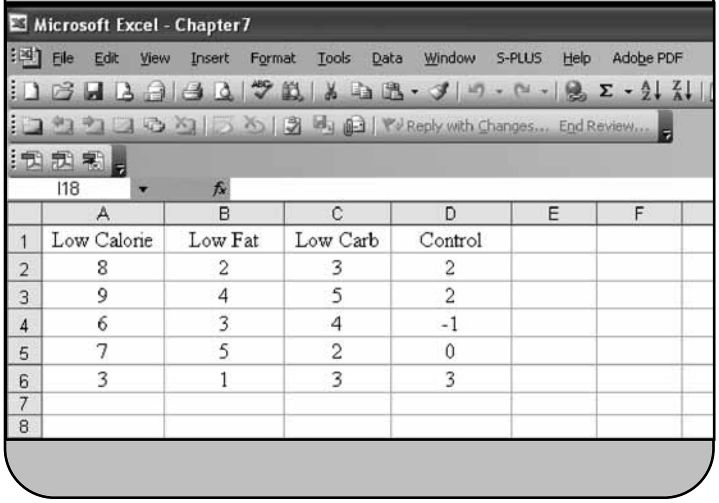
(Find critical value in Table 4, $df_1 = k - 1$, $df_2 = N - k$)

where n_j is the sample size in the jth group, $\overline{X}_j$ is the sample mean in the jth group, and $\overline{X}$ is the overall mean. k represents the number of independent groups ($k > 2$), and N represents the total number of observations in the analysis.

Example 7.8. In Example 7.14 of the textbook, we analyzed data from a clinical trial comparing four weight-loss programs. The outcome of interest was weight loss, defined as the difference in weight at the start of the study (baseline) and weight at the end of the study (8 weeks), in pounds. A total of 20 patients agreed to participate in the study and were randomly assigned to one of the four diet groups. The data are shown in Table 7–4.

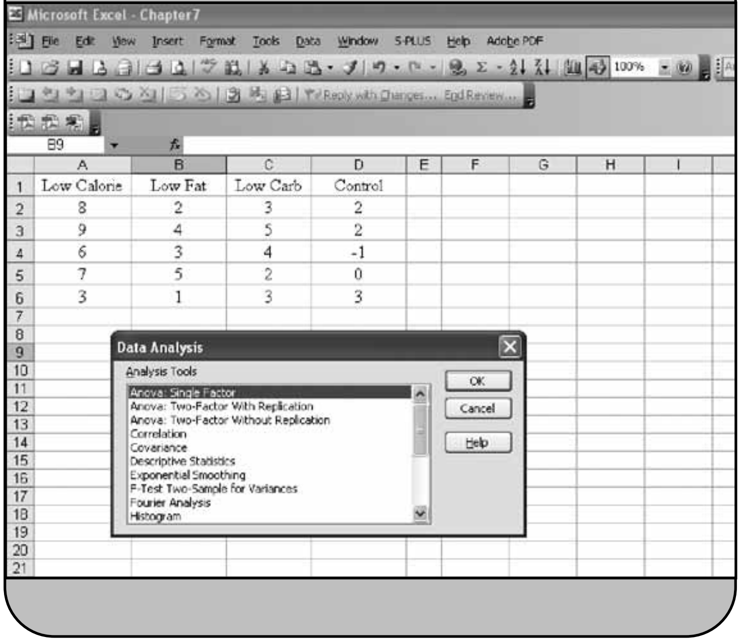
We use Excel to conduct an ANOVA to test whether there is a statistically significant difference in the mean weight loss among the four diets. Excel has an analysis tool to perform an

FIGURE 7–21 Data for ANOVA


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...
I18
fx

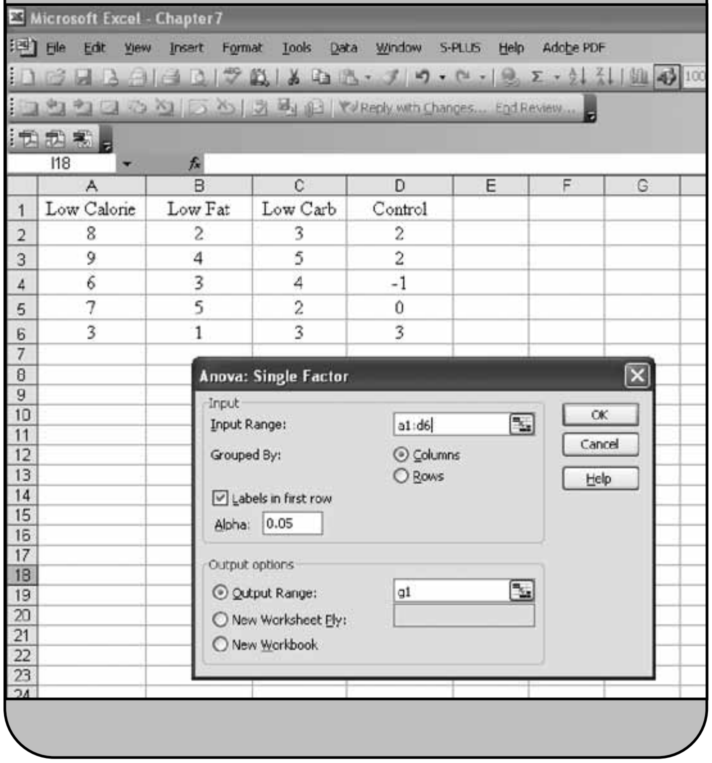
	A	B	C	D	E	F
1	Low Calorie	Low Fat	Low Carb	Control		
2	8	2	3	2		
3	9	4	5	2		
4	6	3	4	-1		
5	7	5	2	0		
6	3	1	3	3		
7						
8						

FIGURE 7–22 ANOVA Data Analysis Tool


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...
B9
fx
Data Analysis
Analysis Tools
Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
OK
Cancel
Help

	A	B	C	D	E	F	G	H	I
1	Low Calorie	Low Fat	Low Carb	Control					
2	8	2	3	2					
3	9	4	5	2					
4	6	3	4	-1					
5	7	5	2	0					
6	3	1	3	3					

FIGURE 7–23 Specifications for ANOVA


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...
I18
Anova: Single Factor
OK
Cancel
Help

	A	B	C	D	E	F	G
1	Low Calorie	Low Fat	Low Carb	Control			
2	8	2	3	2			
3	9	4	5	2			
4	6	3	4	-1			
5	7	5	2	0			
6	3	1	3	3			

Input

Input Range: a1:d6

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Alpha: 0.05

Output options

● Output Range: g1

○ New Worksheet Ply:

○ New Workbook

analysis of variance in its Data Analysis ToolPak. We first enter the data into an Excel worksheet, as shown in Figure 7–21. Under the **Tools/Data Analysis** option, we choose the **ANOVA: Single Factor** analysis tool shown in Figure 7–22. Once we click **OK**, Excel presents the dialog box shown in Figure 7–23.

In the dialog box, we specify the range of the data. For this analysis, we specify the range of cells containing data on all groups. The data are in cell A1 through cell D6. Because we included the first row of labels (A1 through D1), we click the Labels box. We then specify the level of significance, $\alpha = 0.05$, and a location for the results. In Figure 7–23, we request that Excel place the results in the current worksheet and we specify the top-left corner of the results table as G1. The results are shown in Figure 7–24.

Excel first provides summary statistics on each group. Specifically, Excel shows the count (n), sum (ΣX), average ($\overline{X}$), and variance (s^2) in each group. Excel then generates an ANOVA table similar to the one we produced in Chapter 7 of the textbook when conducting an ANOVA by hand. Excel provides sums of squares (SS), degrees of freedom (df), mean squares (MS = SS / df), and the test statistic, F. Excel also produces the p-value and the critical value of F for a test with a 5% level of significance. The p-value is 0.0013, and thus we reject H_0. We have statistically significant evidence at $\alpha = 0.05$ to show that there is a difference in mean weight loss among the four diets.

7.8 TESTS FOR TWO OR MORE INDEPENDENT SAMPLES, CATEGORICAL AND ORDINAL OUTCOMES: THE CHI-SQUARE TEST OF INDEPENDENCE

For a χ^2 test of independence, the hypotheses are:

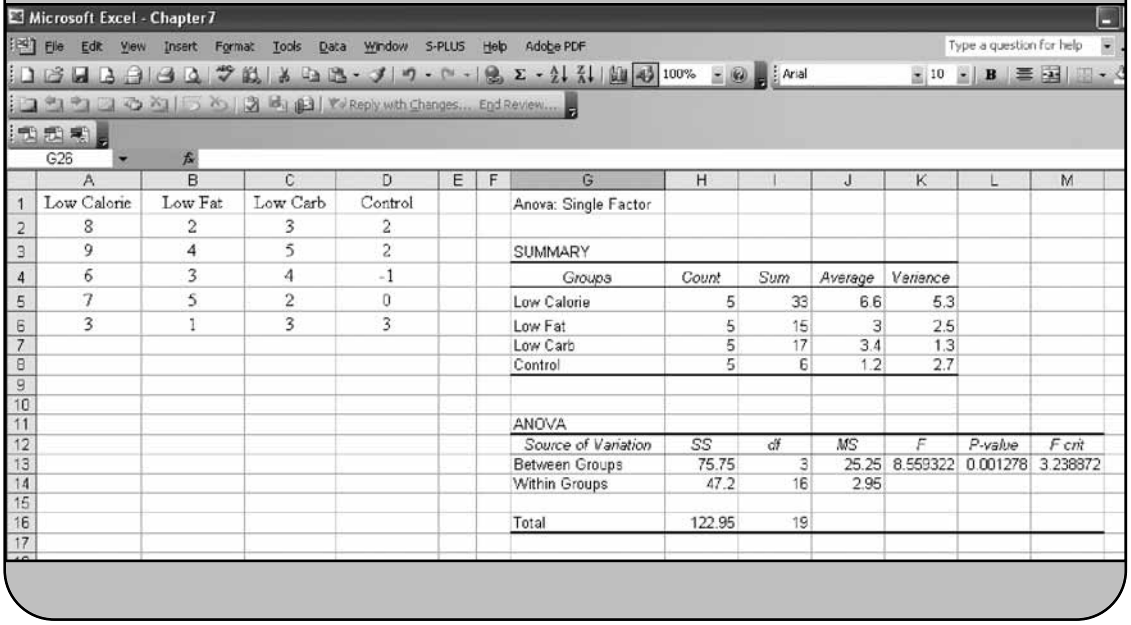
H_0: The distribution of the outcome is
independent of the groups

H_1: H_0 is false

In Chapter 7 of the textbook, we presented the following for the test statistic:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \qquad \text{[Find critical value in Table 3, } df = (r-1)(c-1)\text{]}$$

FIGURE 7–24 Results of ANOVA


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a question for help
100%
Arial
10
B
Reply with Changes... End Review...
G26
fx

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Low Calorie	Low Fat	Low Carb	Control			Anova: Single Factor						
2	8	2	3	2									
3	9	4	5	2			SUMMARY						
4	6	3	4	-1			*Groups*	*Count*	*Sum*	*Average*	*Verience*		
5	7	5	2	0			Low Calorie	5	33	6.6	5.3		
6	3	1	3	3			Low Fat	5	15	3	2.5		
7							Low Carb	5	17	3.4	1.3		
8							Control	5	6	1.2	2.7		
9													
10													
11							ANOVA						
12							*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
13							Between Groups	75.75	3	25.25	8.559322	0.001278	3.238872
14							Within Groups	47.2	16	2.95			
15													
16							Total	122.95	19				
17													

where O = observed frequency (i.e., sample data), E = expected frequency in each of the cells of the table, r = the number of rows in the two-way table, and c = the number of columns in the two-way table (where r and c correspond to the number of comparison groups and the number of response options in the outcome).

When performing the χ^2 test of independence by hand, we compute the expected frequencies for each cell and then compute the test statistic. We then use Table 3 in the Appendix of the textbook to find the appropriate critical value from the χ^2 distribution and compare the test statistic to the critical value to draw a conclusion.

Excel does not have a specific analysis tool to perform the χ^2 test of independence. We will use the CHIDIST function to produce p-values. We used this function for the χ^2 goodness-of-fit test as "=CHIDIST(χ^2, df)".

Again, to use the CHIDIST function we specify the test statistic, χ^2, and the degrees of freedom, df. For the χ^2 test of independence, $df = (r - 1) \times (c - 1)$. The CHIDIST function returns the area in the right tail of the distribution, which is the p-value for the χ^2 test of independence. We now use Excel to conduct a test of independence.

Example 7.9. In Example 7.4 of the Excel workbook, we examined data from a survey of university graduates that assessed, among other things, how frequently students exercised. Here we want to test whether there is a relationship between exercise and students' living arrangements. The data are shown in Table 7–5.

TABLE 7–5 Data from University Survey

	No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Dormitory	32	30	28	90
On-campus apartment	74	64	42	180
Off-campus apartment	110	25	15	150
At home	39	6	5	50
Total	255	125	90	470

The hypotheses are as follows:

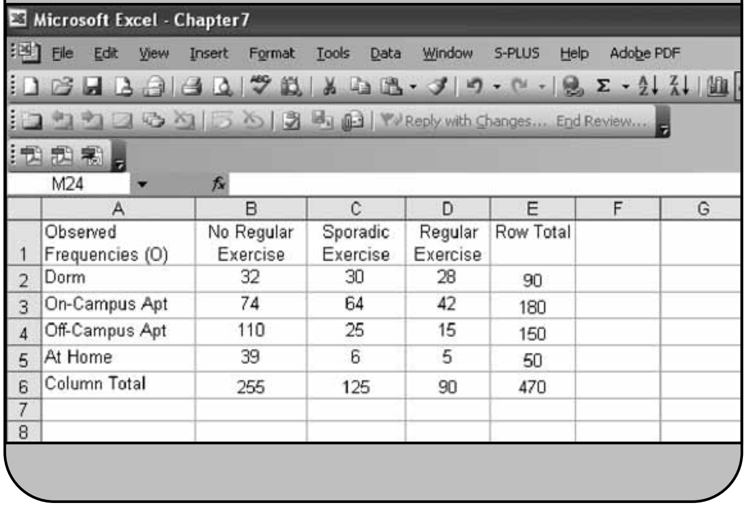
H_0: Living arrangement and exercise are independent

H_1: H_0 is false

$$\alpha = 0.05$$

The appropriate test statistic is $\chi^2 = \sum \frac{(O-E)^2}{E}$.

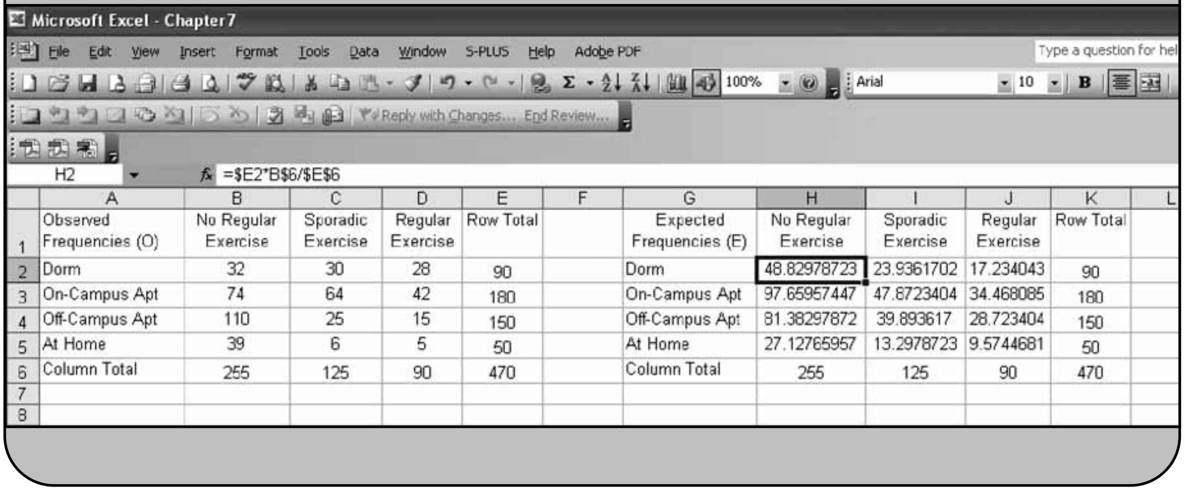
FIGURE 7–25 Data for Chi-Square Test of Independence


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

M24

	A	B	C	D	E	F	G
1	Observed Frequencies (O)	No Regular Exercise	Sporadic Exercise	Regular Exercise	Row Total		
2	Dorm	32	30	28	90		
3	On-Campus Apt	74	64	42	180		
4	Off-Campus Apt	110	25	15	150		
5	At Home	39	6	5	50		
6	Column Total	255	125	90	470		
7							
8							

FIGURE 7–26 Expected Frequencies


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a question for hel
100%
Arial
10
Reply with Changes... End Review...

H2 =$E2*B$6/$E$6

	A	B	C	D	E	F	G	H	I	J	K	L
1	Observed Frequencies (O)	No Regular Exercise	Sporadic Exercise	Regular Exercise	Row Total		Expected Frequencies (E)	No Regular Exercise	Sporadic Exercise	Regular Exercise	Row Total	
2	Dorm	32	30	28	90		Dorm	48.82978723	23.9361702	17.234043	90	
3	On-Campus Apt	74	64	42	180		On-Campus Apt	97.65957447	47.8723404	34.468085	180	
4	Off-Campus Apt	110	25	15	150		Off-Campus Apt	81.38297872	39.893617	28.723404	150	
5	At Home	39	6	5	50		At Home	27.12765957	13.2978723	9.5744681	50	
6	Column Total	255	125	90	470		Column Total	255	125	90	470	
7												
8												

Recall that the expected frequencies (*E*) are computed based on the assumption that H_0 is true. The data for the test are entered into an Excel worksheet, as shown in Figure 7–25. The sample data (i.e., the numbers of students in each cell of the table) are the observed frequencies in the table. The row and column totals are computed using the SUM function. For example, the column total in cell B6 is computed using "=SUM(B2:B5)". The row total in cell E2 is computed using "=SUM(B2:D2)". The grand total in cell E6 is computed by summing the column totals using "=SUM(B6:D6)". This is equivalent to summing the row totals using "=SUM(E2:E5)".

We next compute the expected frequencies using the following:

$$\text{Expected cell frequency} = \frac{\text{Row total} \times \text{column total}}{N}$$

The expected frequencies are placed in another table, as shown in Figure 7–26.

To compute the expected frequencies, we need to use both relative and absolute cell addresses. For example, the expected frequency in cell H2 is computed using "=$E2*B$6/$E$6". The row total is referenced as $E2.

We place a “\$” in front of the column to freeze the column address on row totals that are contained in column E. We do the same for the column total, B\$6, except we freeze the row address on row 6, which contains the column totals. The total sample size is in cell \$E\$6. When we copy this formula from cell H2 to cell H3, for example, the formula is updated to “=\$E3*B\$6/\$E\$6”. The sums of the expected frequencies across rows and down columns are equal to the sums of the observed frequencies across rows and down columns.

We now compute $(O - E)^2 / E$ in each cell of the table. Once we compute these, we sum to produce the χ^2 statistic. The $(O - E)^2$ / E values for each cell are shown in Figure 7–27. For example, in cell F9 the formula is “=(B2−H2)^2/H2”. When we copy this formula to the other cells in the bottom table, the cell references are automatically updated (e.g., the formula in cell H12 is “=(D5−J5)^2/J5”).

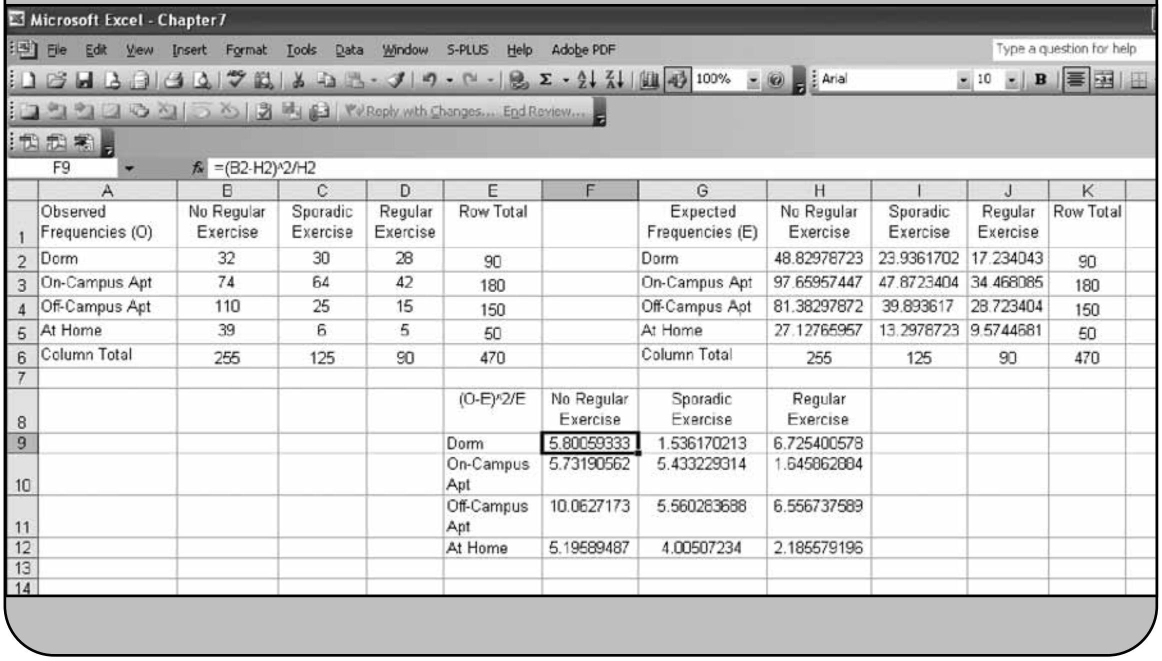
The χ^2 test statistic is computed by summing the $(O - E)^2 / E$ values in the twelve cells. The test statistic is placed in cell G14 and is computed using “=SUM(F9:H12)”. The *p*-value is computed with the CHIDIST function using “=CHIDIST (G14,6)”, where “6” reflects the degrees of freedom, $df = (r - 1) \times (c - 1) = 3(2) = 6$. The test statistic and *p*-value are shown in Figure 7–28.

In this test, the test statistic is $\chi^2 = 60.44$ and the *p*-value is practically zero (3.66×10^{-11}). We reject H_0 because $p = 0 < \alpha = 0.05$. We have statistically significant evidence at $\alpha = 0.05$ to show that H_0 is false, or that living arrangement and exercise are not independent (i.e., they are dependent or related).

7.9 PRACTICE PROBLEMS

1. Data are collected in a clinical trial evaluating a new compound designed to improve wound healing in trauma patients. The new compound is compared against a placebo. After treatment for 5 days with the new compound or placebo, the extent of wound healing is measured and the data are shown in Table 7–6. Is there a difference in the extent of wound healing by treatment? (*Hint:* Are treatment and the percent wound healing independent?) Run the appropriate test at a 5% level of significance.
2. Use the data in Problem 1 and pool the data across the treatments into one sample of size $n = 250$. Use the pooled data to test whether the distribution of the percent wound healing is approximately normal. Specifically, use the following distribution: 30%,

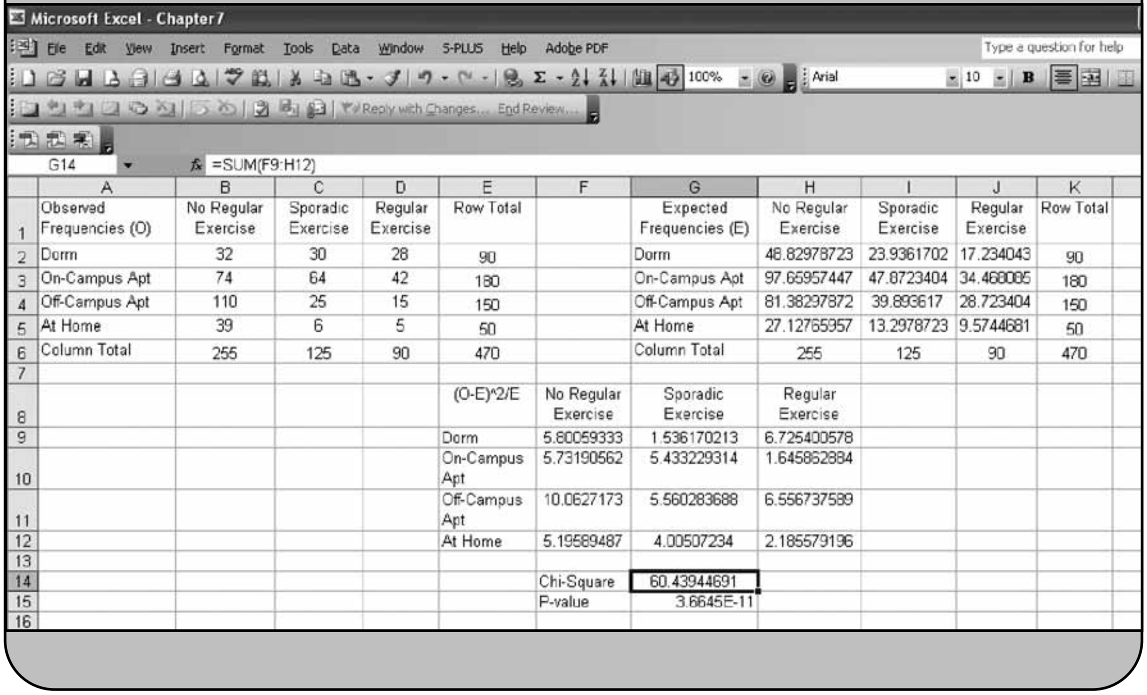
FIGURE 7–27 Computing the Test Statistic


Microsoft Excel - Chapter7
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a question for help
100%
Arial
10
B
Reply with Changes... End Review...

F9 =(B2-H2)^2/H2

	A	B	C	D	E	F	G	H	I	J	K
1	Observed Frequencies (O)	No Regular Exercise	Sporadic Exercise	Regular Exercise	Row Total		Expected Frequencies (E)	No Regular Exercise	Sporadic Exercise	Regular Exercise	Row Total
2	Dorm	32	30	28	90		Dorm	48.82978723	23.9361702	17.234043	90
3	On-Campus Apt	74	64	42	180		On-Campus Apt	97.65957447	47.8723404	34.468085	180
4	Off-Campus Apt	110	25	15	150		Off-Campus Apt	81.38297872	39.893617	28.723404	150
5	At Home	39	6	5	50		At Home	27.12765957	13.2978723	9.5744681	50
6	Column Total	255	125	90	470		Column Total	255	125	90	470
7											
8					(O-E)^2/E	No Regular Exercise	Sporadic Exercise	Regular Exercise			
9					Dorm	5.80059333	1.536170213	6.725400578			
10					On-Campus Apt	5.73190562	5.433229314	1.645862884			
11					Off-Campus Apt	10.0627173	5.560283688	6.556737589			
12					At Home	5.19589487	4.00507234	2.185579196			
13											
14											

FIGURE 7–28 Test Statistic and *p*-Value



G14 =SUM(F9:H12)

	A	B	C	D	E	F	G	H	I	J	K
1	Observed Frequencies (O)	No Regular Exercise	Sporadic Exercise	Regular Exercise	Row Total		Expected Frequencies (E)	No Regular Exercise	Sporadic Exercise	Regular Exercise	Row Total
2	Dorm	32	30	28	90		Dorm	48.82978723	23.9361702	17.234043	90
3	On-Campus Apt	74	64	42	180		On-Campus Apt	97.65957447	47.8723404	34.468085	180
4	Off-Campus Apt	110	25	15	150		Off-Campus Apt	81.38297872	39.893617	28.723404	150
5	At Home	39	6	5	50		At Home	27.12765957	13.2978723	9.5744681	50
6	Column Total	255	125	90	470		Column Total	255	125	90	470
7											
8					(O-E)^2/E	No Regular Exercise	Sporadic Exercise	Regular Exercise			
9					Dorm	5.80059333	1.536170213	6.725400578			
10					On-Campus Apt	5.73190562	5.433229314	1.645862884			
11					Off-Campus Apt	10.0627173	5.560283688	6.556737589			
12					At Home	5.19589487	4.00507234	2.185579196			
13											
14						Chi-Square	60.43944691				
15						P-value	3.6645E-11				
16											

40%, 20%, and 10% and $\alpha = 0.05$ to run the appropriate test.

3. Data are collected in an experiment designed to investigate the impact of different positions of the mother during ultrasound on fetal heart rate. Fetal heart rate is measured by ultrasound in beats per minute. The study includes 20 women who are assigned to one position and have the fetal heart rate measured in that position. Each woman is between 28 weeks and 32 weeks gestation. The data are shown in Table 7–7. Is there a significant difference in mean fetal heart rates by position? Run the test at a 5% level of significance.
4. A clinical trial is conducted comparing a new pain reliever for arthritis to a placebo. Participants are randomly assigned to receive the new treatment or a placebo, and the outcome is pain relief within 30 minutes. The data are shown in Table 7–8. Is there a significant difference in the proportions of patients reporting pain relief? Run the test at a 5% level of significance.

TABLE 7–6 Data for Practice Problems 1 and 2

	Percent Wound Healing			
Treatment	**0–25**	**26–50**	**51–75**	**76–100**
New compound ($n = 125$)	15	37	32	41
Placebo($n = 125$)	36	45	34	10

TABLE 7–7 Data for Practice Problem 3

Back	Side	Sitting	Standing
140	141	144	147
144	143	145	145
146	145	147	148
141	144	148	149
139	136	144	145

TABLE 7–8 Data for Practice Problem 4

	Pain Relief	No Pain Relief
New medication	44	76
Placebo	21	99

5. A clinical trial is planned to compare an experimental medication designed to lower blood pressure to a placebo. Before starting the trial, a pilot study is conducted involving 7 participants. The objective of the study is to assess how systolic blood pressure changes over time untreated. Systolic blood pressures are measured at baseline and again 4 weeks later. Is there a statistically significant difference in blood pressures over time? Run the test at a 5% level of significance.

Baseline:	120	145	130	160	152	143	126
4 Weeks:	122	142	135	158	155	140	130

6. A hypertension trial is mounted and 12 participants are randomly assigned to receive either a new medication or a placebo. Each participant takes the assigned medication and their systolic blood pressure (SBP) is recorded after 6 months on the assigned medication. The data are shown in Table 7–9. Is there a difference in mean SBP between treatments? Run the appropriate test at $\alpha = 0.05$.

TABLE 7–9 Data for Practice Problem 6

Placebo	New Medication
134	114
143	117
148	121
142	124
150	122
160	128

CHAPTER 8

Power and Sample Size Determination

In Chapter 8 in the textbook, we presented various formulas to determine the sample size for statistical inference. In applications where the goal is to generate a confidence interval estimate for an unknown parameter, the sample size is computed to ensure that the margin of error is sufficiently small. In applications where the goal is to perform a test of hypothesis, the sample size is computed to ensure that the test has a high probability of rejecting the null hypothesis when it is false (in other words, to ensure that the test has high power).

Microsoft Office® Excel® does not have a specific analysis tool for determining sample size. However, we can use Excel's probability functions to implement the sample size formulas presented in Chapter 8.

8.1 SAMPLE SIZE ESTIMATES FOR CONFIDENCE INTERVALS WITH A CONTINUOUS OUTCOME IN ONE SAMPLE

In Chapter 8 of the textbook, we presented the following formula to estimate the sample size required to estimate the mean of a continuous outcome variable in a single population:

$$n = \left(\frac{z\sigma}{E}\right)^2$$

where z is the value from the standard normal distribution reflecting the confidence level that will be used (e.g., $z = 1.96$ for 95%), σ is the standard deviation of the outcome variable, and E is the desired margin of error. The preceding formula generates the minimum number of subjects required to ensure that the margin of error in the confidence interval for μ does not exceed E.

To determine sample size requirements with Excel, we use the NORMSINV function to compute z values for confidence intervals as "=NORMSINV(*lower-tail area*)". To use this function for computing sample sizes, we specify the area under the curve in the lower tail of the standard normal distribution. For example, for a 95% confidence interval, the area in the lower tail is 0.975. Figure 8–1 shows the standard normal distribution, z, and the z values that hold the middle 95% of the distribution, $P(-1.96 \leq X \leq 1.96) = 0.95$.

To use the NORMSINV function for computing sample sizes, we specify the probability in the lower tail of the standard normal distribution for the desired confidence level. For example, if a 95% confidence interval is planned, we specify "=NORMSINV(0.975)", which returns 1.96. If a 90% confidence interval is planned, we specify "=NORMSINV(0.95)", which returns 1.645.

Example 8.1. In Example 8.1 of the textbook, we determined the sample size required to estimate the mean systolic blood pressure in children with congenital heart disease who were between 3 and 5 years of age. The analysis was planned to estimate a 95% confidence interval and the investigator decided that a margin of error of 5 units was sufficiently precise. To determine the sample size, the standard deviation was assumed to be 20. The margin of error, standard deviation, and confidence level are input into Excel as shown in Figure 8–2. The z value is estimated using the NORMSINV function as shown in Figure 8–2.

Recall that the argument for the NORMSINV function is the area in the lower tail of the standard normal curve (Figure 8–1). If a 95% confidence interval is planned, the area in the lower tail is computed by first determining the total tail area as $(1 - 0.95)$ and then dividing this by 2 to determine the area in the upper tail. The lower-tail area is computed by subtracting the upper-tail area (i.e., $(1 - 0.95) / 2$) from 1. The

FIGURE 8–1 Z value for 95% Confidence

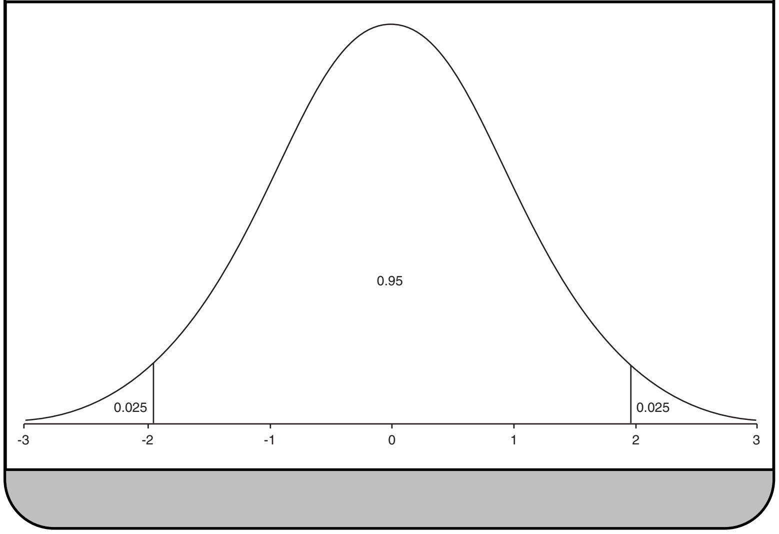
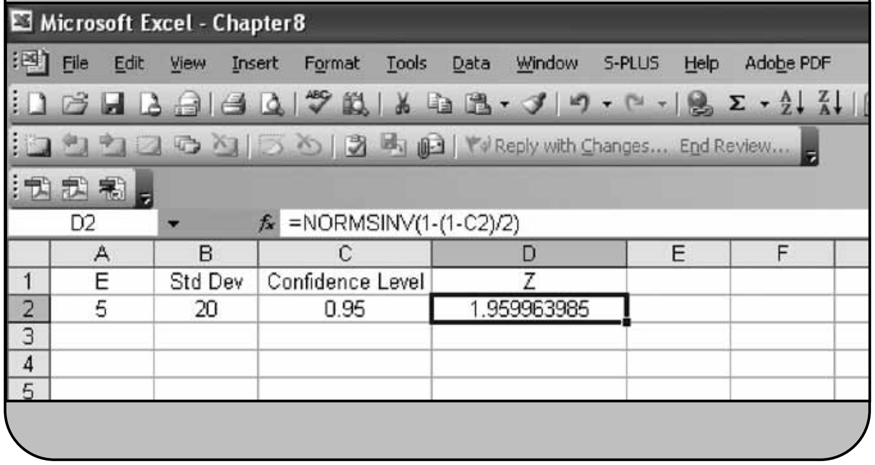


FIGURE 8–2 Data to Estimate Sample Size


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

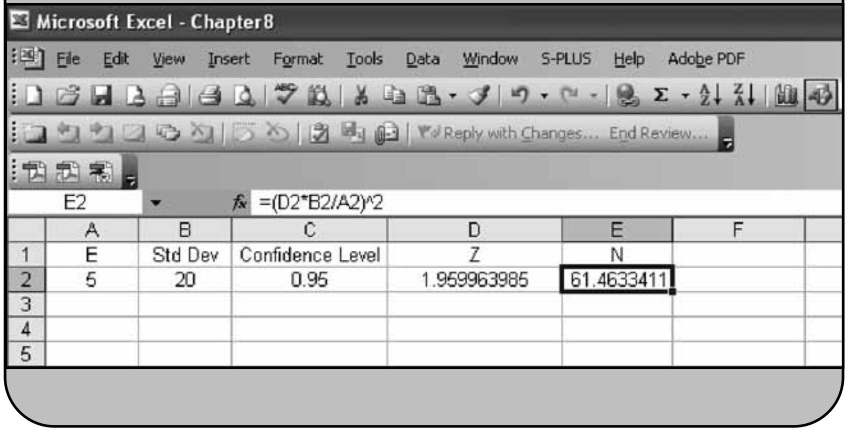
D2 =NORMSINV(1-(1-C2)/2)

	A	B	C	D	E	F
1	E	Std Dev	Confidence Level	Z		
2	5	20	0.95	1.959963985		
3						
4						
5						

argument for the NORMSINV function is "(1−(1−C2)/2)". This returns the value 1.96, which is the *z* value for a 95% confidence interval. The sample size is computed using the preceding formula implemented in Excel as "=(D2*B2/A2)^2". The result is in cell E2 and is shown in Figure 8–3.

Recall that the sample size formula always produces the minimum number of subjects required to ensure that the confidence interval has a margin of error not exceeding *E*. To determine the number of subjects required for the study, we must round up. This is done using Excel's ROUNDUP function. The ROUNDUP function is invoked as "=ROUNDUP (*number to round*, *number of decimal places*)". For sample size computations, we round the value produced by the formula to the nearest integer (i.e., zero decimal places). The sample size required for the study is shown in Figure 8–4. To ensure that the 95% confidence interval estimate of the mean systolic

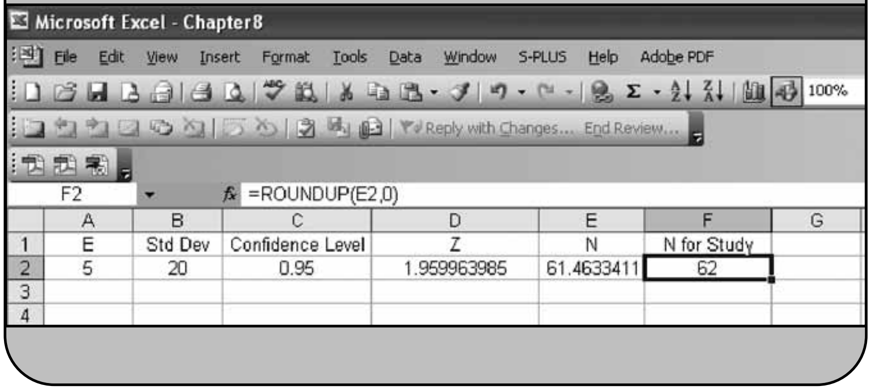
FIGURE 8–3 Determining the Sample Size Required


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E2 =(D2*B2/A2)^2

	A	B	C	D	E	F
1	E	Std Dev	Confidence Level	Z	N	
2	5	20	0.95	1.959963985	61.4633411	
3						
4						
5						

FIGURE 8–4 Sample Size Required for Study


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

F2 =ROUNDUP(E2,0)

	A	B	C	D	E	F	G
1	E	Std Dev	Confidence Level	Z	N	N for Study	
2	5	20	0.95	1.959963985	61.4633411	62	
3							
4							

blood pressure in children between the ages of 3 and 5 years with congenital heart disease is within 5 units of the true mean, a sample of size 62 is needed.

Once the Excel formulas are programmed, other scenarios can be considered. For example, suppose we wish to consider other margins of error (e.g., $E = 5, 4, 3, 2$) and other standard deviations (e.g., 20 and 15). The sample sizes for these other scenarios are determined in Figure 8–5 by copying the formulas from cell D2 through cell F2 to cell D3 through cell F9. The sample sizes are shown in Figure 8–5.

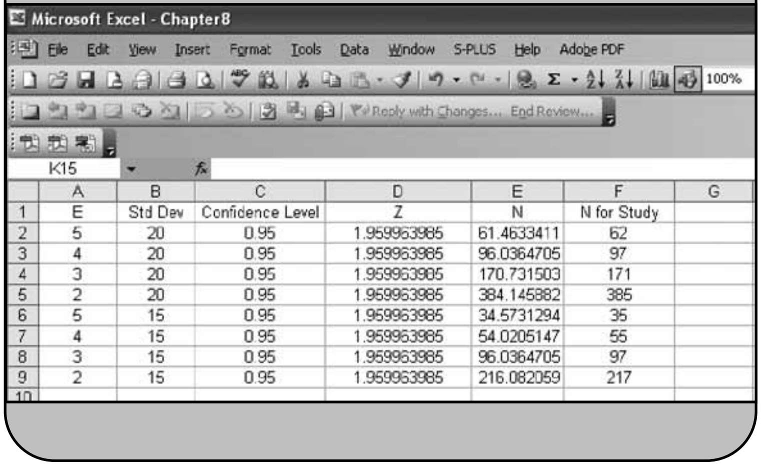
If the standard deviation is 20, then to ensure that a 95% confidence interval estimate of the mean systolic blood pressure in children between the ages of 3 and 5 with congenital heart disease is within 2 units of the true mean, a sample of size 385 is needed. If the standard deviation is 15, a sample of size 217 is needed. It is extremely important to accurately estimate the standard deviation, as it can dramatically affect the sample size.

8.2 SAMPLE SIZE ESTIMATES FOR CONFIDENCE INTERVALS WITH A DICHOTOMOUS OUTCOME IN ONE SAMPLE

In Chapter 8 of the textbook, we presented the following formula to estimate the proportion of successes in a dichotomous outcome variable in a single population:

$$n = p(1-p)\left(\frac{z}{E}\right)^2$$

FIGURE 8–5 Sample Size Estimates for Various Scenarios


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...
K15
fx

	A	B	C	D	E	F	G
1	E	Std Dev	Confidence Level	Z	N	N for Study	
2	5	20	0.95	1.959963985	61.4633411	62	
3	4	20	0.95	1.959963985	96.0364705	97	
4	3	20	0.95	1.959963985	170.731503	171	
5	2	20	0.95	1.959963985	384.145882	385	
6	5	15	0.95	1.959963985	34.5731294	35	
7	4	15	0.95	1.959963985	54.0205147	55	
8	3	15	0.95	1.959963985	96.0364705	97	
9	2	15	0.95	1.959963985	216.082059	217	
10							

where z is the value from the standard normal distribution reflecting the confidence level that will be used (e.g., $z = 1.96$ for 95%), E is the desired margin of error, and p is the proportion of successes in the population. If there is no information available to approximate p, then $p = 0.5$ can be used to generate the most conservative, or largest, sample size.

Example 8.2. In Example 8.3 of the textbook, we determined the sample size required to estimate the proportion of freshmen at a university who currently smoke (i.e., the prevalence of smoking). The investigator wanted to ensure that a 95% confidence interval estimate of the proportion of freshmen who smoke was within 5% of the true proportion. No information was available on the prevalence of smoking, thus $p = 0.5$ was used.

The margin of error ($E = 0.05$), proportion ($p = 0.5$), and confidence level are input into Excel as shown in Figure 8–6. The z value is estimated using the NORMSINV function, as shown in Figure 8–6. Recall that the argument for the NORMSINV function is the area in the lower tail of the standard normal curve (Figure 8–1). The sample size is computed using the formula shown implemented here in Excel as "=B2*(1−B2)*(D2/A2)^2". The result is in cell E2 and is shown in Figure 8–7.

The final step is to round up to the next integer using the ROUNDUP function. The sample size required for the study is shown in Figure 8–8. To ensure that a 95% confidence interval estimate of the proportion of freshmen who smoke is within 5% of the true proportion, a sample of size 385 is needed.

8.3 SAMPLE SIZE ESTIMATES FOR CONFIDENCE INTERVALS WITH A CONTINUOUS OUTCOME IN TWO INDEPENDENT SAMPLES

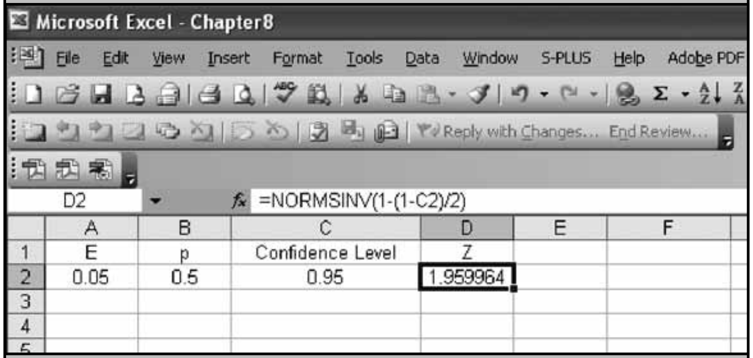
In Chapter 8 of the textbook, we presented the following formula to estimate the sample size required to estimate the difference in means in two independent populations:

$$n_i = 2\left(\frac{z\sigma}{E}\right)^2$$

where n_i is the sample size required in each group ($i = 1,2$), z is the value from the standard normal distribution reflecting the confidence level that will be used (e.g., $z = 1.96$ for 95%), and E is the desired margin of error. σ again reflects the standard deviation of the outcome variable. Recall from Chapter 6 in the textbook, when we generated a confidence interval estimate for the difference in means, we used S_p, the pooled estimate of the common standard deviation, as a measure of variability in the outcome (where S_p is computed as $\text{S}p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$). If data are available on variability of the outcome in each comparison group, then S_p can be computed and used in the sample size formula. However, it is more often the case that data on the variability of the outcome are available from only one group, often the untreated (e.g., placebo/control) or unexposed group. This value can be used to determine the sample sizes.

Example 8.3. In Example 8.6 of the textbook, we determined the sample sizes required to compare two diet programs in obese children. The plan is to enroll children

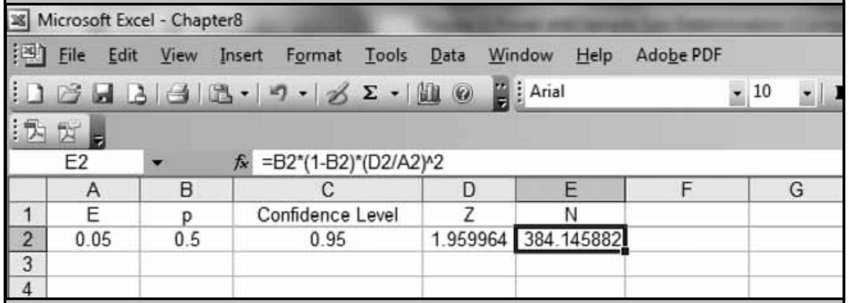
FIGURE 8–6 Data to Estimate Sample Size


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

D2 =NORMSINV(1-(1-C2)/2)

	A	B	C	D	E	F
1	E	p	Confidence Level	Z		
2	0.05	0.5	0.95	1.959964		
3						
4						

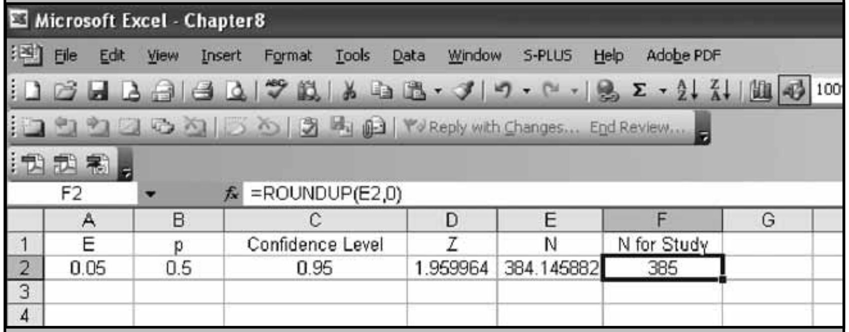
FIGURE 8–7 Determining the Sample Size Required


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10

E2 =B2*(1-B2)*(D2/A2)^2

	A	B	C	D	E	F	G
1	E	p	Confidence Level	Z	N		
2	0.05	0.5	0.95	1.959964	384.145882		
3							
4							

FIGURE 8–8 Sample Size Required for Study


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

F2 =ROUNDUP(E2,0)

	A	B	C	D	E	F	G
1	E	p	Confidence Level	Z	N	N for Study	
2	0.05	0.5	0.95	1.959964	384.145882	385	
3							
4							

and weigh them at the start of the study. Each child will then be randomly assigned to one of the competing diets (low-fat or low-carbohydrate) and followed for 8 weeks, at which time they will again be weighed. The number of pounds lost will be computed for each child. A 95% confidence interval will be estimated to quantify the difference in weight lost between the two diets, and the investigator would like the margin of error to be no more than 3 pounds. Based on adult studies, the common standard deviation was estimated at 8.1 pounds.

The margin of error, standard deviation, and confidence level are input into an Excel worksheet. The *z* value is estimated using the NORMSINV function as shown in Figure 8–9. The sample size required per group is computed using the formula shown here implemented in Excel as "=2*(D2*B2/A2)^2". The result is in cell E2 and is shown in Figure 8–10. The final step is to round up to the next integer using the ROUNDUP function. The sample size required in each group for the study is shown in Figure 8–11.

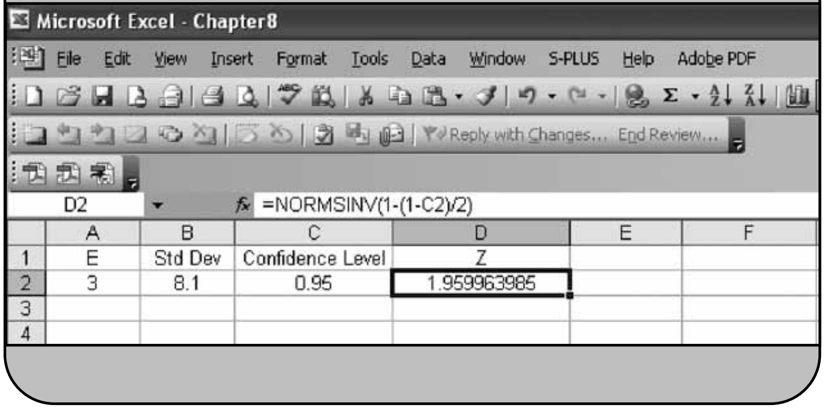
Samples of size $n_1 = 57$ and $n_2 = 57$ will ensure that the 95% confidence interval for the difference in weight lost between diets will have a margin of error of no more than 3 pounds. (Note that in Chapter 8 of the textbook, we estimated the sample size at 56 per group because we carried only 2 decimal places in the by-hand computations. Excel carries more decimal places and therefore rounding up produces sample sizes of 57 per group.)

8.4 SAMPLE SIZE ESTIMATES FOR CONFIDENCE INTERVALS WITH A CONTINUOUS OUTCOME IN MATCHED SAMPLES

In Chapter 8 of the textbook, we presented the following formula to estimate the sample size required to estimate the mean difference of a continuous outcome variable in two matched populations:

$$n = \left(\frac{z\sigma_d}{E}\right)^2$$

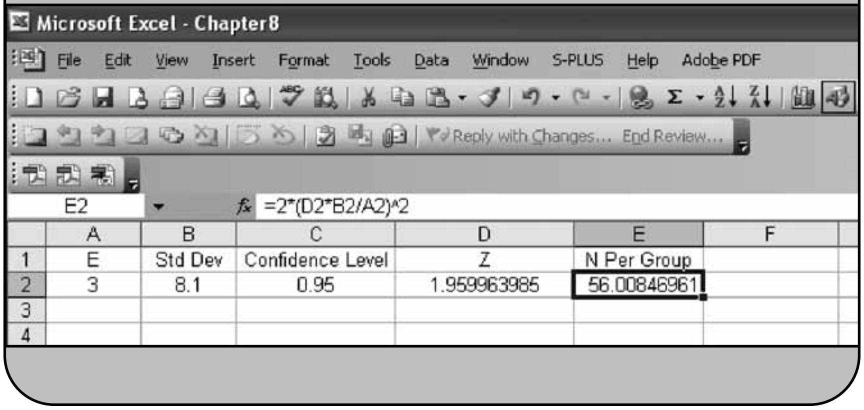
FIGURE 8–9 Data to Estimate Sample Size



D2 =NORMSINV(1-(1-C2)/2)

	A	B	C	D	E	F
1	E	Std Dev	Confidence Level	Z		
2	3	8.1	0.95	1.959963985		
3						
4						

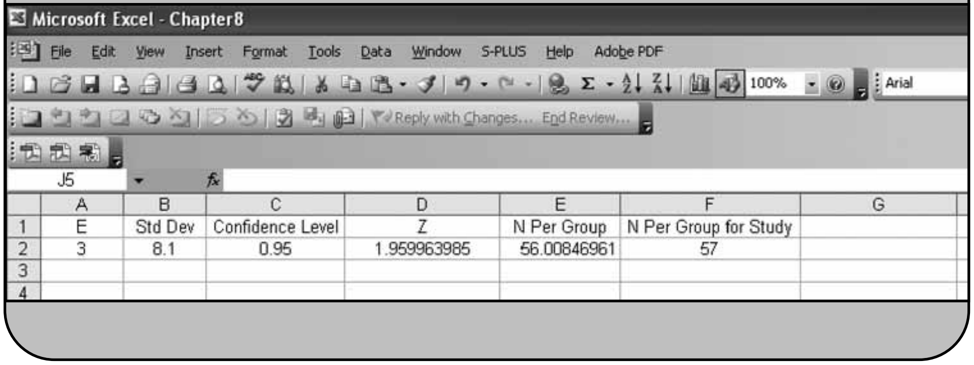
FIGURE 8–10 Determining the Sample Size per Group Required



E2 =2*(D2*B2/A2)^2

	A	B	C	D	E	F
1	E	Std Dev	Confidence Level	Z	N Per Group	
2	3	8.1	0.95	1.959963985	56.00846961	
3						
4						

FIGURE 8–11 Sample Size per Group Required for Study


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...

J5

	A	B	C	D	E	F	G
1	E	Std Dev	Confidence Level	Z	N Per Group	N Per Group for Study	
2	3	8.1	0.95	1.959963985	56.00846961	57	
3							
4							

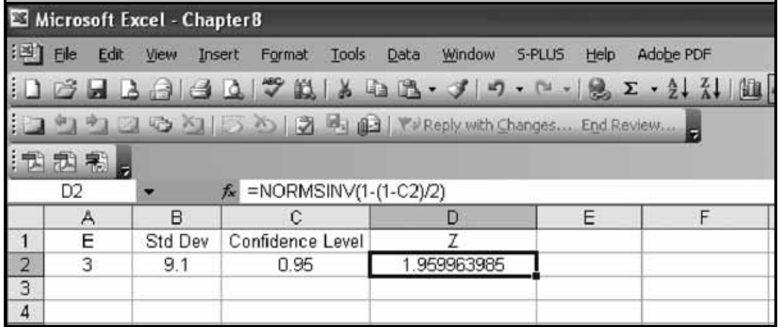
where *z* is the value from the standard normal distribution reflecting the confidence level that will be used (e.g., $z = 1.96$ for 95%), *E* is the desired margin of error, and σ_d is the standard deviation of the difference scores (e.g., the difference based on measurements over time or the difference between matched pairs). It is extremely important that the standard deviation of the difference scores is used here to appropriately estimate the sample size.

Example 8.4. Consider again the diet study proposed in Example 8.3 of the Excel workbook (and in Example 8.7 in the textbook). The investigator considers an alternative design, a crossover trial, where each participant will follow each diet for 8 weeks. At the end of each 8-week period, the weight lost during that period will be measured. The difference in weight lost on the low-fat diet and the low-carbohydrate diet will be computed for each child and a confidence interval for the mean difference in weight lost will be computed. The investigator wants to determine the sample size required to ensure that a 95% confidence interval estimate of the mean difference in weight lost between diets is within 3 units of the true mean difference. Suppose that the standard deviation of the difference in weight loss between a low-fat diet and a low-carbohydrate diet is approximately 9.1 lbs. based on a cross-over trial conducted in adults.

The margin of error, standard deviation of the differences in weights, and the confidence level are input into an Excel worksheet. The *z* value is estimated using the NORMSINV function as shown in Figure 8–12. The sample size required is computed using the formula shown here implemented in Excel as “=(D2*B2/A2)^2”. The result is in cell E2 and is shown in Figure 8–13.

The final step is to round up to the next integer using the ROUNDUP function. The sample size required for the study is shown in cell F2 in Figure 8–13. To ensure that the 95% confidence interval estimate of the mean difference in weight lost between diets is within 3 units of the true mean, a sample of size 36 children is needed.

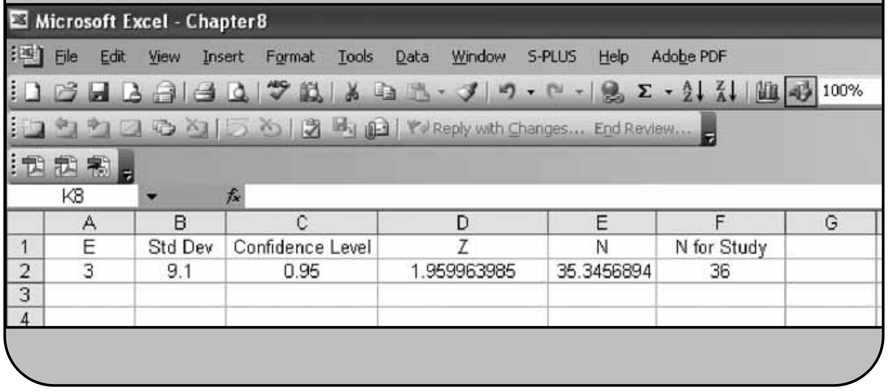
FIGURE 8–12 Data to Estimate Sample Size


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

D2 =NORMSINV(1-(1-C2)/2)

	A	B	C	D	E	F
1	E	Std Dev	Confidence Level	Z		
2	3	9.1	0.95	1.959963985		
3						
4						

FIGURE 8–13 Sample Size Required for Study


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...
K8
fx

	A	B	C	D	E	F	G
1	E	Std Dev	Confidence Level	Z	N	N for Study	
2	3	9.1	0.95	1.959963985	35.3456894	36	
3							
4							

8.5 SAMPLE SIZE ESTIMATES FOR CONFIDENCE INTERVALS WITH A DICHOTOMOUS OUTCOME IN TWO INDEPENDENT SAMPLES

In Chapter 8 of the textbook, we presented the following formula to estimate the difference in proportions between two independent populations (i.e., to estimate the risk difference):

$$n_i = \left[p_1(1-p_1) + p_2(1-p_2)\right]\left(\frac{z}{E}\right)^2$$

where n_i is the sample size required in each group ($i = 1,2$), z is the value from the standard normal distribution reflecting the confidence level that will be used (e.g., $z = 1.96$ for 95%), E is the desired margin of error, and p_1 and p_2 are the proportions of successes in each comparison group. Again, here we are planning a study to generate a 95% confidence interval for the difference in unknown proportions, and the formula to estimate the sample sizes needed requires p_1 and p_2. To estimate the sample size, we need approximate values of p_1 and p_2. The values of p_1 and p_2 that maximize the sample size are $p_1 = p_2 = 0.5$. Thus, if there is no information available to approximate p_1 and p_2, then 0.5 can be used to generate the most conservative, or largest, sample sizes.

Example 8.5. In Example 8.9 in the textbook, an investigator determined the sample size to estimate the impact of smoking on the incidence of prostate cancer. Men who are free of prostate cancer will be enrolled at age 50 and followed for 30 years. The plan is to enroll approximately equal numbers of smokers and nonsmokers in the study and to follow them prospectively for the outcome of interest, a diagnosis of prostate cancer. The plan is to generate a 95% confidence interval for the difference in proportions of smoking and nonsmoking men who develop prostate cancer. How many men should be enrolled in the study to ensure that the 95% confidence interval for the difference in proportions has a margin of error of no more than 5%? Estimates of the incidence of prostate cancer from a previous study were used to design the current study: $p_1 = 0.34$ and $p_2 = 0.17$.

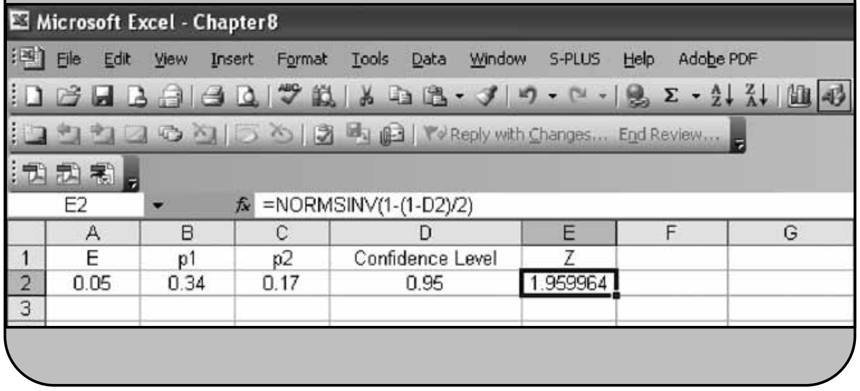
The margin of error, estimates of proportions, and the confidence level are input into an Excel worksheet. The z value is estimated using the NORMSINV function as shown in Figure 8–14. The sample size required per group is computed using the formula shown here implemented in Excel as "=(B2*(1−B2)+C2*(1−C2))*(E2/A2)^2". The result is in cell F2 and is shown in Figure 8–15.

The final step is to round up to the next integer using the ROUNDUP function. The sample size required in each group for the study is shown in Figure 8–16. Samples of size $n_1 = 562$ men who smoke and $n_2 = 562$ men who do not smoke will ensure that the 95% confidence interval for the difference in incidence of prostate cancer will have a margin of error of no more than 5%.

8.6 ISSUES IN ESTIMATING SAMPLE SIZE FOR HYPOTHESIS TESTING

In Chapter 8 of the textbook, we presented formulas to determine the sample size required to ensure a specified power in a test of hypothesis. Excel does not have an analysis tool to perform the computations, but the formulas can be programmed into Excel to determine the appropriate sample sizes. The sample size formulas for hypothesis testing depend on the nature of the outcome variable (e.g., continuous or dichotomous) and also the number of comparison groups involved (e.g., one, two independent, or two matched).

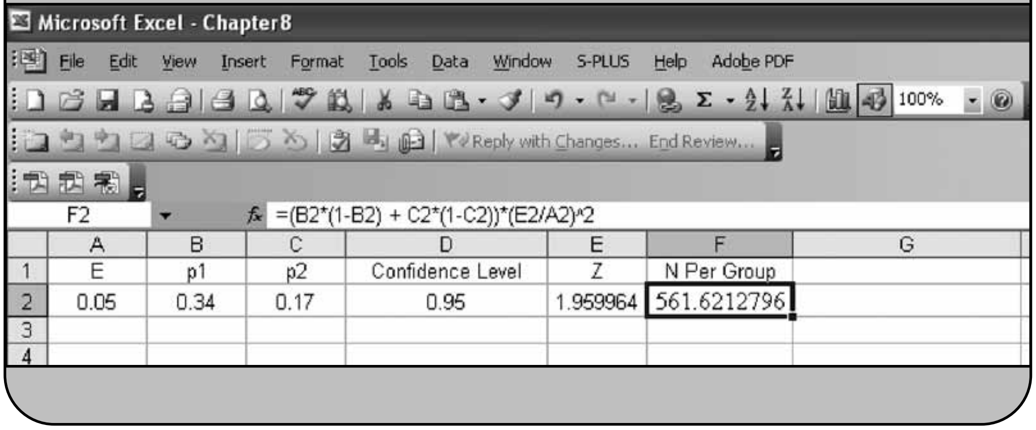
FIGURE 8–14 Data to Estimate Sample Size


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E2 =NORMSINV(1-(1-D2)/2)

	A	B	C	D	E	F	G
1	E	p1	p2	Confidence Level	Z		
2	0.05	0.34	0.17	0.95	1.959964		
3							

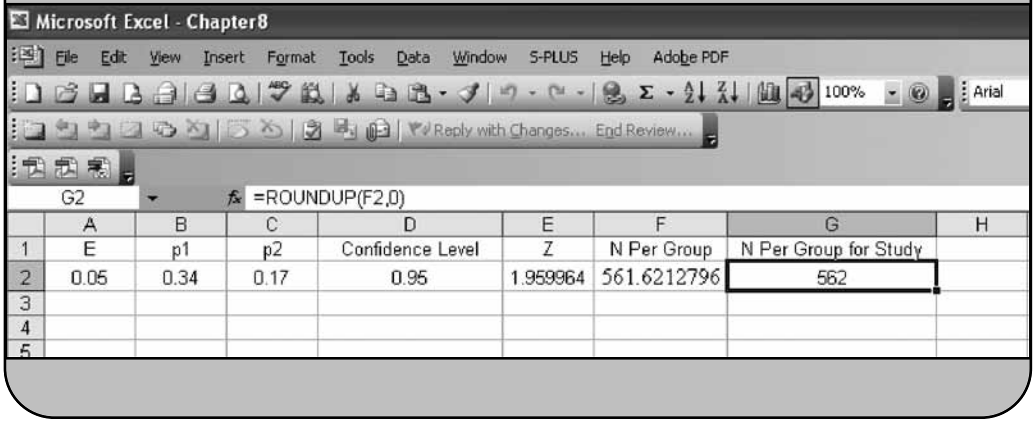
FIGURE 8–15 Determining the Sample Size per Group Required


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

F2 =(B2*(1-B2) + C2*(1-C2))*(E2/A2)^2

	A	B	C	D	E	F	G
1	E	p1	p2	Confidence Level	Z	N Per Group	
2	0.05	0.34	0.17	0.95	1.959964	561.6212796	
3							
4							

FIGURE 8–16 Sample Size per Group Required for Study


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...

G2 =ROUNDUP(F2,0)

	A	B	C	D	E	F	G	H
1	E	p1	p2	Confidence Level	Z	N Per Group	N Per Group for Study	
2	0.05	0.34	0.17	0.95	1.959964	561.6212796	562	
3								
4								
5								

All of the sample size formulas contain the following two terms: $z_{1-\alpha/2}$ and $z_{1-\beta}$ where α is the probability of a Type I error or the specified level of significance (e.g., 0.05), β is the probability of a Type II error, and $1 - \beta$ is the specified power (e.g., 0.80, 0.90). $z_{1-\alpha/2}$ is the value from the standard normal distribution holding $1- \alpha/2$ below it and $z_{1-\beta}$ is the value from the standard normal distribution holding $1-\beta$ below it.

The NORMSINV function is used to compute these values. The NORMSINV function returns the value from the standard normal distribution, z, which holds a specified area below it (i.e., in the lower tail): "=NORMSINV(*lower-tail area*)". For example, if $\alpha = 0.05$, then $z_{1-\alpha/2} = z_{0.975}$ is computed by "=NORMSINV(0.975)". If power = 0.80, then $z_{0.80}$ is computed by "=NORMSINV(0.80)".

8.7 SAMPLE SIZE ESTIMATES FOR TESTS OF MEANS IN ONE SAMPLE

In Chapter 8 of the textbook, we presented a formula to determine the sample size required to ensure adequate power to test the following hypotheses about the mean of a continuous outcome variable in a single population:

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

where μ_0 is the known mean (e.g., an historical control). The formula for determining sample size to ensure that the test has a specified power is given below:

$$n = \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\text{ES}}\right)^2$$

where α is the selected level of significance and $z_{1-\alpha/2}$ is the value from the standard normal distribution holding $1 - \alpha/2$ below it. $1 - \beta$ is the selected power and $z_{1-\beta}$ is the value from the standard normal distribution holding $1 - \beta$ below it. ES is the effect size, defined as follows:

$$\text{ES} = \frac{|\mu_1 - \mu_0|}{\sigma}$$

where μ_1 is the mean under the alternative hypothesis, μ_0 is the mean under the null hypothesis, and σ is the standard deviation of the outcome of interest.

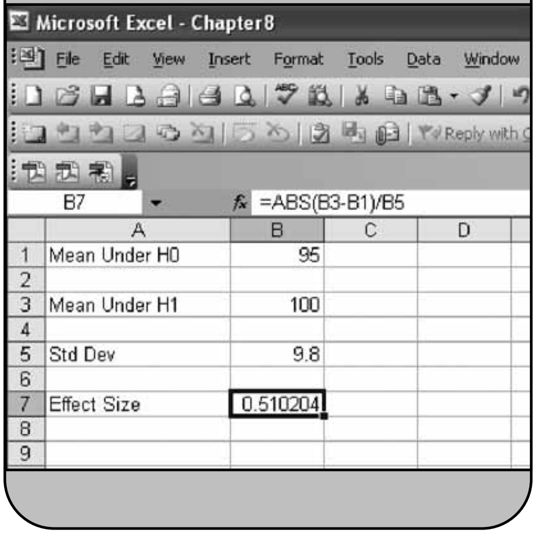
Example 8.6. In Example 8.10 of the textbook, we determined the sample size required to test whether the mean blood glucose level in people who drink at least two cups of coffee per day is different from the reported mean of 95 mg/dl (the standard deviation was 9.8 mg/dl). Investigators wanted a sample size that would ensure 80% power to detect a mean of 100 mg/dl. A two-sided test is planned with a 5% level of significance.

Before we compute the sample size, we first must compute the effect size. This is done by entering the mean under the null hypothesis, the mean under the alternative hypothesis, and the standard deviation into an Excel worksheet, as shown in Figure 8–17.

The effect size is shown in cell B7 and is computed as "=ABS(B3−B1)/B5", where ABS is the Excel function to compute the absolute value of the difference in means under the null and alternative hypotheses. The next step is to compute the z value for the selected level of significance (i.e., $z_{1-\alpha/2}$) and the z value for the desired power (i.e., $z_{1-\beta}$). We first enter the level of significance, α, and the desired power. This is shown in Figure 8–18.

Recall that the argument for the NORMSINV function is the area in the lower tail of the standard normal curve (Figure 8–1). If a two-sided test is planned (which is generally the case for sample size planning) with a 5% level of significance, the area in the lower tail is defined as $(1-\alpha/2)$. Thus, we specify 1−B9/2 as the argument to the NORMSINV function as shown in Figure 8–18. $z_{1-\beta}$ is determined in the same way using "=NORMSINV(B11)". The computations are shown in Figure 8–19. The next step is to compute the sample size based on the effect size and the appropriate z values for the selected α and power. This is shown in Figure 8–20.

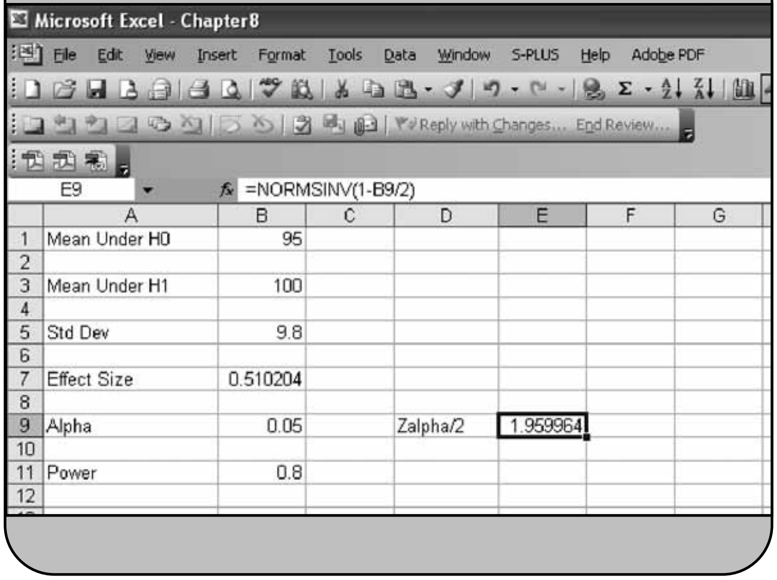
FIGURE 8–17 Data to Estimate Sample Size



B7 =ABS(B3-B1)/B5

	A	B	C	D
1	Mean Under H0	95		
2				
3	Mean Under H1	100		
4				
5	Std Dev	9.8		
6				
7	Effect Size	0.510204		
8				
9				

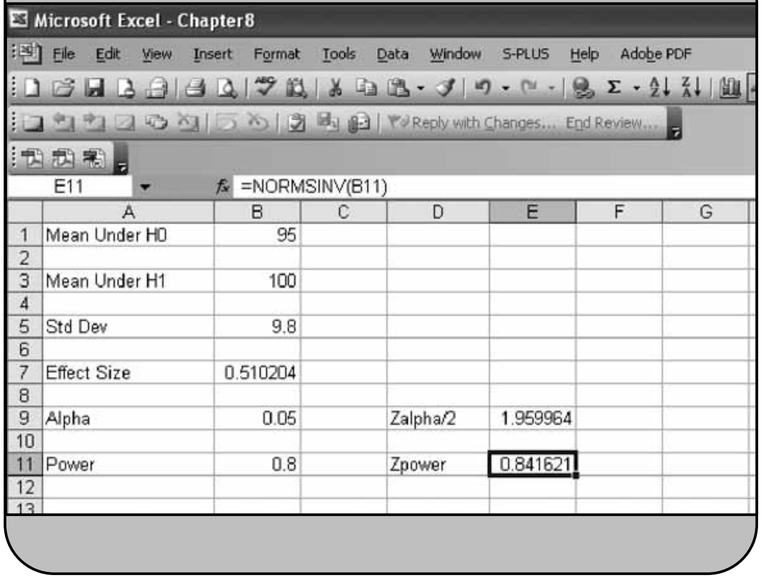
FIGURE 8–18 Level of Significance and Power


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E9 =NORMSINV(1-B9/2)

	A	B	C	D	E	F	G
1	Mean Under H0	95					
2							
3	Mean Under H1	100					
4							
5	Std Dev	9.8					
6							
7	Effect Size	0.510204					
8							
9	Alpha	0.05		Zalpha/2	1.959964		
10							
11	Power	0.8					
12							

FIGURE 8–19 Computing *z* Values

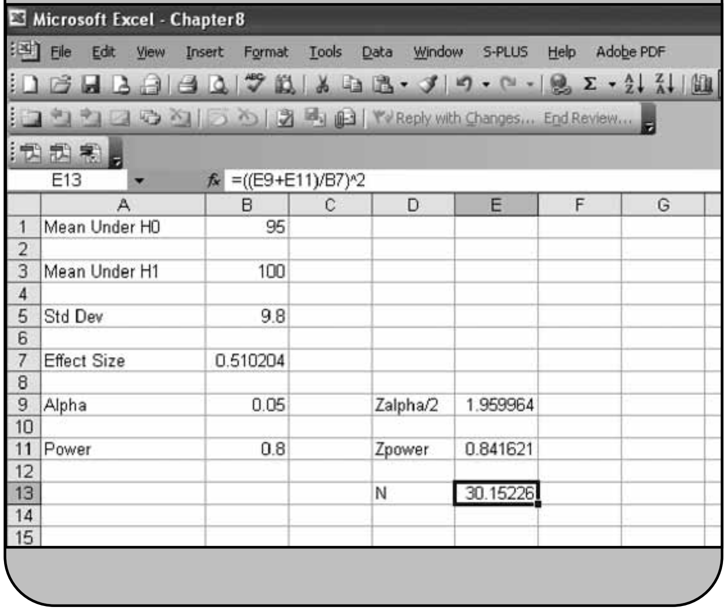

Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E11 =NORMSINV(B11)

	A	B	C	D	E	F	G
1	Mean Under H0	95					
2							
3	Mean Under H1	100					
4							
5	Std Dev	9.8					
6							
7	Effect Size	0.510204					
8							
9	Alpha	0.05		Zalpha/2	1.959964		
10							
11	Power	0.8		Zpower	0.841621		
12							
13							

Because the sample size formula always produces the minimum number of subjects required to ensure that the test has the specified power to detect the desired effect size at the specified level of significance, to determine the number of subjects required for the study, we must round up. This is done using Excel's ROUNDUP function. The sample size required for the study is shown in Figure 8–21. A sample of size $n = 31$ will ensure that a two-sided test with $\alpha = 0.05$ has 80% power to detect a 5-mg/dl difference in mean fasting blood glucose levels.

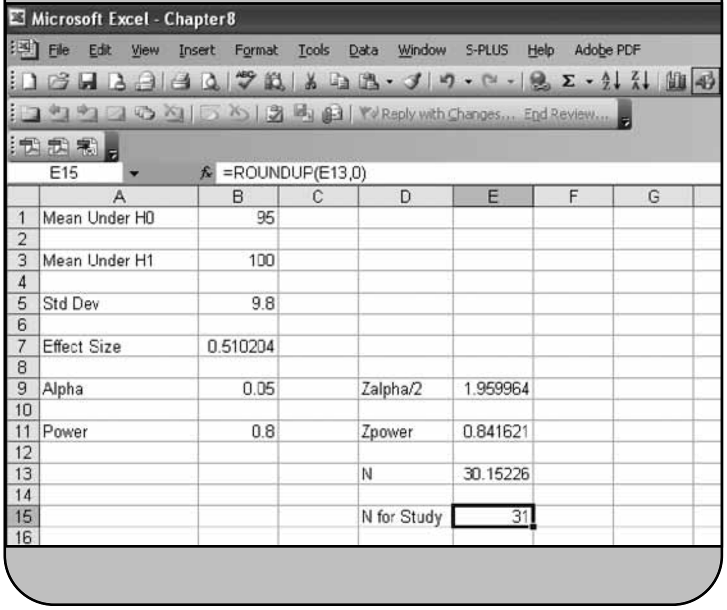
FIGURE 8–20 Determining the Sample Size Required


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E13 =((E9+E11)/B7)^2

	A	B	C	D	E	F	G
1	Mean Under H0	95					
2							
3	Mean Under H1	100					
4							
5	Std Dev	9.8					
6							
7	Effect Size	0.510204					
8							
9	Alpha	0.05		Zalpha/2	1.959964		
10							
11	Power	0.8		Zpower	0.841621		
12							
13				N	30.15226		
14							
15							

FIGURE 8–21 Sample Size Required for Study


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E15 =ROUNDUP(E13,0)

	A	B	C	D	E	F	G
1	Mean Under H0	95					
2							
3	Mean Under H1	100					
4							
5	Std Dev	9.8					
6							
7	Effect Size	0.510204					
8							
9	Alpha	0.05		Zalpha/2	1.959964		
10							
11	Power	0.8		Zpower	0.841621		
12							
13				N	30.15226		
14							
15				N for Study	31		
16							

8.8 SAMPLE SIZE ESTIMATES FOR TESTS OF PROPORTIONS IN ONE SAMPLE

In Chapter 8 of the textbook, we presented a formula to determine the sample size required to ensure adequate power to test the following hypotheses about the proportion of successes in a dichotomous outcome variable in a single population:

$$H_0: p = p_0$$

$$H_1: p \neq p_0$$

where p_0 is the known proportion (e.g., an historical control). The formula for determining sample size to ensure that the test has a specified power is given below:

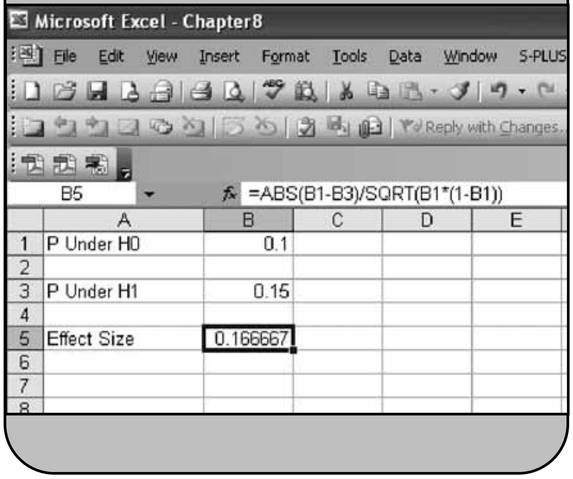
$$n = \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\text{ES}}\right)^2$$

where α is the selected level of significance and $z_{1-\alpha/2}$ is the value from the standard normal distribution holding $1 - \alpha/2$ below it. $1 - \beta$ is the selected power and $z_{1-\beta}$ is the value from the standard normal distribution holding $1-\beta$ below it. ES is the effect size, defined as:

$$\text{ES} = \frac{|p_1 - p_0|}{\sqrt{p_0(1-\text{p}_0)}}$$

where p_0 is the proportion of successes under H_0 and p_1 is the proportion of successes under H_1. The numerator of the effect size, the absolute value of the difference in proportions $|p_1 - p_0|$, again represents what is considered a clinically meaningful or practically important difference in proportions.

FIGURE 8–22 Data to Estimate Sample Size


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS
Reply with Changes...

B5 =ABS(B1-B3)/SQRT(B1*(1-B1))

	A	B	C	D	E
1	P Under H0	0.1			
2					
3	P Under H1	0.15			
4					
5	Effect Size	0.166667			
6					
7					
8					

Example 8.7. In Example 8.13 in the textbook, we determined the sample size required to test whether the percentage of defective stents produced by a manufacturer in one of his plants was more than 10%. The manufacturer wanted the test to have 90% power to detect an absolute difference in proportions of 0.05 (i.e., from 0.10 to 0.15 defectives). How many stents must be evaluated? A two-sided test will be used with a 5% level of significance.

Before we compute the sample size, we first must compute the effect size. This is done by entering the proportion under the null hypothesis and the proportion under the alternative hypothesis into an Excel worksheet, as shown in Figure 8–22.

The effect size is shown in cell B5 and is computed as "=ABS(B3−B1)/SQRT(B1*(1−B1))", where ABS is the Excel function to compute the absolute value of the difference in proportions under the null and alternative hypotheses. The next step is to compute the z value for the selected level of significance (i.e., $z_{1-\alpha/2}$) and the z value for the desired power (i.e., $z_{1-\beta}$). We first enter the level of significance, α, and the desired power. We then use the NORMSINV function twice to compute $z_{1-\alpha/2}$ and $z_{1-\beta}$. This is shown in Figure 8–23.

The next step is to compute the sample size based on the effect size and the appropriate z values for the selected α and power. This is shown in Figure 8–24. As the final step, we round up to the next integer using the ROUNDUP function. The sample size for the study is shown in Figure 8–25.

A sample of size $n = 379$ stents will ensure that a two-sided test with $\alpha = 0.05$ has 90% power to detect a 5% difference in the proportion of defective stents produced. (When we computed the sample size by hand in the textbook, we determined that $n = 364$ stents were needed. The difference is because Excel is carrying more decimal places in the computations.)

8.9 SAMPLE SIZE ESTIMATES FOR TESTS OF DIFFERENCES IN MEANS IN TWO INDEPENDENT SAMPLES

In Chapter 8 of the textbook, we presented a formula to determine the sample size required to ensure adequate power to test the following hypotheses about the difference in means in two independent populations:

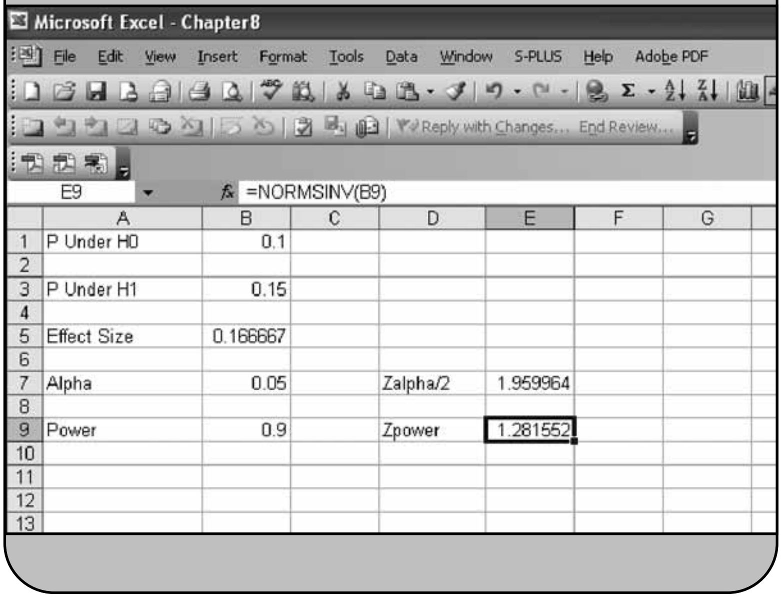
$$H_0: \mu_1 = \mu_2$$

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

where μ_1 and μ_2 are the means in the two comparison populations. The formula for determining sample size required in each group to ensure that the test has a specified power follows:

$$n_i = 2\left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\text{ES}}\right)^2$$

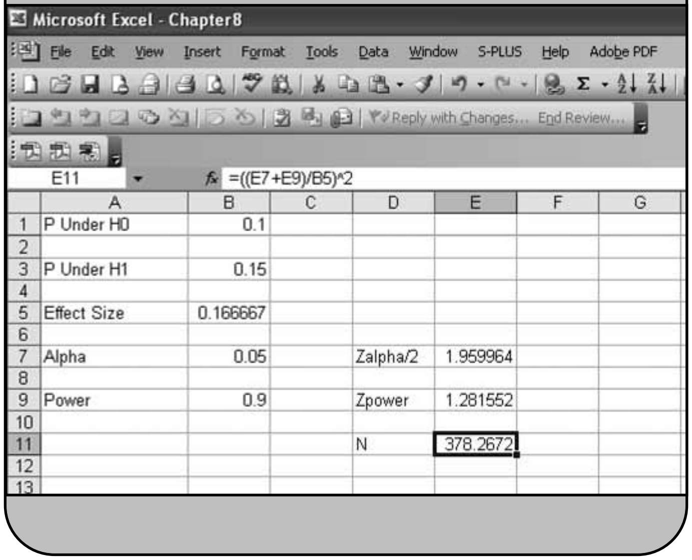
FIGURE 8–23 Computing *z* Values


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E9 =NORMSINV(B9)

	A	B	C	D	E	F	G
1	P Under H0	0.1					
2							
3	P Under H1	0.15					
4							
5	Effect Size	0.166667					
6							
7	Alpha	0.05		Zalpha/2	1.959964		
8							
9	Power	0.9		Zpower	1.281552		
10							
11							
12							
13							

FIGURE 8–24 Determining the Sample Size Required


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E11 =((E7+E9)/B5)^2

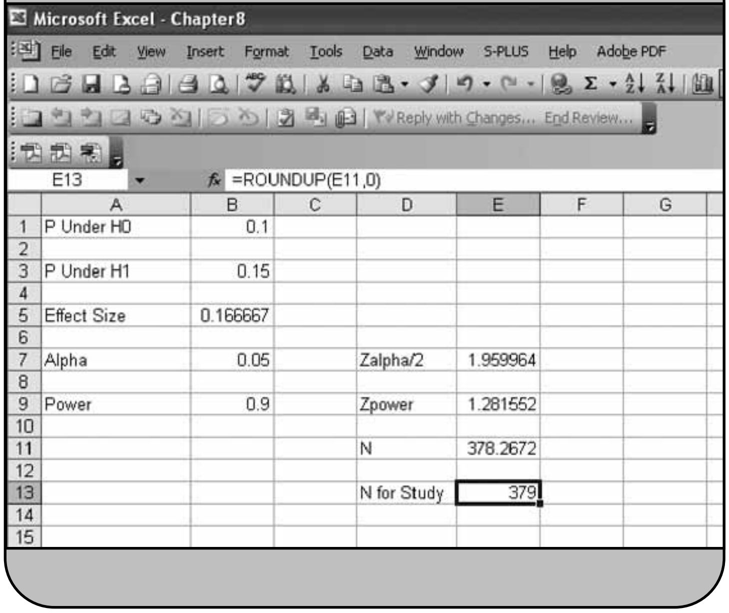
	A	B	C	D	E	F	G
1	P Under H0	0.1					
2							
3	P Under H1	0.15					
4							
5	Effect Size	0.166667					
6							
7	Alpha	0.05		Zalpha/2	1.959964		
8							
9	Power	0.9		Zpower	1.281552		
10							
11				N	378.2672		
12							
13							

where n_i is the sample size required in each group ($i = 1,2$), α is the selected level of significance, $z_{1-\alpha/2}$ is the value from the standard normal distribution holding $1 - \alpha/2$ below it, $1 - \beta$ is the selected power, and $z_{1-\beta}$ is the value from the standard normal distribution holding $1 - \beta$ below it. ES is the effect size, defined as follows:

$$\text{ES} = \frac{|\mu_1 - \mu_2|}{\sigma}$$

where $|\mu_1 - \mu_2|$ is the absolute value of the difference in means between the two groups representing what is considered a clinically meaningful or practically important difference in means. σ is the standard deviation of the outcome of interest. If data

FIGURE 8–25 Sample Size Required for Study


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E13 =ROUNDUP(E11,0)

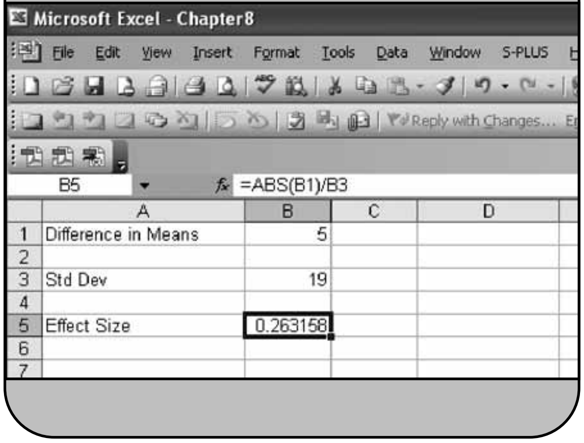
	A	B	C	D	E	F	G
1	P Under H0	0.1					
2							
3	P Under H1	0.15					
4							
5	Effect Size	0.166667					
6							
7	Alpha	0.05		Zalpha/2	1.959964		
8							
9	Power	0.9		Zpower	1.281552		
10							
11				N	378.2672		
12							
13				N for Study	379		
14							
15							

are available on variability of the outcome in each comparison group, then S_p (the pooled estimate of the common standard deviation) can be computed and used to generate the sample sizes. However, it is more often the case that data on the variability of the outcome are available from only one group, usually the untreated (e.g., placebo/control) or unexposed group.

Example 8.8. In Example 8.14 in the textbook, we determined the sample sizes required for a clinical trial to evaluate the efficacy of a new drug designed to reduce systolic blood pressure. The plan was to enroll participants and to randomly assign them to receive either the new drug or a placebo and to measure systolic blood pressure in each participant after 12 weeks on the assigned treatment. Investigators indicated that a 5-unit difference in mean systolic blood pressure would represent a clinically meaningful difference. How many patients should be enrolled in the trial to ensure that the power of the test is 80% to detect this difference? A two-sided test is planned with a 5% level of significance and the standard deviation is assumed to be 19.0, based on data from the Framingham Heart Study.

We first compute the effect size based on the hypothesized difference in means under the alternative hypothesis and the standard deviation. The data are entered into an Excel worksheet as shown in Figure 8–26. The effect size is shown in cell B5 and is computed as "=ABS(B1)/B3". Notice that the hypothesized difference in means is specified in Figure 8–26. In some applications, the means under the null and alternative hypotheses are specified, in which case the difference is computed and used as the numerator in the computation of the effect size. We next enter the level of significance, α, and the desired power to compute $z_{1-\alpha/2}$ and the $z_{1-\beta}$. This is shown in Figure 8–27. The next step is to compute the sample size per group based on the effect size and the appropriate z values for the selected α and power. This is shown in Figure 8–28.

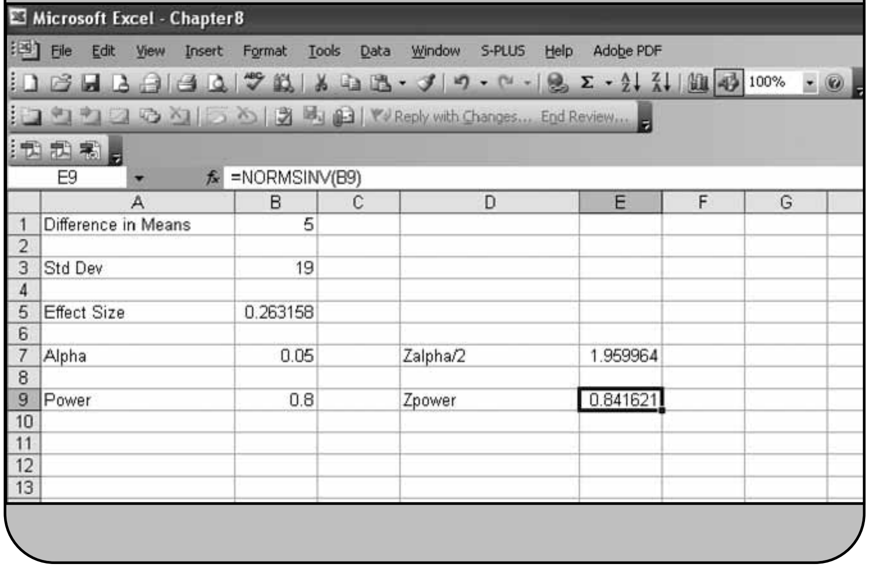
FIGURE 8–26 Data to Estimate Sample Size


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS
Reply with Changes...

B5 =ABS(B1)/B3

	A	B	C	D
1	Difference in Means	5		
2				
3	Std Dev	19		
4				
5	Effect Size	0.263158		
6				
7				

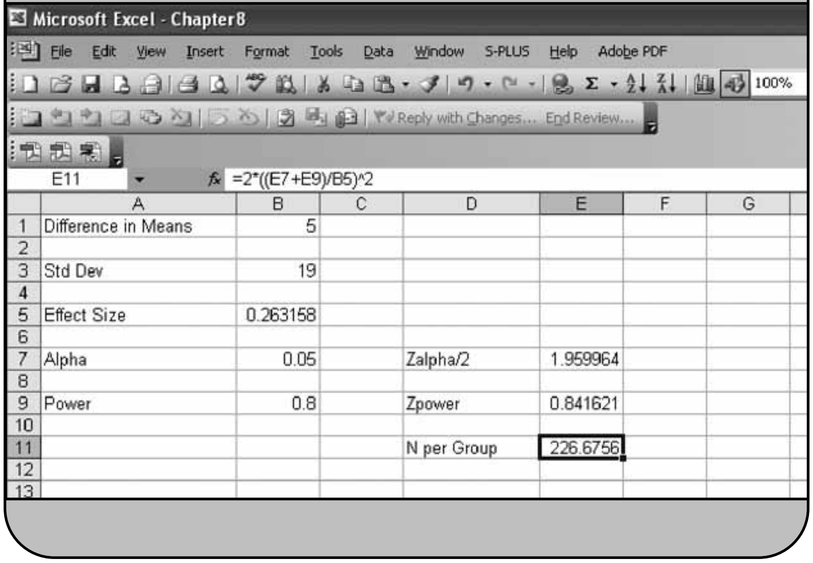
FIGURE 8–27 Computing *z* Values


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

E9 =NORMSINV(B9)

	A	B	C	D	E	F	G
1	Difference in Means	5					
2							
3	Std Dev	19					
4							
5	Effect Size	0.263158					
6							
7	Alpha	0.05		Zalpha/2	1.959964		
8							
9	Power	0.8		Zpower	0.841621		
10							
11							
12							
13							

FIGURE 8–28 Determining the Sample Size Required per Group


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

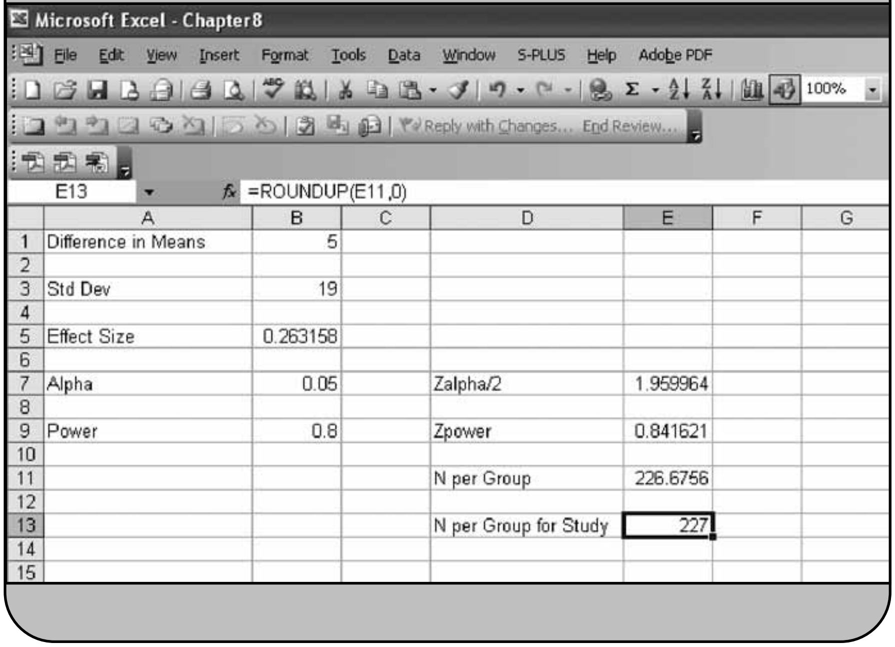
E11 =2*((E7+E9)/B5)^2

	A	B	C	D	E	F	G
1	Difference in Means	5					
2							
3	Std Dev	19					
4							
5	Effect Size	0.263158					
6							
7	Alpha	0.05		Zalpha/2	1.959964		
8							
9	Power	0.8		Zpower	0.841621		
10							
11				N per Group	226.6756		
12							
13							

Finally, because the sample size formula always produces the minimum number of subjects per group required to ensure that the test has the specified power to detect the desired effect size at the specified level of significance, to determine the numbers of subjects per group required for the study, we must round up. This is done using Excel's ROUNDUP function. The sample sizes required per group are shown in Figure 8–29.

Samples of size $n_1 = 227$ and $n_2 = 227$ will ensure that the test of hypothesis will have 80% power to detect a 5-unit difference in mean systolic blood pressures in patients receiving the new drug as compared to patients receiving the placebo.

FIGURE 8–29 Sample Size Required per Group for Study


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

E13 =ROUNDUP(E11,0)

	A	B	C	D	E	F	G
1	Difference in Means	5					
2							
3	Std Dev	19					
4							
5	Effect Size	0.263158					
6							
7	Alpha	0.05		Zalpha/2	1.959964		
8							
9	Power	0.8		Zpower	0.841621		
10							
11				N per Group	226.6756		
12							
13				N per Group for Study	227		
14							
15							

8.10 SAMPLE SIZE ESTIMATES FOR TESTS OF MEAN DIFFERENCES IN MATCHED SAMPLES

In Chapter 8 of the textbook, we presented a formula to determine the sample size required to ensure adequate power to test the following hypotheses about the mean difference in a continuous outcome based on matched populations:

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

where μ_d is the mean difference in the population. The formula for determining the sample size (i.e., number of participants, each of whom will be measured twice) required to ensure that the test has a specified power is as follows:

$$n = \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\text{ES}}\right)^2$$

where α is the selected level of significance, $z_{1-\alpha/2}$ is the value from the standard normal distribution holding $1 - \alpha/2$ below it, $1 - \beta$ is the selected power, and $z_{1-\beta}$ is the value from the standard normal distribution holding $1 - \beta$ below it. ES is the effect size, defined as follows:

$$\text{ES} = \frac{\mu_d}{\sigma_d}$$

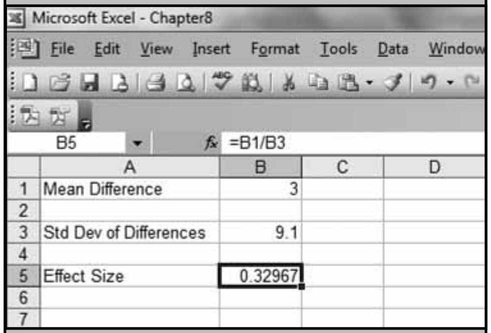
where μ_d is the mean difference expected under the alternative hypothesis, H_1, and σ_d is the standard deviation of the difference in the outcome (e.g., the difference based on measurements over time or the difference between matched pairs).

Example 8.9. In Example 8.16 of the textbook, we generated sample size requirements for a crossover trial to compare two diet programs for their effectiveness in promoting weight loss. The proposed study will have each child follow each diet for 8 weeks, and at the end of each 8-week period, the weight lost during that period will be measured. The difference in weight lost between the diets will be computed for each child and the plan is to test if there is a statistically significant difference in weight loss between the diets. How many children are required to ensure that a two-sided test with a 5% level of significance has 80% power to detect a mean difference of 3 pounds in weight lost between the two diets? Based on a previous study, the standard deviation in the differences in weight loss is estimated at 9.1 pounds.

We first compute the effect size based on the hypothesized mean difference between weight-loss programs and the standard deviation of the differences in weight loss. The data are entered into an Excel worksheet as shown in Figure 8–30.

The effect size is shown in cell B5 and is computed as "=B1/B3". We next enter the level of significance, α, and the

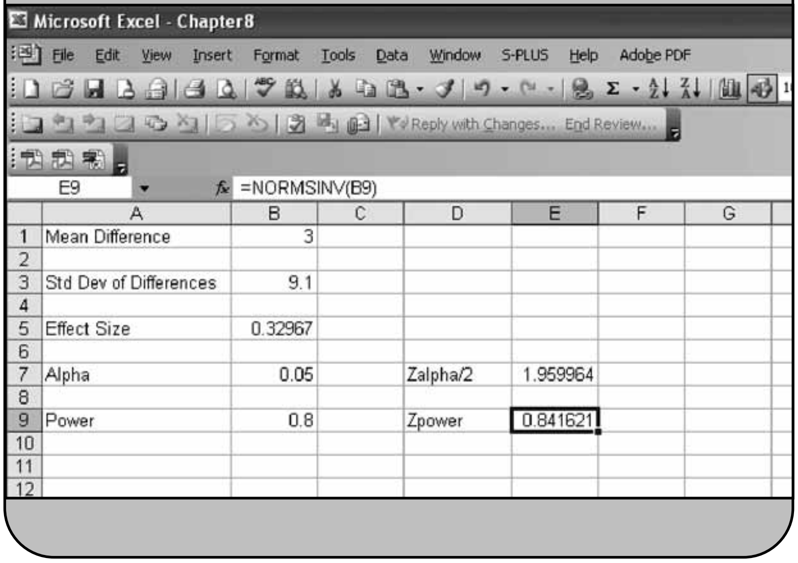
FIGURE 8–30 Data to Estimate Sample Size


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window

B5 =B1/B3

	A	B	C	D
1	Mean Difference	3		
2				
3	Std Dev of Differences	9.1		
4				
5	Effect Size	0.32967		
6				
7				

FIGURE 8–31 Computing *z* Values


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E9 =NORMSINV(B9)

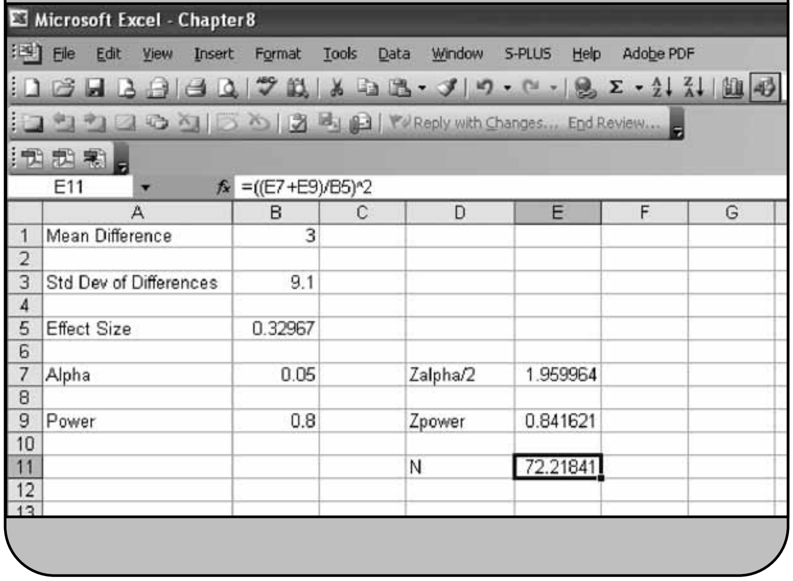
	A	B	C	D	E	F	G
1	Mean Difference	3					
2							
3	Std Dev of Differences	9.1					
4							
5	Effect Size	0.32967					
6							
7	Alpha	0.05		Zalpha/2	1.959964		
8							
9	Power	0.8		Zpower	0.841621		
10							
11							
12							

desired power to compute $z_{1-\alpha/2}$ and $z_{1-\beta}$. This is shown in Figure 8–31. The next step is to compute the sample size based on the effect size and the appropriate *z* values for the selected α and power. This is shown in Figure 8–32.

Finally, because the sample size formula always produces the minimum number of subjects required to ensure that the test has the specified power to detect the desired effect size at the specified level of significance, to determine the numbers of subjects required for the study, we must round up. This is done using Excel's ROUNDUP function. The sample sizes required per group are shown in Figure 8–33.

A sample of size $n = 73$ children will ensure that a two-sided test with $\alpha = 0.05$ has 80% power to detect a mean dif-

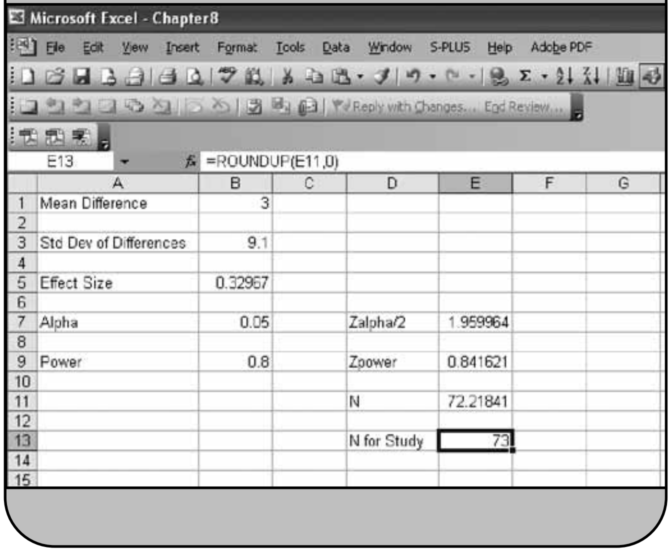
FIGURE 8–32 Determining the Sample Size Required


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E11 =((E7+E9)/B5)^2

	A	B	C	D	E	F	G
1	Mean Difference	3					
2							
3	Std Dev of Differences	9.1					
4							
5	Effect Size	0.32967					
6							
7	Alpha	0.05		Zalpha/2	1.959964		
8							
9	Power	0.8		Zpower	0.841621		
10							
11				N	72.21841		
12							

FIGURE 8–33 Sample Size Required for Study


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E13 =ROUNDUP(E11,0)

	A	B	C	D	E	F	G
1	Mean Difference	3					
2							
3	Std Dev of Differences	9.1					
4							
5	Effect Size	0.32967					
6							
7	Alpha	0.05		Zalpha/2	1.959964		
8							
9	Power	0.8		Zpower	0.841621		
10							
11				N	72.21841		
12							
13				N for Study	73		
14							
15							

ference of 3 pounds between diets using a crossover trial (i.e., each child will be measured on each diet).

8.11 SAMPLE SIZE ESTIMATES FOR TESTS OF PROPORTIONS IN TWO INDEPENDENT SAMPLES

In Chapter 8 of the textbook, we presented a formula to determine the sample size required to ensure adequate power to test the following hypotheses about the difference in proportions in two independent populations:

$$H_0: p_1 = p_2$$
$$H_1: p_1 \neq p_2$$

where p_1 and p_2 are the proportions in the two comparison populations. The formula for determining sample size required in each group to ensure that the test has a specified power is as follows:

$$n_i = 2\left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\text{ES}}\right)^2$$

where n_i is the sample size required in each group ($i = 1,2$), α is the selected level of significance, $z_{1-\alpha/2}$ is the value from the standard normal distribution holding $1 - \alpha/2$ below it, $1 - \beta$ is the selected power, and $z_{1-\beta}$ is the value from the standard normal distribution holding $1 - \beta$ below it. ES is the effect size, defined as follows:

$$\text{ES} = \frac{|p_1 - p_2|}{\sqrt{p(1-p)}}$$

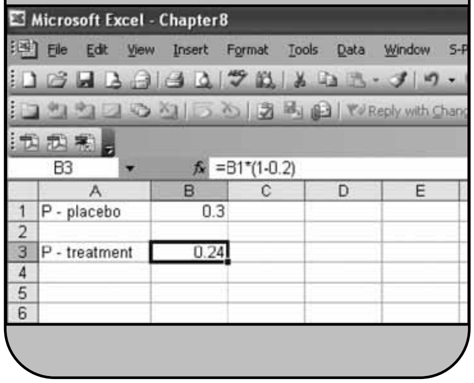
where $|p_1 - p_2|$ is the absolute value of the difference in proportions between the two groups expected under the alternative hypothesis, H_1, and p is the overall proportion, based on pooling the data from the two comparison groups. (p can be computed by taking the mean of the proportions in the two comparison groups, assuming that the groups will be of approximately equal size.)

Example 8.10. In Example 8.18 of the textbook, we determined the sample size needed for a clinical trial proposed to evaluate the efficacy of a new drug designed to reduce systolic blood pressure. The primary outcome is diagnosis of hypertension (true/false), defined as a systolic blood pressure above 140 or a diastolic blood pressure above 90. In planning the trial, investigators hypothesized that 30% of the participants would meet the criteria for hypertension in the placebo group and that the new drug would be considered efficacious if there was a 20% reduction in the proportion of patients receiving the new drug who meet the criteria for hypertension (i.e., if the proportion is 24% among patients receiving the new drug). How many patients should be enrolled in the trial to ensure that the power of the test is 80% to detect this difference in the proportions of patients with hypertension? A two-sided test will be used with a 5% level of significance.

We first compute the effect size based on the hypothesized difference in proportions. The proportion expected in the placebo group is entered into cell B1 and the proportion expected in the treatment group is computed as a 20% reduction in B3 using "=B1*(1−0.2)". The data in the Excel worksheet are shown in Figure 8–34.

Before computing the effect size, we need to compute the overall proportion. This is done by taking the mean of the proportions in the two treatment groups using "=(B1+B3)/2". The computation is shown in Figure 8–35. In Figure 8–36, we compute the effect size.

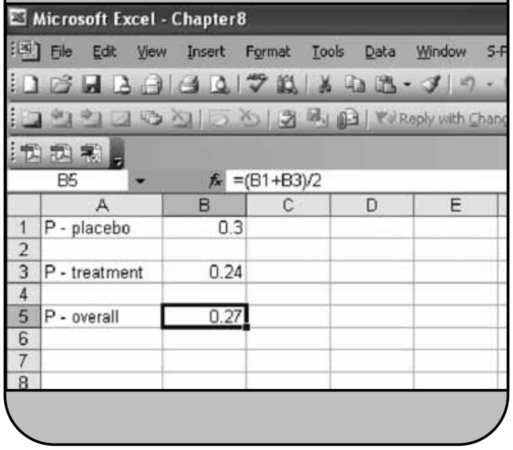
FIGURE 8–34 Sample Proportions


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-P
Reply with Chang

B3 =B1*(1-0.2)

	A	B	C	D	E
1	P - placebo	0.3			
2					
3	P - treatment	0.24			
4					
5					
6					

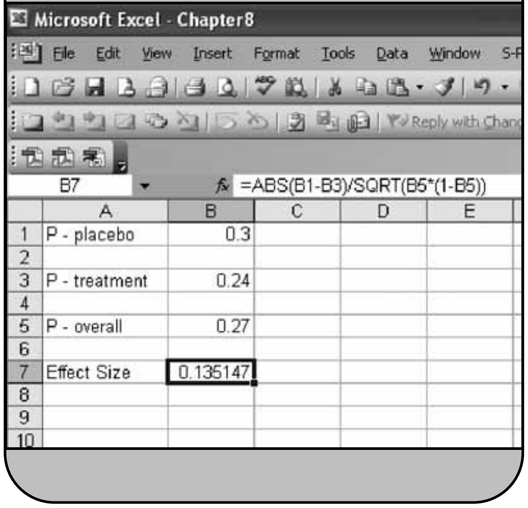
FIGURE 8–35 Computing Overall Proportions


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-P
Reply with Chang

B5 =(B1+B3)/2

	A	B	C	D	E
1	P - placebo	0.3			
2					
3	P - treatment	0.24			
4					
5	P - overall	0.27			
6					
7					
8					

FIGURE 8–36 Computing the Effect Size


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-P
Reply with Chang

B7 =ABS(B1-B3)/SQRT(B5*(1-B5))

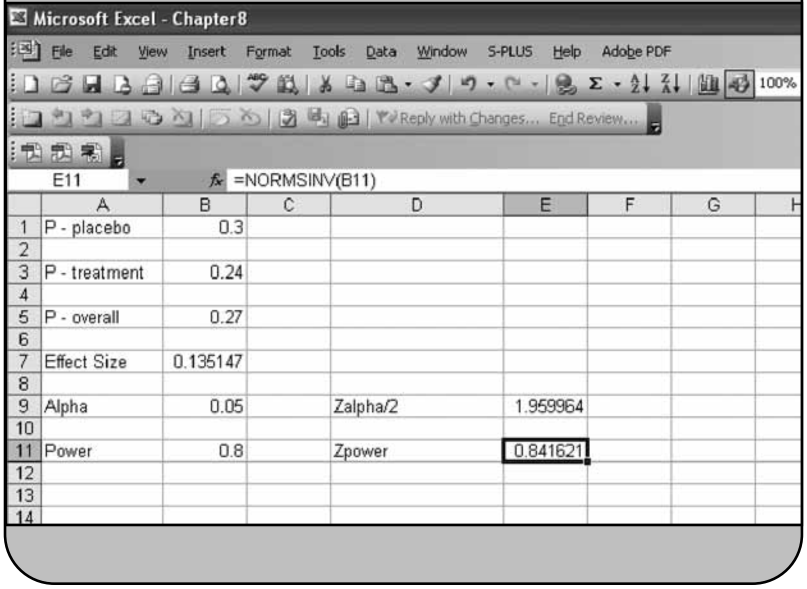
	A	B	C	D	E
1	P - placebo	0.3			
2					
3	P - treatment	0.24			
4					
5	P - overall	0.27			
6					
7	Effect Size	0.135147			
8					
9					
10					

We next enter the level of significance, α, and the desired power to compute $z_{1-\alpha/2}$ and the $z_{1-\beta}$. This is shown in Figure 8–37. The next step is to compute the sample size per group based on the effect size and the appropriate z values for the selected α and power. This is shown in Figure 8–38.

Finally, because the sample size formula always produces the minimum number of subjects per group required to ensure that the test has the specified power to detect the desired effect size at the specified level of significance, to determine the numbers of subjects per group required for the study, we must round up. This is done using Excel's ROUNDUP function. The sample sizes required per group are shown in Figure 8–39.

Samples of size $n_1 = 860$ patients on the new drug and $n_2 = 860$ patients on placebo will ensure that the test of hypothesis will have 80% power to detect a 20% reduction

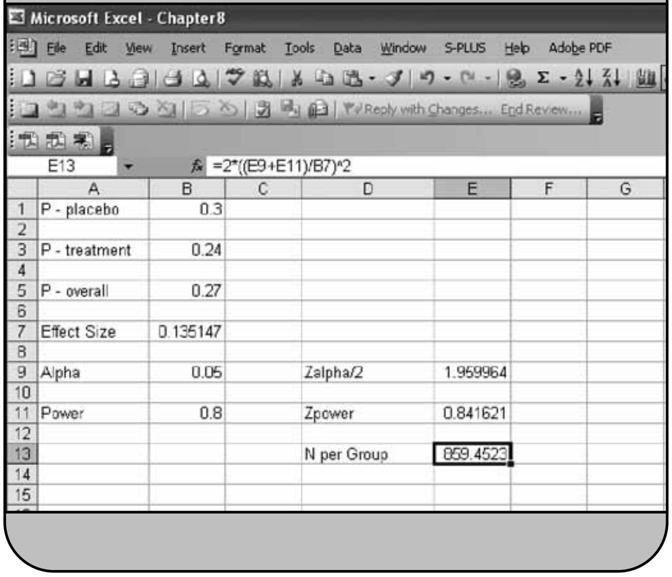
FIGURE 8–37 Computing *z* Values


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...

E11 =NORMSINV(B11)

	A	B	C	D	E	F	G	H
1	P - placebo	0.3						
2								
3	P - treatment	0.24						
4								
5	P - overall	0.27						
6								
7	Effect Size	0.135147						
8								
9	Alpha	0.05		Zalpha/2	1.959964			
10								
11	Power	0.8		Zpower	0.841621			
12								
13								
14								

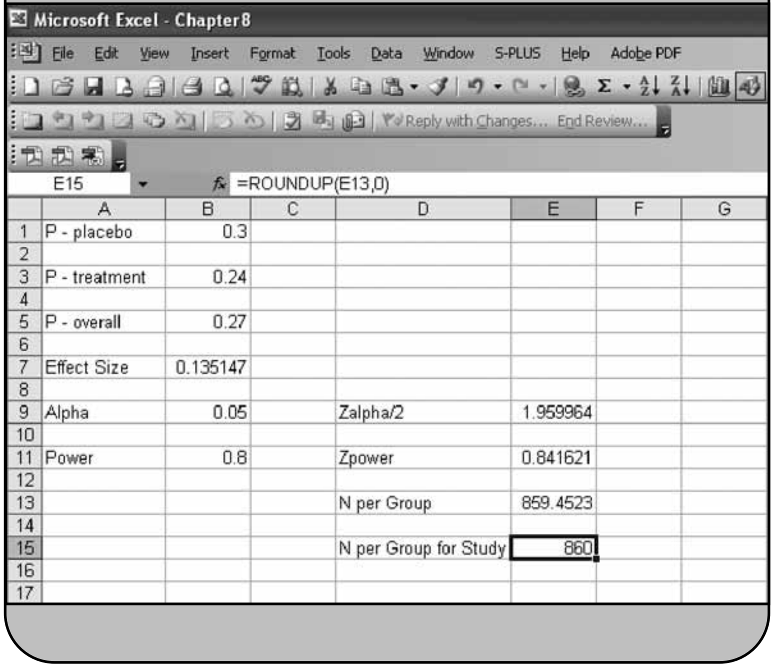
FIGURE 8–38 Determining the Sample Size Required per Group


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E13 =2*((E9+E11)/B7)^2

	A	B	C	D	E	F	G
1	P - placebo	0.3					
2							
3	P - treatment	0.24					
4							
5	P - overall	0.27					
6							
7	Effect Size	0.135147					
8							
9	Alpha	0.05		Zalpha/2	1.959964		
10							
11	Power	0.8		Zpower	0.841621		
12							
13				N per Group	859.4523		
14							
15							

FIGURE 8–39 Sample Size Required per Group for Study


Microsoft Excel - Chapter8
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Reply with Changes... End Review...

E15 =ROUNDUP(E13,0)

	A	B	C	D	E	F	G
1	P - placebo	0.3					
2							
3	P - treatment	0.24					
4							
5	P - overall	0.27					
6							
7	Effect Size	0.135147					
8							
9	Alpha	0.05		Zalpha/2	1.959964		
10							
11	Power	0.8		Zpower	0.841621		
12							
13				N per Group	859.4523		
14							
15				N per Group for Study	860		
16							
17							

in the proportions of patients who meet the criteria for hypertension.

Once the Excel formulas are programmed to compute the sample sizes required to ensure a specified power in a test of hypothesis, other scenarios can be considered easily by changing the inputs (e.g., α, the desired power, the difference in the parameter reflecting a clinically meaningful change, or the standard deviation).

8.12 PRACTICE PROBLEMS

1. We want to design a new placebo-controlled trial to evaluate an experimental medication to increase lung capacity. The primary outcome is peak expiratory flow rate, a continuous variable measured in liters per minute. The primary outcome will be measured after 6 months on treatment. The mean peak expiratory flow rate in adults is 300 with a standard deviation of 50. How many subjects should be enrolled to ensure 80% power to detect a difference of 15 liters per minute with a two-sided test and $\alpha = 0.05$?
2. An investigator wants to estimate caffeine consumption in high school students. How many students would be required to ensure that a 95% confidence interval estimate for the mean caffeine intake (measured in mg) is within 15 mg of the true mean? Assume that the standard deviation in caffeine intake is 68 mg.
3. Consider the study proposed in Problem 2. How many students would be required to estimate the proportion of students who consume coffee? Suppose we want the estimate to be within 5% of the true proportion with 95% confidence.
4. A clinical trial was conducted comparing a new compound designed to improve wound healing in trauma patients to a placebo. After treatment for 5 days, 58% of the patients taking the new compound had a substantial reduction in the size of their wound as compared to 44% in the placebo group. The trial failed to show significance. How many subjects would be required to detect the difference in proportions observed in the trial with 80% power? A two-sided test is planned at $\alpha = 0.05$.
5. A crossover trial is planned to evaluate the impact of an educational intervention program to reduce alcohol consumption in patients determined to be at risk for alcohol problems. The plan is to measure alcohol consumption (the number of drinks on a typical drinking day) before the intervention and then again

after participants complete the educational intervention program. How many participants would be required to ensure that a 95% confidence interval for the mean difference in the number of drinks is within two drinks of the true mean difference? Assume that the standard deviation of the difference in the mean number of drinks is 6.7 drinks.

6. An investigator wants to design a study to estimate the difference in the proportions of men and women who develop early-onset cardiovascular disease (defined as cardiovascular disease before age 50). A study conducted 10 years ago found that 15% and 8% of men and women, respectively, developed early onset cardiovascular disease. How many men and women are needed to generate a 95% confidence interval estimate for the difference in proportions with a margin of error not exceeding 4%?
7. The mean body mass index (BMI) for boys age 12 is 23.6. An investigator wants to test if the BMI is higher in boys age 12 living in New York City. How many boys are needed to ensure that a two-sided test of hypothesis has 80% power to detect a difference of 2 units in BMI? Assume that the standard deviation in BMI is 5.7.
8. An investigator wants to design a study to estimate the difference in the mean BMI between 12-year-old boys and girls living in New York City. How many boys and girls are needed to ensure that a 95% confidence interval estimate for the difference in mean BMI between boys and girls has a margin of error not exceeding 2 units? Use the estimate of the variability in BMI from Problem 7.

CHAPTER 9

Regression Analysis

In Chapter 9 of the textbook, we introduced regression analysis. We noted that regression analysis is a very general and widely applied technique. In the textbook, we focused more on the use of regression analysis to assess confounding and effect modification. We limit our focus here to estimating simple linear and multiple linear regression models using the linear regression tool in the Data Analysis ToolPak.

We use data collected from $n = 40$ randomly selected participants of the Sixth Examination of the Framingham Offspring Study to illustrate regression analysis using Microsoft Office® Excel®. The data are shown in Table 9–1 and include the participant's age (in years), gender (which is coded 1 for males and 0 for females), body mass index (BMI), systolic and diastolic blood pressures, total cholesterol, HDL cholesterol, diabetes (coded 1 for participants diagnosed with diabetes and 0 otherwise), and current smoking status (coded 1 for current smokers and 0 otherwise).

9.1 SIMPLE LINEAR REGRESSION ANALYSIS

In Chapter 9 of the textbook, we introduced simple linear regression analysis as a technique for estimating the equation that best describes the linear association between a continuous dependent or outcome variable, y, and a single independent or predictor variable, x. The independent variable can be continuous or dichotomous (sometimes called an indicator variable). The regression equation is as follows:

$$\hat{y} = b_0 + b_1x$$

where $\hat{y}$ is the predicted or expected value of the dependent or outcome variable, b_0 is the estimated y-intercept, and b_1 is the estimated slope. Excel has an analysis tool that can be used to estimate the y-intercept and slope.

Example 9.1. Suppose we wish to estimate the equation of the line that best describes the relationship between systolic blood pressure (SBP) and age. The data from Table 9–1 are entered into an Excel worksheet as shown in Figure 9–1. Note that the Excel worksheet contains $n = 40$ observations; only the first 20 are shown in Figure 9–1.

To estimate the simple linear regression equation, we use the **Tools/Data Analysis** menu option. We select the **Regression Analysis** tool as shown in Figure 9–2 and click **OK**. Excel then requests specification of the variables for analysis in the dialog box shown in Figure 9–3.

We first specify the dependent or outcome variable (y). In our example, the dependent variable is systolic blood pressure, which is contained in cell D1 through cell D41. We then specify the independent variable (x), which in this example is age. The age data is contained in cell A1 through cell A41. Because we include the first row of labels (A1 and D1), we click on the **Labels** box to indicate that the labels are contained in these cells. We then specify a location for the results of the regression analysis. For this example, we request that Excel place the results in a new worksheet entitled *Simple regression*. Excel offers a number of additional details, such as analysis of residuals and normal probability plots. These are used to examine the fit of the regression equation, and are called regression diagnostics. (We introduced only the basic applications of regression analysis in the textbook, and we restrict our attention to the same in the Excel applications.) The results of the regression analysis are shown in Figure 9–4.

TABLE 9–1 Data from $n = 40$ Randomly Selected Participants of the Sixth Examination of the Framingham Offspring Study

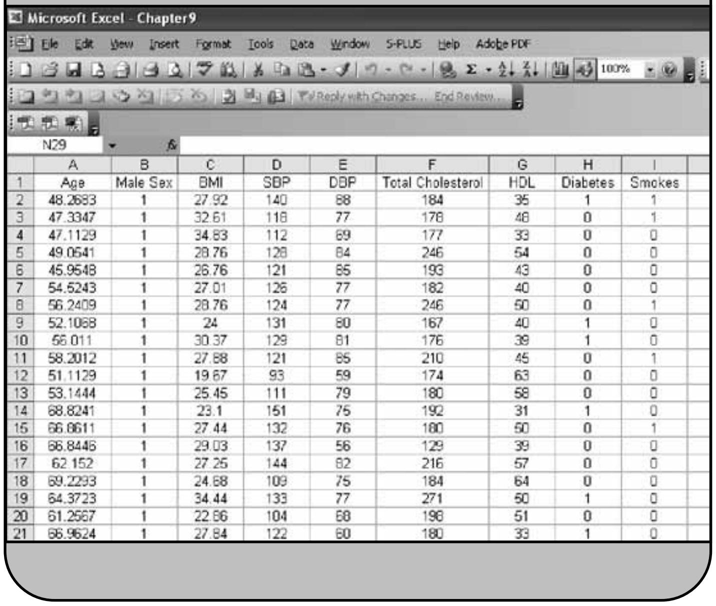
Age	Male	BMI	SBP	DBP	Total Cholesterol	HDL	Diabetes	Smoke
48.2683	1	27.92	140	88	184	35	1	1
47.3347	1	32.61	118	77	178	48	0	1
47.1129	1	34.83	112	69	177	33	0	0
49.0541	1	28.76	128	84	246	54	0	0
45.9548	1	26.76	121	85	193	43	0	0
54.5243	1	27.01	126	77	182	40	0	0
56.2409	1	28.76	124	77	246	50	0	1
52.1068	1	24	131	80	167	40	1	0
56.011	1	30.37	129	81	176	39	1	0
58.2012	1	27.88	121	85	210	45	0	1
51.1129	1	19.67	93	59	174	63	0	0
53.1444	1	25.45	111	79	180	58	0	0
68.8241	1	23.1	151	75	192	31	1	0
66.8611	1	27.44	132	76	180	50	0	1
66.8446	1	29.03	137	56	129	39	0	0
62.152	1	27.25	144	82	216	57	0	0
69.2293	1	24.68	109	75	184	64	0	0
64.3723	1	34.44	133	77	271	50	1	0
61.2567	1	22.86	104	68	198	51	0	0
66.9624	1	27.84	122	60	180	33	1	0
71.7454	1	27.7	137	81	198	44	1	0
71.0089	1	31.04	136	75	213	62	0	0
77.4456	1	34.06	110	57	181	45	0	0
34.6557	0	21.8	99	60	178	33	0	0
59.0773	0	23.59	124	76	212	47	0	0
45.7659	0	22.39	118	77	258	56	0	0
55.9808	0	26.18	110	66	263	50	0	0
47.5729	0	24.86	103	66	183	47	0	0
59.4798	0	32.89	123	85	203	40	0	0
58.3381	0	24.47	118	61	230	81	0	0
50.0589	0	21.98	110	68	168	72	0	1
52.6845	0	25.12	105	67	201	61	0	0
51.7016	0	39.93	131	80	197	43	0	0
58.8255	0	38.14	107	69	224	29	0	0
64.5859	0	25.86	138	68	205	53	0	0
67.0418	0	30.95	135	72	210	36	0	0
62.642	0	31.99	123	65	209	70	0	0
71.8248	0	19.03	103	50	206	63	0	0
76.6899	0	21.8	137	85	176	74	0	0
73.9932	0	33.07	135	80	254	57	0	0

Excel produces a number of statistics and analyses in its standard regression analysis. We again focus only on the analyses discussed in the textbook. Specifically, the estimates of the regression coefficients are in the last section of the results under the column headed *Coefficients*. The estimate of the y-intercept is $b_0 = 89.40$ and the estimate of the slope is $b_1 = 0.56$. (Notice that the slope is the coefficient associated with age.) The regression equation relating age to systolic blood pressure is:

$$\hat{y} = 89.40 + 0.56(\text{Age})$$

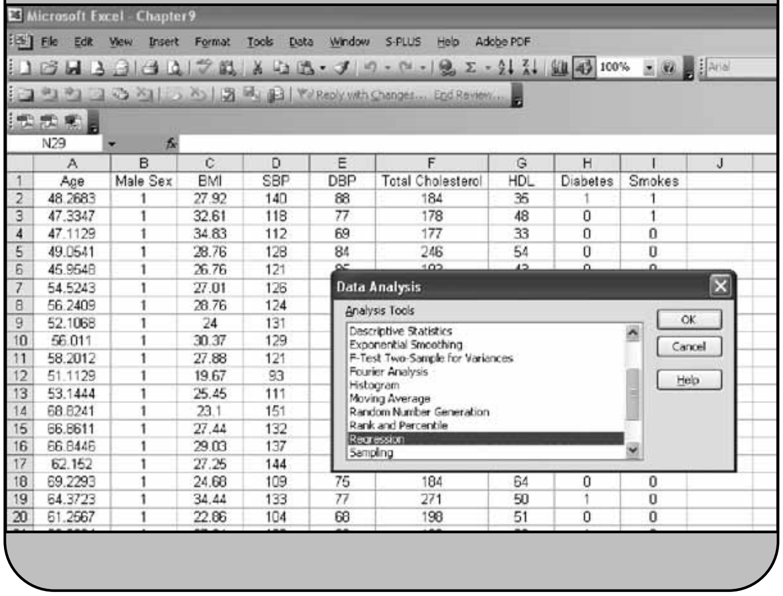
where $\hat{y}$ is the predicted or expected SBP. Excel also provides standard errors of the regression coefficients, t statistics, and

FIGURE 9–1 Data for Regression Analysis


Microsoft Excel - Chapter9
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Reply with Changes... End Review...
N29
fx

	A	B	C	D	E	F	G	H	I
1	Age	Male Sex	BMI	SBP	DBP	Total Cholesterol	HDL	Diabetes	Smokes
2	48.2683	1	27.92	140	88	184	35	1	1
3	47.3347	1	32.61	118	77	178	48	0	1
4	47.1129	1	34.83	112	69	177	33	0	0
5	49.0641	1	28.76	128	84	246	54	0	0
6	45.9548	1	26.76	121	85	193	43	0	0
7	54.5243	1	27.01	126	77	182	40	0	0
8	56.2409	1	28.76	124	77	246	50	0	1
9	52.1068	1	24	131	80	167	40	1	0
10	56.011	1	30.37	129	81	176	39	1	0
11	58.2012	1	27.88	121	85	210	45	0	1
12	51.1129	1	19.67	93	59	174	63	0	0
13	53.1444	1	25.45	111	79	180	58	0	0
14	68.8241	1	23.1	151	75	192	31	1	0
15	66.8611	1	27.44	132	76	180	50	0	1
16	66.8446	1	29.03	137	56	129	39	0	0
17	62.152	1	27.25	144	82	216	57	0	0
18	69.2293	1	24.68	109	75	184	64	0	0
19	64.3723	1	34.44	133	77	271	50	1	0
20	61.2567	1	22.86	104	68	198	51	0	0
21	66.9624	1	27.84	122	80	180	33	1	0

FIGURE 9–2 Regression Analysis Tool


Microsoft Excel - Chapter9
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
N29
fx
Data Analysis
OK
Cancel
Help

	A	B	C	D	E	F	G	H	I	J
1	Age	Male Sex	BMI	SBP	DBP	Total Cholesterol	HDL	Diabetes	Smokes	
2	48.2683	1	27.92	140	88	184	35	1	1	
3	47.3347	1	32.61	118	77	178	48	0	1	
4	47.1129	1	34.83	112	69	177	33	0	0	
5	49.0641	1	28.76	128	84	246	54	0	0	
6	45.9548	1	26.76	121						
7	54.5243	1	27.01	126						
8	56.2409	1	28.76	124						
9	52.1068	1	24	131						
10	56.011	1	30.37	129						
11	58.2012	1	27.88	121						
12	51.1129	1	19.67	93						
13	53.1444	1	25.45	111						
14	68.8241	1	23.1	151						
15	66.8611	1	27.44	132						
16	66.8446	1	29.03	137						
17	62.152	1	27.25	144						
18	69.2293	1	24.68	109	75	184	64	0	0	
19	64.3723	1	34.44	133	77	271	50	1	0	
20	61.2567	1	22.86	104	68	198	51	0	0	

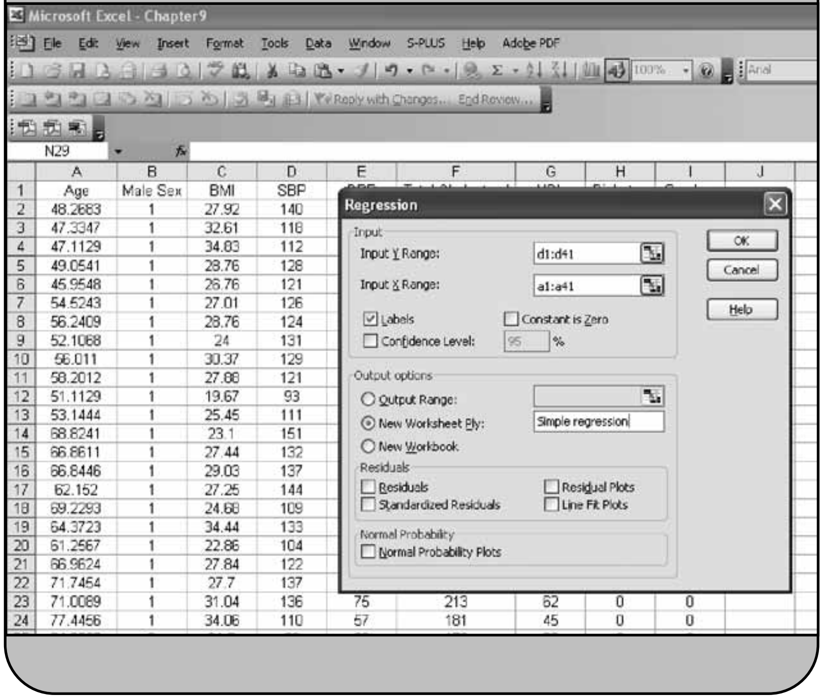
Analysis Tools

- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling

p-values to test whether the regression coefficients are statistically significantly different from zero. Usually, we are not interested in whether the intercept is significantly different from zero. However, it is of interest to test whether the slope in the population is significantly different from zero.

Specifically, we test H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$. Excel provides a *p*-value of 0.0107, indicating that there is a statistically significant association between age and systolic blood pressure. The regression equation indicates that each additional year of age is associated with a 0.56-unit increase in

FIGURE 9–3 Specification of Variables for Simple Linear Regression Analysis


Microsoft Excel - Chapter9
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
Reply with Changes... End Review...
N29
fx

	A	B	C	D	E	F	G	H	I	J
1	Age	Male Sex	BMI	SBP						
2	48.2683	1	27.92	140						
3	47.3347	1	32.61	118						
4	47.1129	1	34.83	112						
5	49.0541	1	28.76	128						
6	45.9548	1	26.76	121						
7	54.5243	1	27.01	126						
8	56.2409	1	28.76	124						
9	52.1068	1	24	131						
10	56.011	1	30.37	129						
11	58.2012	1	27.88	121						
12	51.1129	1	19.67	93						
13	53.1444	1	25.45	111						
14	68.8241	1	23.1	151						
15	66.8611	1	27.44	132						
16	66.8446	1	29.03	137						
17	62.152	1	27.25	144						
18	69.2293	1	24.68	109						
19	64.3723	1	34.44	133						
20	61.2567	1	22.86	104						
21	66.9624	1	27.84	122						
22	71.7454	1	27.7	137						
23	71.0089	1	31.04	136	75	213	62	0	0	
24	77.4456	1	34.06	110	57	181	45	0	0	

Regression

Input

Input Y Range: d1:d41

Input X Range: a1:a41

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply: Simple regression

○ New Workbook

Residuals

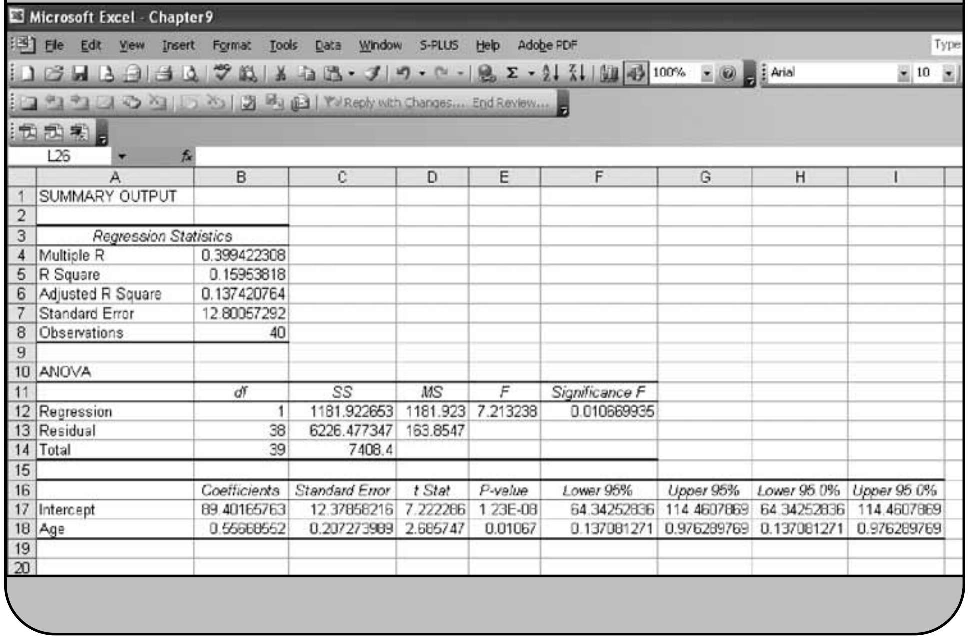
☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

FIGURE 9–4 Results of Simple Linear Regression Analysis


Microsoft Excel - Chapter9
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type
100%
Arial
10
Reply with Changes... End Review...
L26
fx

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.399422308							
5	R Square	0.15953818							
6	Adjusted R Square	0.137420764							
7	Standard Error	12.80057292							
8	Observations	40							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	1181.922653	1181.923	7.213238	0.010669935			
13	Residual	38	6226.477347	163.8547					
14	Total	39	7408.4						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	89.40165763	12.37858216	7.222286	1.23E-08	64.34252836	114.4607869	64.34252836	114.4607869
18	Age	0.55668552	0.207273989	2.685747	0.01067	0.137081271	0.976289769	0.137081271	0.976289769
19									
20									

systolic blood pressure. (The other analyses that Excel generates are useful and interested readers should see some of the references at the end of Chapter 9 in the textbook for more details.)

Example 9.2. Suppose we wish to assess whether there is an association between systolic blood pressure and current smoking status using the data in Table 9–1, which were entered into an Excel worksheet as shown in Figure 9–1. We again use the **Tools/Data Analysis** menu option and select the **Regression Analysis** tool. When we click **OK**, Excel requests specification of the variables for analysis in the dialog box shown in Figure 9–5.

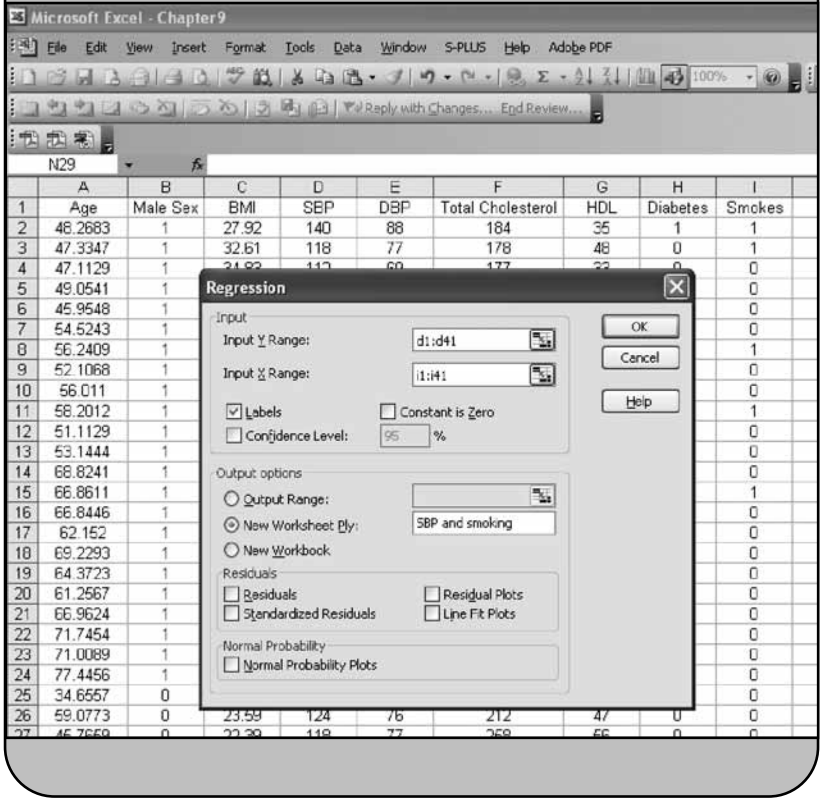
We again specify the dependent or outcome variable (y). In this example, the outcome is systolic blood pressure, which is contained in cell D1 through cell D41. We then specify the independent variable (x), which in this example is smoking status. The smoking data is contained in cell I1 through cell I41. Because we included the first row of labels (I1 and D1), we click on the **Labels** box to indicate that the labels are contained in these cells. We then specify a location for the results of the regression analysis. For this example, we request that Excel place the results in a new worksheet entitled *SBP and smoking*. The results of the regression analysis are shown in Figure 9–6.

The estimate of the y-intercept is $b_0 = 121.85$ and the estimate of the slope is $b_1 = 2.31$. The regression equation relating current smoking status to systolic blood pressure is:

$$\hat{y} = 121.85 + 2.31 \text{ (Current Smoking Status)}$$

where $\hat{y}$ is the predicted or expected SBP. The p-value for the test of signif-icance for the slope is $p = 0.7098$, indicating that there is no statistically significant association between current smoking status and systolic blood pressure in the population. The regression equation indicates that smokers have higher systolic blood pressures by approximately 2.31 units, as compared to nonsmokers. However, this difference is not statistically significantly different from zero (because $p = 0.7098$).

FIGURE 9–5 Specification of Variables for Simple Linear Regression Analysis



N29

	A	B	C	D	E	F	G	H	I
1	Age	Male Sex	BMI	SBP	DBP	Total Cholesterol	HDL	Diabetes	Smokes
2	48.2683	1	27.92	140	88	184	35	1	1
3	47.3347	1	32.61	118	77	178	48	0	1
4	47.1129	1							0
5	49.0541	1							0
6	45.9548	1							0
7	54.5243	1							0
8	56.2409	1							1
9	52.1068	1							0
10	56.011	1							0
11	58.2012	1							1
12	51.1129	1							0
13	53.1444	1							0
14	68.8241	1							0
15	66.8611	1							1
16	66.8446	1							0
17	62.152	1							0
18	69.2293	1							0
19	64.3723	1							0
20	61.2567	1							0
21	66.9624	1							0
22	71.7454	1							0
23	71.0089	1							0
24	77.4456	1							0
25	34.6557	0							0
26	59.0773	0	23.59	124	76	212	47	0	0

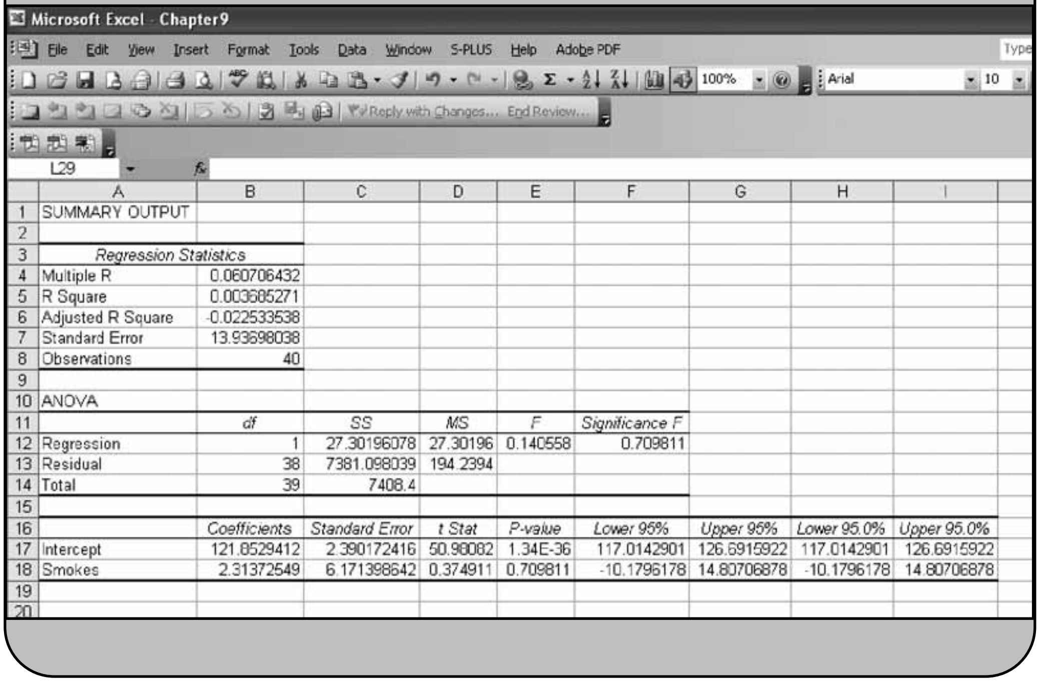
Input
Input Y Range: d1:d41
Input X Range: i1:i41
☑ Labels □ Constant is Zero
□ Confidence Level: 95 %

Output options
○ Output Range:
● New Worksheet Ply: SBP and smoking
○ New Workbook

Residuals
□ Residuals □ Residual Plots
□ Standardized Residuals □ Line Fit Plots

Normal Probability
□ Normal Probability Plots

FIGURE 9–6 Results of Simple Linear Regression Analysis


Microsoft Excel - Chapter9
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type
100%
Arial
10
Reply with Changes... End Review...
L29
fx

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.060706432							
5	R Square	0.003685271							
6	Adjusted R Square	-0.022533538							
7	Standard Error	13.93698038							
8	Observations	40							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	27.30196078	27.30196	0.140558	0.709811			
13	Residual	38	7381.098039	194.2394					
14	Total	39	7408.4						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	121.8529412	2.390172416	50.98082	1.34E-36	117.0142901	126.6915922	117.0142901	126.6915922
18	Smokes	2.31372549	6.171398642	0.374911	0.709811	-10.1796178	14.80706878	-10.1796178	14.80706878
19									
20									

9.2 MULTIPLE LINEAR REGRESSION ANALYSIS

In Chapter 9 of the textbook, we introduced multiple linear regression analysis as a technique for estimating the equation that best describes the association between a continuous outcome variable y and a set of independent variables, x_1, x_2, ..., x_p. The independent variables can be continuous or dichotomous. The regression equation is as follows:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p,$$

where $\hat{y}$ is the predicted or expected value of the dependent variable, x_1 through x_p are p distinct independent or predictor variables, b_0 is the value of y when all of the independent variables (x_1 through x_p) are equal to zero, and b_1 through b_p are the estimated regression coefficients. Excel has an analysis tool that can be used to estimate the coefficients of a multiple regression equation.

Example 9.3. Suppose we again consider systolic blood pressure as our dependent or outcome variable. We now wish to assess the association between age and sex, considered simultaneously, and SBP using the data in Table 9–1. We again use the **Tools/Data Analysis** menu option and select the **Regression Analysis** tool. When we click **OK**, Excel requests specification of the variables for analysis in the dialog box shown in Figure 9–7.

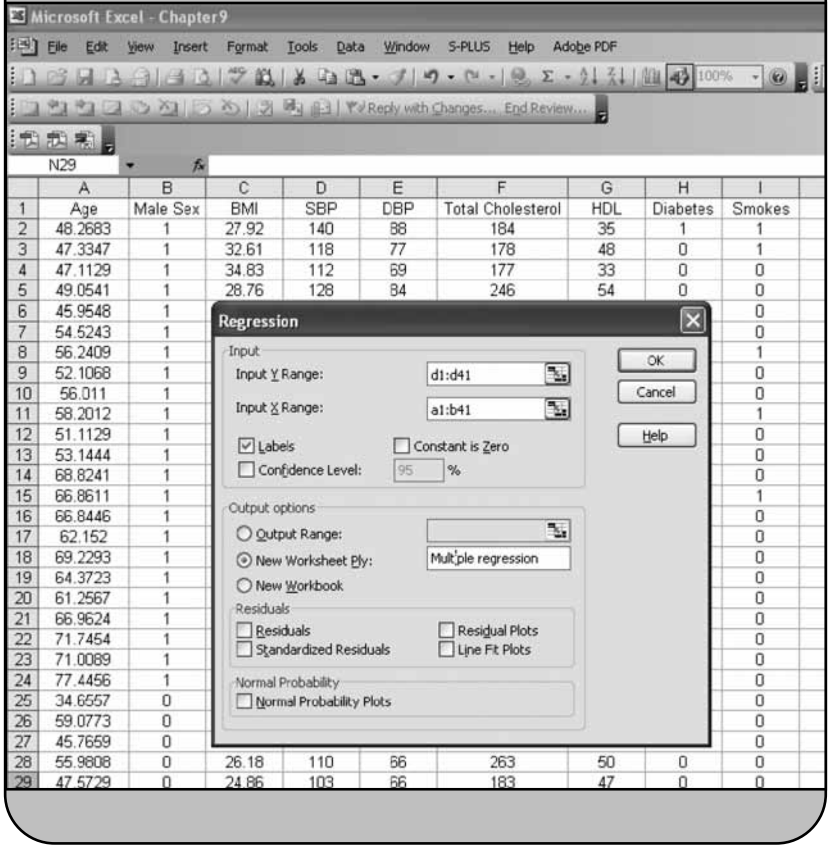
We again specify the dependent or outcome variable (y). In Example 9.3, the outcome is systolic blood pressure, which is contained in cell D1 through cell D41. We then specify the independent variables (x_1 and x_2), which in this example are age and sex. The age data is contained in cell A1 through cell A41 and the sex data is contained in cell B1 though cell B41 (in this example, the variable is an indicator of male gender: 1 = male, 0 = female). The range "A1:B41" includes both independent variables. Because we included the first row of labels (A1, B1, and D1), we click on the **Labels** box to indicate that the labels are contained in these cells. We then specify a location for the results of the regression analysis. For this example, we request that Excel place the results in a new worksheet entitled *Multiple regression*. The results of the regression analysis are shown in Figure 9–8.

The estimate of the coefficients of the multiple regression equation are $b_0 = 87.19$, $b_1 = 0.54$, and $b_2 = 5.38$. The regression equation relating age and sex to systolic blood pressure is

$$\hat{y} = 87.19 + 0.54(\text{Age}) + 5.38(\text{Male sex})$$

where $\hat{y}$ is the estimated SBP. The p-values for the tests of significance for the regression coefficients associated with

FIGURE 9–7 Specification of Variables for Multiple Linear Regression Analysis



	A	B	C	D	E	F	G	H	I
1	Age	Male Sex	BMI	SBP	DBP	Total Cholesterol	HDL	Diabetes	Smokes
2	48.2683	1	27.92	140	88	184	35	1	1
3	47.3347	1	32.61	118	77	178	48	0	1
4	47.1129	1	34.83	112	69	177	33	0	0
5	49.0541	1	28.76	128	84	246	54	0	0
6	45.9548	1							0
7	54.5243	1							0
8	56.2409	1							1
9	52.1068	1							0
10	56.011	1							0
11	58.2012	1							1
12	51.1129	1							0
13	53.1444	1							0
14	68.8241	1							0
15	66.8611	1							1
16	66.8446	1							0
17	62.152	1							0
18	69.2293	1							0
19	64.3723	1							0
20	61.2567	1							0
21	66.9624	1							0
22	71.7454	1							0
23	71.0089	1							0
24	77.4456	1							0
25	34.6557	0							0
26	59.0773	0							0
27	45.7659	0							0
28	55.9808	0	26.18	110	66	263	50	0	0
29	47.5729	0	24.86	103	66	183	47	0	0

Regression

Input

Input Y Range: d1:d41

Input X Range: a1:b41

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply: Mult'ple regression

○ New Workbook

Residuals

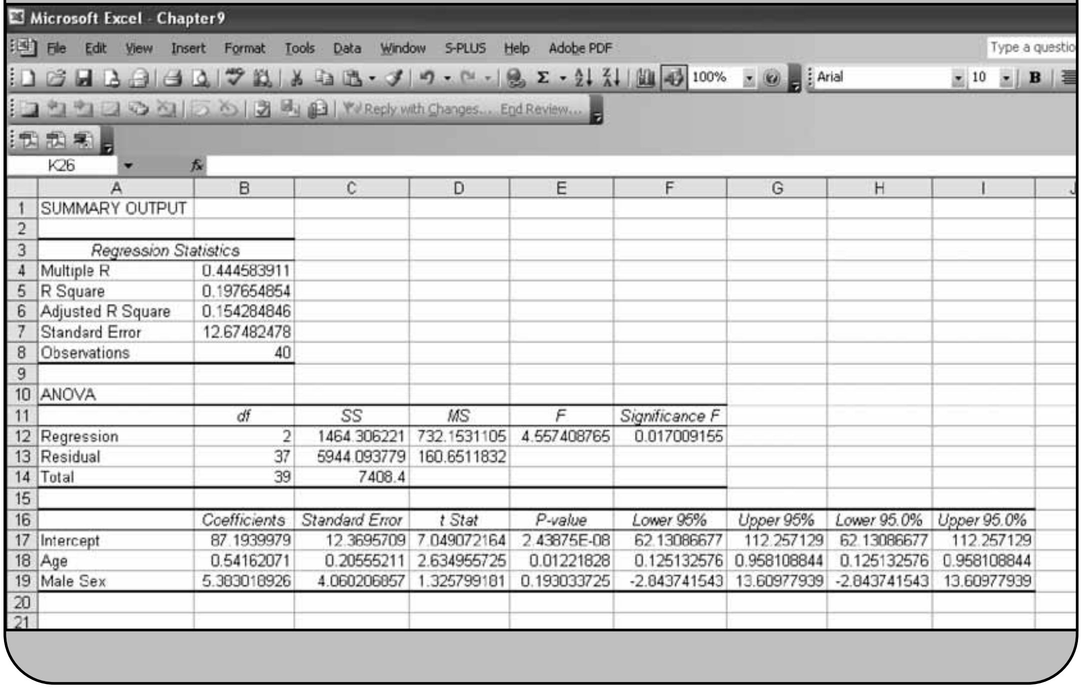
☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

FIGURE 9–8 Results of Multiple Linear Regression Analysis



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.444583911							
5	R Square	0.197654854							
6	Adjusted R Square	0.154284846							
7	Standard Error	12.67482478							
8	Observations	40							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	1464.306221	732.1531105	4.557408765	0.017009155			
13	Residual	37	5944.093779	160.6511832					
14	Total	39	7408.4						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	87.1939979	12.3695709	7.049072164	2.43875E-08	62.13086677	112.257129	62.13086677	112.257129
18	Age	0.54162071	0.20555211	2.634955725	0.01221828	0.125132576	0.958108844	0.125132576	0.958108844
19	Male Sex	5.383018926	4.060206857	1.325799181	0.193033725	-2.843741543	13.60977939	-2.843741543	13.60977939
20									
21									

age and sex are $p = 0.0122$ and $p = 0.1930$, respectively. The p-values indicate that there is a statistically significant association between age and systolic blood pressure accounting for sex, but not between sex and systolic blood pressure, once age is considered. The multiple regression equation indicates that each additional year of age is associated with a 0.54-unit increase in systolic blood pressure, holding sex constant, and that men have higher systolic blood pressures than women by about 5.38 units, holding age constant.

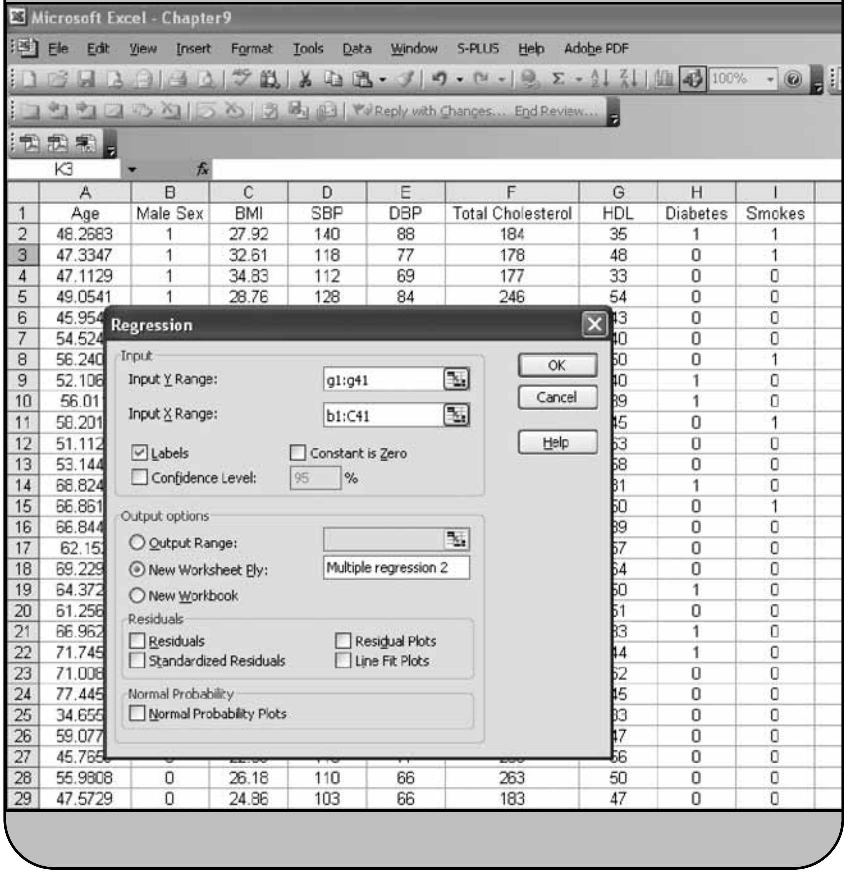
Example 9.4. We now consider HDL as our dependent or outcome variable and want to assess the association between BMI and sex, considered simultaneously, and HDL using the data in Table 9–1. We again use the **Tools/Data Analysis** menu option and select the **Regression Analysis** tool. When we click **OK**, Excel requests specification of the variables for analysis in the dialog box shown in Figure 9–9.

We first specify the location of the data for our dependent or outcome variable (y). In this example, the outcome is HDL, which is contained in cell G1 through cell G41. We then specify the independent variables (x_1 and x_2), which in this example are sex and BMI. The sex data is contained in cell B1 through cell B41 and the BMI data is contained in cell C1 though cell C41. The range "B1:C41" includes both independent variables. Because we included the first row of labels (B1, C1, and G1), we click on the **Labels** box to indicate that the labels are contained in these cells. We then specify a location for the results of the regression analysis. For this example, we request that Excel place the results in a new worksheet entitled *Multiple regression 2*. The results of the regression analysis are shown in Figure 9–10.

The estimates of the coefficients of the multiple regression equation are as follows: $b_0 = 79.38$, $b_1 = -6.31$, and $b_2 = -0.94$. The regression equation relating sex and BMI to HDL is:

$$\hat{y} = 79.38 - 6.31(\text{Male sex}) - 0.94(\text{BMI})$$

FIGURE 9–9 Specification of Variables for Multiple Linear Regression Analysis



	A	B	C	D	E	F	G	H	I
1	Age	Male Sex	BMI	SBP	DBP	Total Cholesterol	HDL	Diabetes	Smokes
2	48.2683	1	27.92	140	88	184	35	1	1
3	47.3347	1	32.61	118	77	178	48	0	1
4	47.1129	1	34.83	112	69	177	33	0	0
5	49.0541	1	28.76	128	84	246	54	0	0
6	45.954						43	0	0
7	54.524						40	0	0
8	56.240						50	0	1
9	52.106						40	1	0
10	56.01						39	1	0
11	58.201						45	0	1
12	51.112						53	0	0
13	53.144						58	0	0
14	68.824						31	1	0
15	66.861						50	0	1
16	66.844						39	0	0
17	62.15						57	0	0
18	69.229						54	0	0
19	64.372						50	1	0
20	61.256						51	0	0
21	66.962						33	1	0
22	71.745						44	1	0
23	71.006						52	0	0
24	77.445						45	0	0
25	34.655						33	0	0
26	59.077						47	0	0
27	45.765						56	0	0
28	55.9808	0	26.18	110	66	263	50	0	0
29	47.5729	0	24.86	103	66	183	47	0	0

Input

Input Y Range: g1:g41

Input X Range: b1:C41

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply: Multiple regression 2

○ New Workbook

Residuals

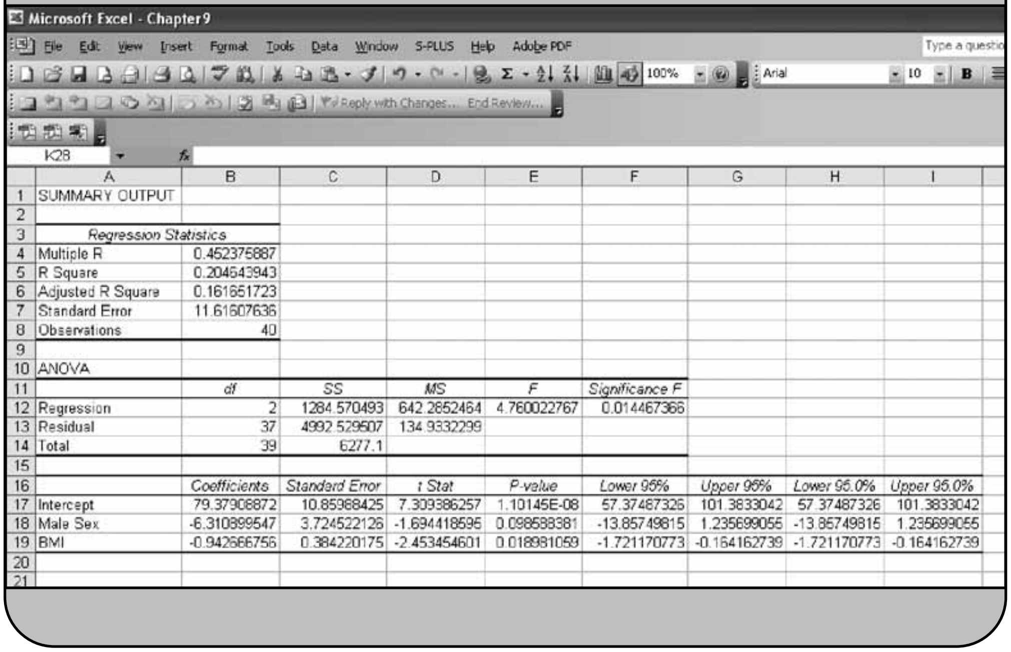
☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

FIGURE 9–10 Results of Multiple Linear Regression Analysis


Microsoft Excel - Chapter9
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
Type a questio
100%
Arial
10
B
Reply with Changes... End Review...
K28
fx

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.452375887							
5	R Square	0.204643943							
6	Adjusted R Square	0.161651723							
7	Standard Error	11.61607636							
8	Observations	40							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	1284.570493	642.2852464	4.760022767	0.014467366			
13	Residual	37	4992.529507	134.9332299					
14	Total	39	6277.1						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	79.37908872	10.85988425	7.309386257	1.10145E-08	57.37487326	101.3833042	57.37487326	101.3833042
18	Male Sex	-6.310899547	3.724522126	-1.694418595	0.098588381	-13.85749815	1.235699055	-13.85749815	1.235699055
19	BMI	-0.942666756	0.384220175	-2.453454601	0.018981059	-1.721170773	-0.164162739	-1.721170773	-0.164162739
20									
21									

where $\hat{y}$ is the predicted or expected HDL. The p-values for the tests of significance for the regression coefficients associated with gender and BMI are $p = 0.0986$ and $p = 0.0190$, respectively. The p-values indicate that there is a marginally significant association between sex and HDL (often when p-values fall in the range of 0.05 to 0.10, they are described as marginally significant), accounting for BMI, and a statistically significant association between BMI and HDL, accounting for sex. The multiple regression equation indicates that men have lower HDL than women by about 6.31 units, holding BMI constant, and that each additional unit of BMI is associated with a 0.94-unit reduction in HDL.Thus, increased BMI is associated with decreased HDL. Recall that HDL is the "good cholesterol" and that higher values are healthier.

It is important to note that for multiple regression analysis, the independent or predictor variables (x_1, x_2, x_3, ... , x_p) must be in adjacent columns in the Excel worksheet. When we specify the location of the cells containing the independent variables (i.e., Input *x* Range text field, Figure 9–9), we specify the locations of the first and last cells in the adjacent columns containing the data. For example, suppose in Example 9.3 we wished to consider sex and diabetes as the independent variables. To use the Regression Analysis Tool (to correctly specify these independent variables), we would need to reorganize the data in the Excel worksheet so that gender and diabetes are in adjacent columns. This can be done in several different ways. An easy way is to copy the data from column B and column H into column K and column L, as shown in Figure 9–11.

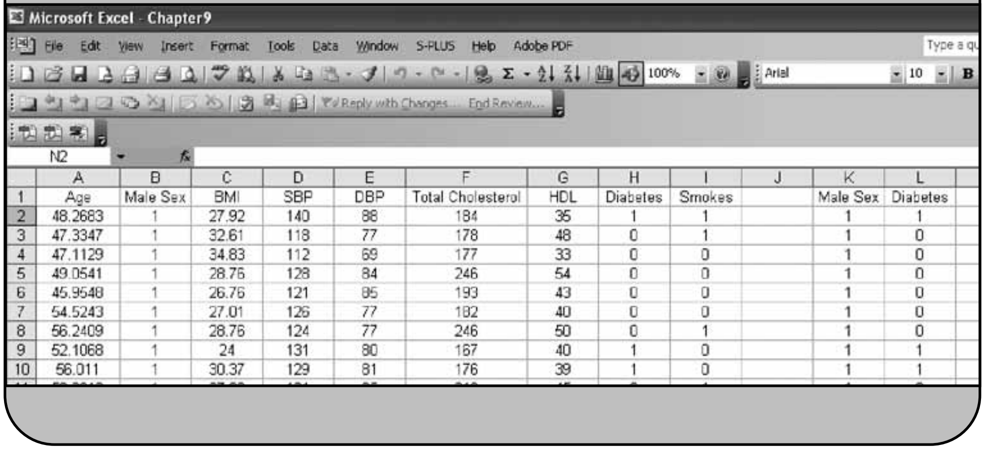
To estimate the multiple regression equation, we use the **Tools/Data Analysis** menu option and select the **Regression Analysis** tool. When we click **OK**, Excel requests specification of the variables for analysis in the dialog box shown in Figure 9–12.

We again specify the location of the data for our dependent or outcome variable (y = HDL), which is contained in cell G1 through cell G41. We then specify the independent variables (x_1 and x_2, or sex and diabetes), which are now contained in cell K1 through cell L41. The analysis is performed as described in Example 9.3 and Example 9.4.

9.3 PRACTICE PROBLEMS

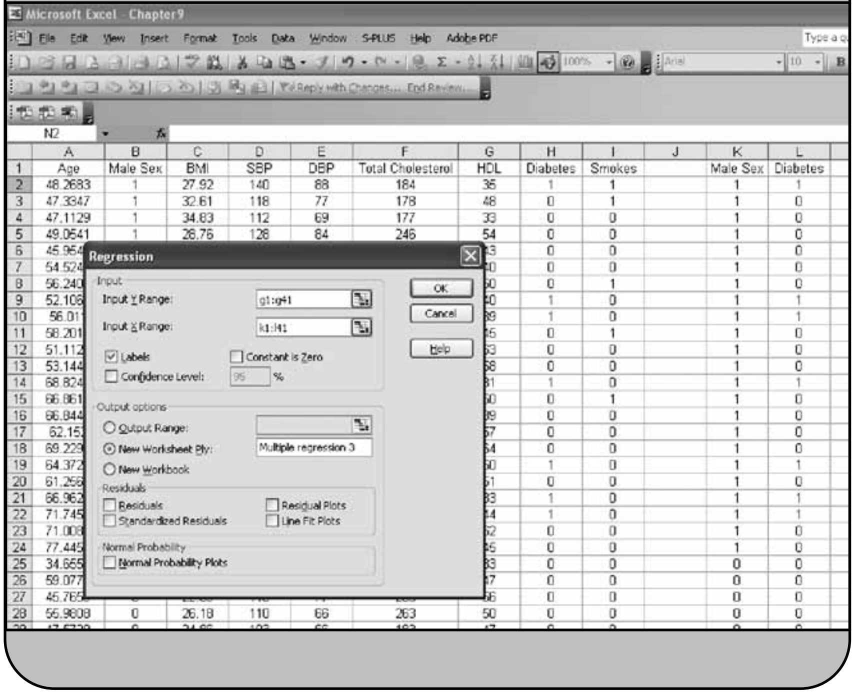
1. Consider the data shown in Table 9–2 measured in a sample of $n = 25$ undergraduates in an on-campus survey of health behaviors. Enter the data into an Excel worksheet for analysis.

FIGURE 9–11 Organizing the Data for a Multiple Regression Analysis


Microsoft Excel - Chapter9
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
10
Reply with Changes... End Review...
N2

	A	B	C	D	E	F	G	H	I	J	K	L
1	Age	Male Sex	BMI	SBP	DBP	Total Cholesterol	HDL	Diabetes	Smokes		Male Sex	Diabetes
2	48.2683	1	27.92	140	88	184	35	1	1		1	1
3	47.3347	1	32.61	118	77	178	48	0	1		1	0
4	47.1129	1	34.83	112	69	177	33	0	0		1	0
5	49.0541	1	28.76	128	84	246	54	0	0		1	0
6	45.9548	1	26.76	121	85	193	43	0	0		1	0
7	54.5243	1	27.01	126	77	182	40	0	0		1	0
8	56.2409	1	28.76	124	77	246	50	0	1		1	0
9	52.1068	1	24	131	80	167	40	1	0		1	1
10	56.011	1	30.37	129	81	176	39	1	0		1	1

FIGURE 9–12 Specification of Variables for Multiple Linear Regression Analysis


Microsoft Excel - Chapter9
File Edit View Insert Format Tools Data Window S-PLUS Help Adobe PDF
100%
Arial
10
Reply with Changes... End Review...
N2
Regression
OK
Cancel
Help

	A	B	C	D	E	F	G	H	I	J	K	L
1	Age	Male Sex	BMI	SBP	DBP	Total Cholesterol	HDL	Diabetes	Smokes		Male Sex	Diabetes
2	48.2683	1	27.92	140	88	184	35	1	1		1	1
3	47.3347	1	32.61	118	77	178	48	0	1		1	0
4	47.1129	1	34.83	112	69	177	33	0	0		1	0
5	49.0541	1	26.76	128	84	246	54	0	0		1	0
6	45.954							0	0		1	0
7	54.524							0	0		1	0
8	56.240							0	1		1	0
9	52.106							1	0		1	1
10	56.01							1	0		1	1
11	58.201							0	1		1	0
12	51.112							0	0		1	0
13	53.144							0	0		1	0
14	68.824							1	0		1	1
15	66.861							0	1		1	0
16	66.844							0	0		1	0
17	62.15							0	0		1	0
18	69.229							0	0		1	0
19	64.372							1	0		1	1
20	61.256							0	0		1	0
21	66.962							1	0		1	1
22	71.745							1	0		1	1
23	71.008							0	0		1	0
24	77.445							0	0		1	0
25	34.655							0	0		0	0
26	59.077							0	0		0	0
27	45.765							0	0		0	0
28	55.9808	0	26.18	110	66	263	50	0	0		0	0

Input

Input Y Range: g1:g41

Input X Range: k1:l41

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply: Multiple regression 3

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

TABLE 9–2 Data for Practice Problems

ID	Age	Female	Year in School	GPA	Current Smoker	Exercise per Week (h)	Average Drinks per Week	Cups of Coffee per Week
1	18	1	Fr	3.85	1	7	3	3
2	21	0	Jr	3.27	1	3	2	4
3	19	1	So	2.90	0	0	4	7
4	22	0	Sr	3.65	1	0	2	4
5	21	1	Sr	3.41	1	0	1	3
6	20	0	Jr	3.20	0	2	5	8
7	19	1	Jr	2.89	1	1	4	10
8	17	0	Fr	3.75	0	6	0	0
9	18	0	So	4.00	0	6	2	6
10	17	1	So	3.18	0	3	5	7
11	21	0	Jr	2.58	1	3	12	12
12	22	1	Sr	2.98	0	2	3	4
13	19	0	Fr	3.16	1	2	0	6
14	21	1	Jr	3.36	1	3	1	2
15	22	1	So	3.72	0	6	3	0
16	19	0	So	3.30	1	4	0	6
17	16	0	Fr	3.28	0	4	0	5
18	22	0	Sr	2.98	0	0	8	5
19	17	1	Fr	3.90	0	7	0	2
20	20	1	Sr	3.78	1	4	6	2
21	21	1	So	3.26	1	2	3	4
22	23	0	Jr	3.01	0	1	9	7
23	23	0	Sr	3.83	1	5	4	4
24	17	1	Fr	3.76	0	5	2	1
25	22	1	Sr	3.05	0	1	5	5

2. Using the data in Table 9–2, estimate the simple linear regression equation relating number of cups of coffee per week to GPA (consider GPA the dependent or outcome variable).
3. Using the data in Table 9–2, estimate the simple linear regression equation relating female gender to GPA (consider GPA the dependent or outcome variable).
4. Using the data in Table 9–2, estimate the multiple linear regression equation relating number of cups of coffee per week, female gender, and number of hours of exercise per week, considered simultaneously, to GPA (consider GPA the dependent or outcome variable).

CHAPTER 10

Nonparametric Procedures

In Chapter 10 of the textbook we presented hypothesis testing procedures for situations with small sample sizes and outcomes that are ordinal, ranked, or continuous and cannot be assumed to be normally distributed. Nonparametric tests are based on ranks that are assigned to the ordered data. The tests involve the same five steps as parametric tests: specifying the null hypothesis and alternative or research hypothesis, selecting and computing an appropriate test statistic, setting up a decision rule, and drawing a conclusion.

Four tests were presented in Chapter 10: the Mann–Whitney *U* Test for comparing a continuous outcome in two independent samples, the Sign and Signed Rank tests for comparing a continuous outcome in two matched or paired samples, and the Kruskal–Wallis test for comparing a continuous outcome in more than two independent samples.

For each test we used the same general approach, a five-step approach, which is outlined below:

- ***Step 1:*** Set up hypotheses (H_0 and H_1) and select a level of significance, α.
- ***Step 2:*** Choose the appropriate test statistic (e.g., *U*, the smaller of the number of positive or negative signs, or W or H).
- ***Step 3:*** Determine the critical value(s) and set up the decision rule (which depends on α, the test statistic, and whether the test is upper-, lower-, or two-tailed).
- ***Step 4:*** Compute the test statistic based on observed sample data.
- ***Step 5:*** Draw a conclusion by comparing the test statistic to the critical value.

The test statistic (Step 2) varies depending on the specific test. When we conducted nonparametric tests of hypothesis by hand in Chapter 10 of the textbook, we ultimately drew a conclusion by comparing the test statistic to the critical value, which was derived from an appropriate probability distribution table. Critical values can be found in Table 5 through Table 8 of the Appendix. We now use Microsoft Office® Excel® 2003 to compute the test statistics for each test. Before illustrating the computation of test statistics with Excel, we first illustrate how to use Excel to rank data, which is a key component of all nonparametric tests.

10.1 RANKING DATA

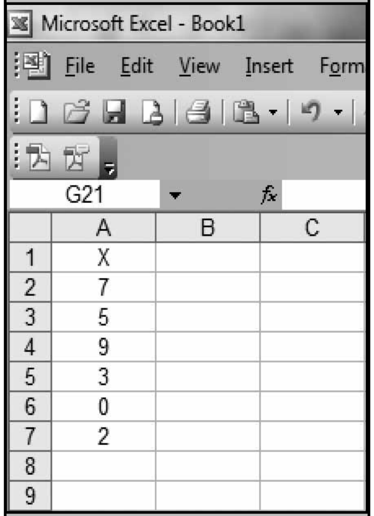
The nonparametric procedures that we describe here follow the same general procedure. The outcome variable (ordinal, interval, or continuous) is ranked from lowest to highest, and the analysis focuses on ranks as opposed to the measured or raw values. For example, suppose we measure self-reported pain using a visual analog scale with anchors at 0 (no pain) and 10 (agonizing pain) and record the following in a sample of six participants ($n = 6$) :

7 5 9 3 0 2

The data are entered into Excel as shown in Figure 10-1.

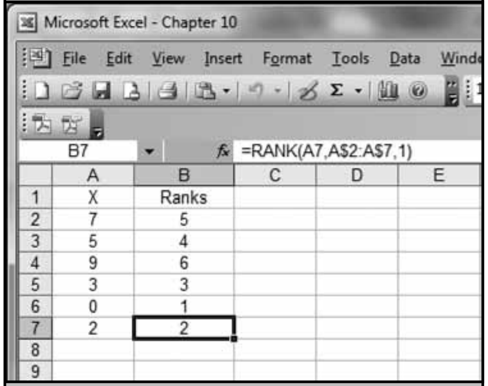
The ranks, which are used to perform a nonparametric test, are assigned to the data, which are ordered from smallest to largest. Notice that the data do not have to be ordered when they are entered into Excel. The smallest value is assigned a rank of 1, the next smallest is assigned a rank of 2, and so on.

FIGURE 10-1 Observed Data



	A	B	C
1	X		
2	7		
3	5		
4	9		
5	3		
6	0		
7	2		
8			
9			

FIGURE 10-2 Ranked Data


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Winde
B7
fx

=RAND(A7,A$2:A$7,1)

	A	B	C	D	E
1	X	Ranks			
2	7	5			
3	5	4			
4	9	6			
5	3	3			
6	0	1			
7	2	2			
8					
9					

The largest value is assigned a rank of n (in this example, $n = 6$). Excel has a built-in RANK function that assigns ranks to data. The function is used as follows:

=RANK(*value*, *data range*, *order*)

The *value* is the address of the cell that we wish to rank. The *data range* contains the addresses of the cells containing the first and last observations in the dataset, separated by a colon. For example, the data in Figure 10-1 occupy the range A2:A7. *Order* is specified as either 0 or 1. *Order* = 0 indicates that ranks are assigned in descending order, and *order* = 1 indicates that ranks are assigned in ascending order. Because we want to assign ranks from smallest to largest, we specify *order* = 1. To rank the data shown in Figure 10-1 in ascending order, we use the RANK function as follows: "=RANK(*value*,A$2:A$7,1)", where *value* changes depending on which data point we wish to rank (i.e., which cell in cell A2 through cell A7). Notice that we use absolute cell references to indicate the data range (A$2:A$7) to ensure that the same data range is considered as we copy the RANK function from cell to cell. The rank function is specified in cell B2 as "=RANK(A2,A$2:A$7,1)" and is copied from cell B2 to cell B3 through cell B7 as shown in Figure 10-2 (Notice the formula in cell B7 showing in the menu bar).

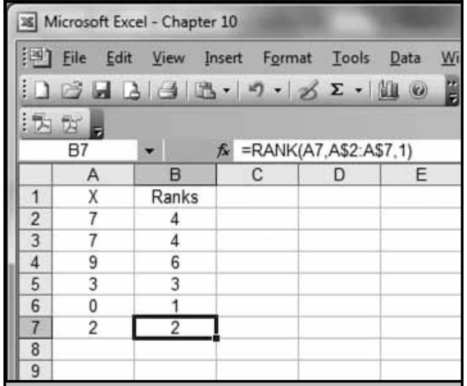
Using Excel, we rank the data in one step, as opposed to first ordering the data and then assigning ranks as we did when we assigned ranks by hand. The RANK function ranks the data in ascending or descending order. However, if there are ties, the same ranks are assigned to the tied values. For example, suppose the following are recorded in a sample of six participants ($n = 6$):

7 7 9 3 0 2

The data are entered into Excel as shown in Figure 10-3, and the rank function is again invoked as "=RANK(value,A$2:A$7,1)". The ranks are shown in Figure 10-3.

Notice that the 4th and 5th ordered values are both assigned ranks of 4. In nonparametric testing, we wish to assign

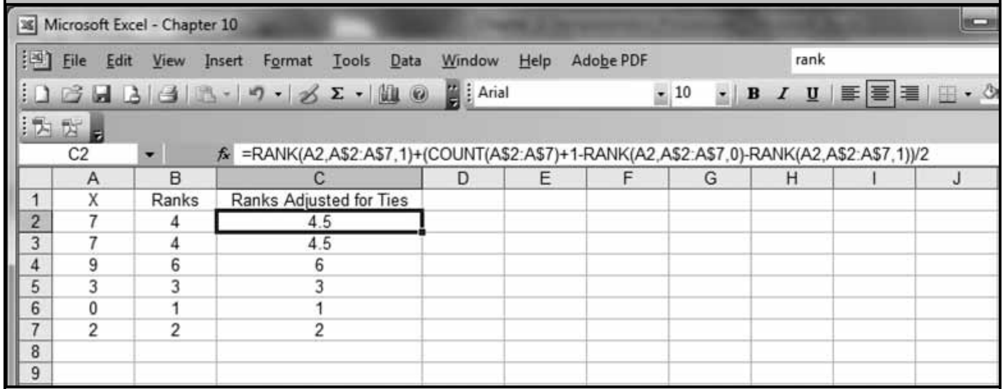
FIGURE 10-3 Ranks in Data with Ties


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Wi
B7
fx

=RANK(A7,A$2:A$7,1)

	A	B	C	D	E
1	X	Ranks			
2	7	4			
3	7	4			
4	9	6			
5	3	3			
6	0	1			
7	2	2			
8					
9					

FIGURE 10-4 Assigning Ranks Accounting for Ties


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
rank
Arial
10

C2 =RANK(A2,A$2:A$7,1)+(COUNT(A$2:A$7)+1-RANK(A2,A$2:A$7,0)-RANK(A2,A$2:A$7,1))/2

	A	B	C
1	X	Ranks	Ranks Adjusted for Ties
2	7	4	4.5
3	7	4	4.5
4	9	6	6
5	3	3	3
6	0	1	1
7	2	2	2
8			
9			

the mean rank to values that are tied. Specifically, we assign ranks of 4.5 to the two values of 7. To assign mean ranks to the tied values, we use the following as a correction. In each cell (see cell C2 in Figure 10-4), we enter "=RANK(A2,A$2:A$7,1)+(COUNT(A$2:A$7)+1−RANK(A2,A$2:A$7,0)−RANK(A2,A$2:A$7,1))/2." This formula is then copied from cell C2 to cell C3 through cell C7 as shown in Figure 10-4. (Notice the formula in cell C2 showing in the menu bar.)

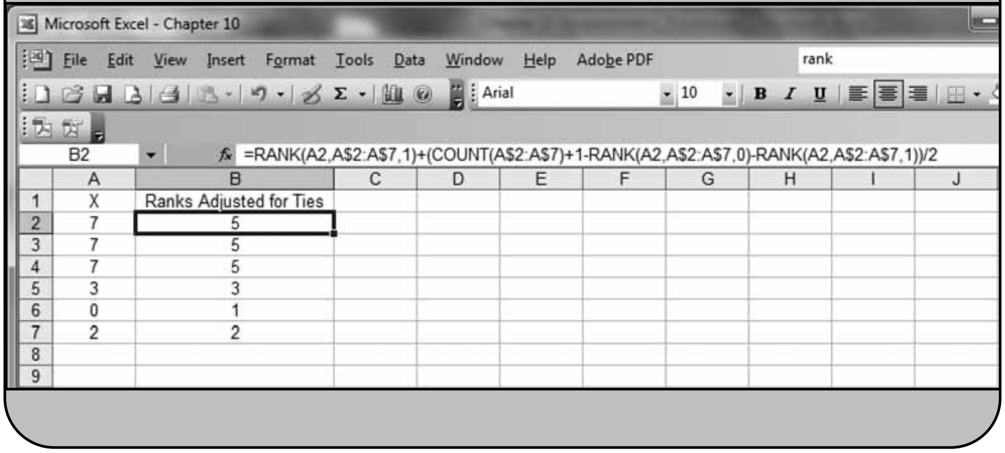
The corrected formula assigns the mean rank to tied values. As we work through applications, only the formula shown in column C is needed to assign ranks, as it makes the appropriate adjustment to assign mean ranks, when there are ties. Suppose the following are recorded in a sample of six participants ($n = 6$):

7 7 7 3 0 2

The data are entered into Excel as shown in Figure 10-5, and the rank function with the correction factor is specified as "=RANK(A2,A$2:A$7,1)+(COUNT(A$2:A$7)+1−RANK(A2,A$2:A$7,0)−RANK(A2,A$2:A$7,1))/2" in cell B2. This formula is then copied from cell B2 to cell B3 through cell B7 as shown in Figure 10-5.

Notice that there are three values of 7. We assign a rank of 5 (the mean of 4, 5, and 6) to the 4th, 5th, and 6th ordered

FIGURE 10-5 Ranking Data with Ties


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
rank
Arial
10

B2 =RANK(A2,A$2:A$7,1)+(COUNT(A$2:A$7)+1-RANK(A2,A$2:A$7,0)-RANK(A2,A$2:A$7,1))/2

	A	B
1	X	Ranks Adjusted for Ties
2	7	5
3	7	5
4	7	5
5	3	3
6	0	1
7	2	2
8		
9		

values. Using this approach of assigning the mean rank when there are ties ensures that the sum of the ranks is the same in each sample and always equal to $n(n + 1)/2$. When conducting nonparametric tests, it is useful to check the sum of the ranks before proceeding with the analysis.

10.2 TESTS WITH TWO INDEPENDENT SAMPLES

The Mann–Whitney *U* test is used to compare a continuous outcome in two independent samples with small sample sizes and outcomes that are ordinal, ranked, or continuous and cannot be assumed to be normally distributed. The hypotheses, with a two-sided alternative (as is generally the case) are as follows:

H_0: The two populations are equal.

H_1: The two populations are not equal.

In Chapter 10 of the textbook, we presented the following formula for the test statistic, *U*, in the Mann–Whitney *U* Test: *U* is the smaller of $U_1 = n_1n_2 + \frac{n_1(n_1+1)}{2} - R_1$ and $U_2 = n_1n_2 + \frac{n_2(n_2+1)}{2} - R_2$, where R_1 and R_2 are the sums of the ranks in groups 1 and 2, respectively.

When performing the test of hypothesis by hand, we computed the test statistic *U* and found the appropriate critical value in Table 5 in the Appendix to set up the decision rule: Reject H_0 if $U \leq$ critical value from Table 5. Excel does not have a specific analysis tool for the Mann–Whitney *U* test. However, Excel can be used to assign ranks and compute the test statistic. The conclusion of the test is based on a comparison of the test statistic to the appropriate critical value from Table 5 in the Appendix.

Example 10.1. In Example 10.1 in the textbook we analyzed data from a Phase II clinical trial designed to investigate the effectiveness of a new drug to reduce symptoms of asthma in children. A total of 10 participants ($n = 10$) were randomized to receive either the new drug or a placebo. Participants recorded the number of episodes of shortness of breath over a 1-week period following receipt of the assigned treatment. The data are shown below.

Placebo	7	5	6	4	12
New Drug	3	6	4	2	1

The question of interest is whether there is a difference in the number of episodes of shortness of breath over a 1-week period in participants receiving the new drug compared to those receiving the placebo. The hypotheses to be tested are given below and we run the test at the 5% level of significance (i.e., $\alpha = 0.05$).

H_0: The two populations are equal.

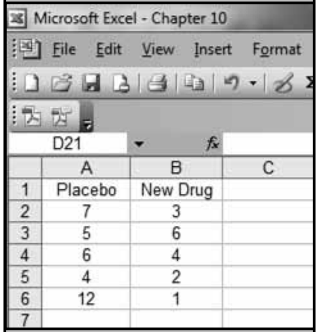
H_1: The two populations are not equal.

The data are entered into Excel as shown in Figure 10-6.

The first step is to assign ranks, and this is done on the combined or total sample (i.e., pooling the data from the two treatment groups), and assigning ranks from 1 to 10. Using Excel, we enter the data for each comparison group and then rank the pooled sample ($n = 10$). This is done using the procedure outlined in Section 10.1. Specifically, we use the rank function with the correction factor to assign ranks to the observed data in the placebo group and new-drug group, combined. Recall that we must maintain the group assignments. To perform the ranking, we create two new columns—one to contain the ranks of the data in the placebo group and the other to contain the ranks of the data in the new-drug group (See Figure 10-7). The first observation in the placebo group is in cell A2. To rank this value, in cell C2 we specify "=RANK(A2,$A$2:$B$6,1)+(COUNT($A$2:$B$6)+1−RANK(A2,$A$2:$B$6,0)−RANK(A2,$A$2:$B$6,1))/2". This formula is then copied from cell C2 to cell C3 through cell D6 as shown in Figure 10-7. Notice that we specify the range of the data as $A$2:$B$6. Specifically, we use absolute cell references for both the row and column addresses of the cells containing the data so that when the formula is copied we continue to rank the observed data, which is located in cell A2 through cell B6. See the formula for cell D6 showing in the menu bar.

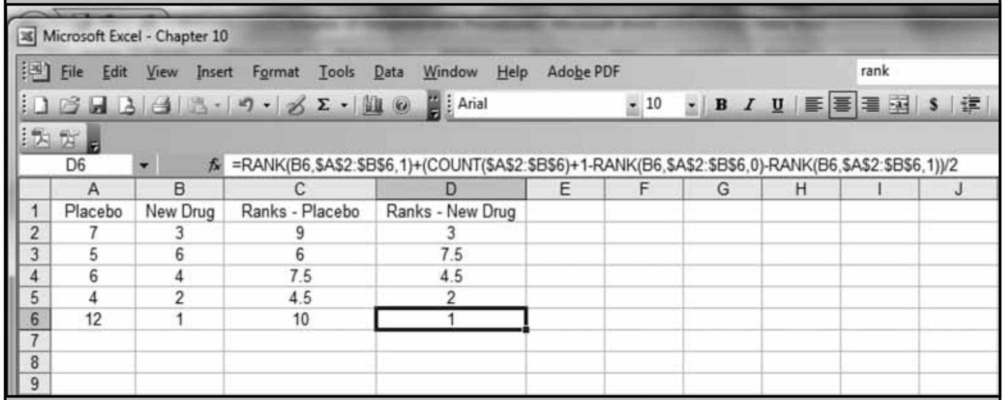
Before proceeding, we check the assignment of the ranks by summing the ranks in each group. These are denoted R_1 and R_2, for the placebo and new-drug groups, respectively.

FIGURE 10-6 Data for Mann–Whitney *U* Test



	A	B	C
1	Placebo	New Drug	
2	7	3	
3	5	6	
4	6	4	
5	4	2	
6	12	1	
7			

FIGURE 10-7 Ranking Data in the Pooled Sample


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
rank
Arial
10

D6 =RANK(B6,$A$2:$B$6,1)+(COUNT($A$2:$B$6)+1-RANK(B6,$A$2:$B$6,0)-RANK(B6,$A$2:$B$6,1))/2

	A	B	C	D
1	Placebo	New Drug	Ranks - Placebo	Ranks - New Drug
2	7	3	9	3
3	5	6	6	7.5
4	6	4	7.5	4.5
5	4	2	4.5	2
6	12	1	10	1

Recall that the sum of the ranks will always equal $n(n + 1)/2 = 10(11)/2 = 55$. This is shown in Figure 10-8.

Thus, $R_1 = 37$ and $R_2 = 18$. The test statistic for the Mann–Whitney *U* Test is denoted *U* and is the smaller of U_1 and U_2, defined below.

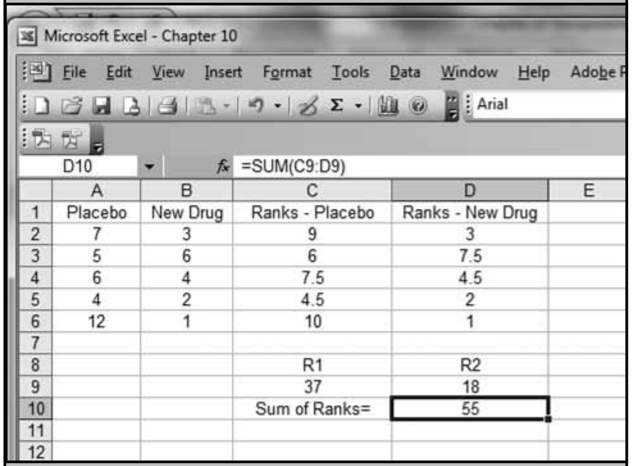
$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

To compute the test statistic in Excel, we first compute the sample sizes, n_1 and n_2, using the COUNT function. We specify the cells we wish to count as follows: "=COUNT(A2:A6)". The totals the number of cells in the range of A2 through A6 that contain numeric data (numbers). The computation of sample sizes is shown in Figure 10-9.

In this example, the sample size of group 1 (n_1) is in cell C13 and is determined by "=COUNT(A2:A6)". The sample size of group 2 (n_2) is in cell D13 and is determined by "=COUNT(B2:B6)". We now use R_1, R_2, n_1, and n_2 to com-

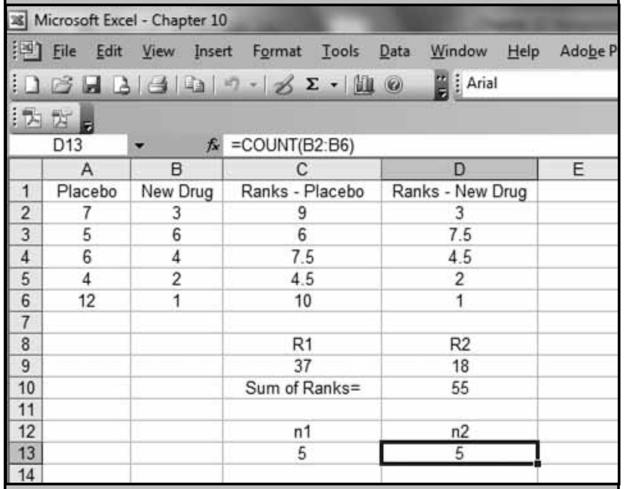
FIGURE 10-8 Summing the Ranks in Each Group


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe P
Arial

D10 =SUM(C9:D9)

	A	B	C	D
1	Placebo	New Drug	Ranks - Placebo	Ranks - New Drug
2	7	3	9	3
3	5	6	6	7.5
4	6	4	7.5	4.5
5	4	2	4.5	2
6	12	1	10	1
7				
8			R1	R2
9			37	18
10			Sum of Ranks=	55

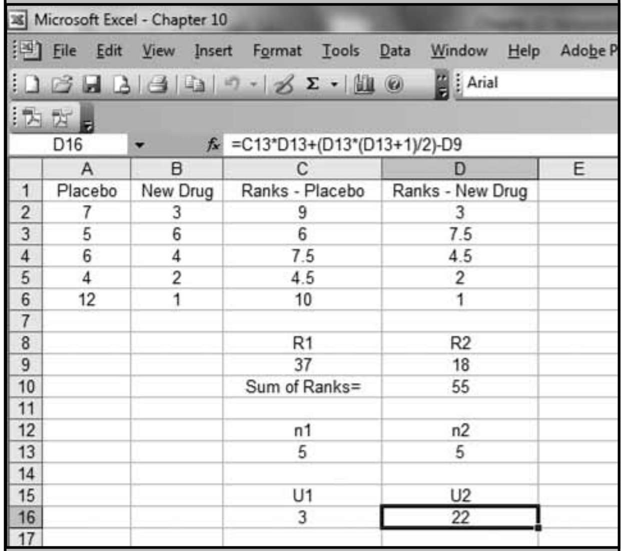
FIGURE 10-9 Determining Sample Sizes


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe P
Arial

D13 =COUNT(B2:B6)

	A	B	C	D
1	Placebo	New Drug	Ranks - Placebo	Ranks - New Drug
2	7	3	9	3
3	5	6	6	7.5
4	6	4	7.5	4.5
5	4	2	4.5	2
6	12	1	10	1
7				
8			R1	R2
9			37	18
10			Sum of Ranks=	55
11				
12			n1	n2
13			5	5

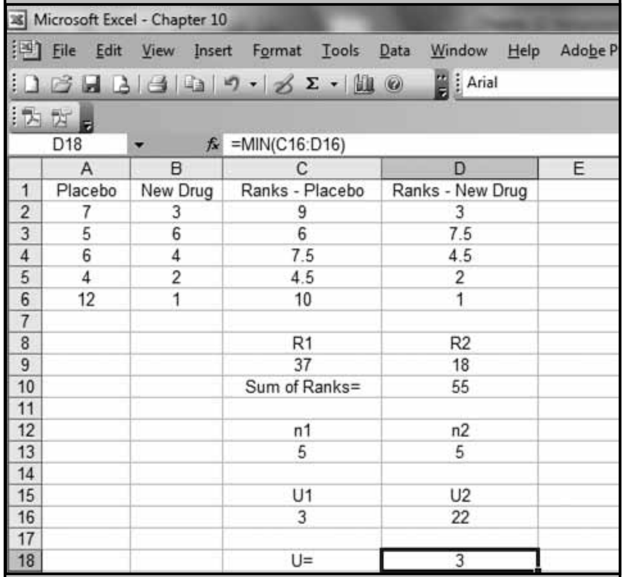
FIGURE 10-10 Computing U_1 and U_2



D16 =C13*D13+(D13*(D13+1)/2)-D9

	A	B	C	D	E
1	Placebo	New Drug	Ranks - Placebo	Ranks - New Drug	
2	7	3	9	3	
3	5	6	6	7.5	
4	6	4	7.5	4.5	
5	4	2	4.5	2	
6	12	1	10	1	
7					
8			R1	R2	
9			37	18	
10			Sum of Ranks=	55	
11					
12			n1	n2	
13			5	5	
14					
15			U1	U2	
16			3	22	
17					

FIGURE 10-11 Computing the Test Statistic U



D18 =MIN(C16:D16)

	A	B	C	D	E
1	Placebo	New Drug	Ranks - Placebo	Ranks - New Drug	
2	7	3	9	3	
3	5	6	6	7.5	
4	6	4	7.5	4.5	
5	4	2	4.5	2	
6	12	1	10	1	
7					
8			R1	R2	
9			37	18	
10			Sum of Ranks=	55	
11					
12			n1	n2	
13			5	5	
14					
15			U1	U2	
16			3	22	
17					
18			U=	3	

pute U_1 and U_2 using Excel. The computations are shown in Figure 10-10. The formula used to compute U_2 in cell D16 is shown in the top menu bar.

The final step is to produce the test statistic, U, which is the smaller of U_1 and U_2. This is computed using the MIN (minimum) function and shown in Figure 10-11 .

Thus, $U = 3$. To draw a final conclusion in the test, we must determine whether the observed test statistic, U, supports the null or the research hypothesis. This is done by determining a critical value of U such that if the observed value of U is less than or equal to the critical value, we reject H_0 in favor of H_1, and if the observed value of U exceeds the critical value we do not reject H_0. The critical value of U is found in Table 5 in the Appendix. For $n_1 = n_2 = 5$ and a two-sided level of significance $\alpha = 0.05$, the critical value is 2, and the decision rule is to reject H_0 if $U \leq 2$. We do not reject H_0 because $3 > 2$. We do not have statistically significant evidence at $\alpha = 0.05$ to show that the two populations of numbers of episodes of shortness of breath are not equal.

10.3 TESTS WITH MATCHED SAMPLES

The Sign test and the Wilcoxon Signed Rank test are used to compare a continuous outcome in two matched or paired samples with small sample sizes and outcomes that are ordinal, ranked, or continuous and cannot be assumed to be normally distributed. Recall that when data are matched or paired, we compute difference scores for each individual and analyze difference scores. The hypotheses for both tests, with a two-sided alternative (as is generally the case), are as follows:

H_0: The median difference is zero.

H_1: The median difference is not zero.

In Chapter 10 of the textbook, we presented the Sign test, where the test statistic is the smaller of the number of positive or negative signs of the difference scores. When performing the test of hypothesis by hand, we found the appropriate critical value in Table 6 in the Appendix to set up the decision rule: Reject H_0 if the smaller of the number of positive or negative signs $\leq$ critical value from Table 6. Excel does not have a specific analysis tool for the Sign test. However, Excel can be used to compute difference scores, to count the numbers of positive and negative signs, and to determine the smaller of the two, which is the test statistic. The conclusion of the test is based on a comparison of the test statistic to the appropriate critical value from Table 6 in the Appendix.

In Chapter 10 of the textbook, we also presented the Wilcoxon Signed Rank test and the test statistic W, which is defined as the smaller of $W+$ and $W-$, which are the sums of the positive and negative ranks of the difference scores, respectively. When performing the test of hypothesis by hand, we found the appropriate critical value in Table 7 in the Appendix to set up the decision rule: Reject H_0 if $W \leq$ critical value from Table 7. Excel does not have a specific analysis tool for the Wilcoxon Signed Rank test. However, Excel can be used to compute difference scores, to rank the differences, to attach the signs, to determine $W+$ and $W-$, and to compute the test statistic W. The conclusion of the test is based on a comparison of the test statistic to the appropriate critical value from Table 7 in the Appendix.

Example 10.2. In Example 10.5 in the textbook, we analyzed data from a study to assess quality of life (QOL) in patients with breast cancer following a new chemotherapy treatment. QOL was measured on an ordinal scale, and for analysis purposes, numbers were assigned to each response category as follows: 1 = Poor, 2 = Fair, 3 = Good, 4 = Very Good, 5 = Excellent. The data are shown in Table 10-1.

The question of interest is whether there is a difference in QOL after chemotherapy treatment as compared to before. The test is run at a 5% level of significance and the hypotheses are as follows:

$$H_0\text{: The median difference is zero.}$$

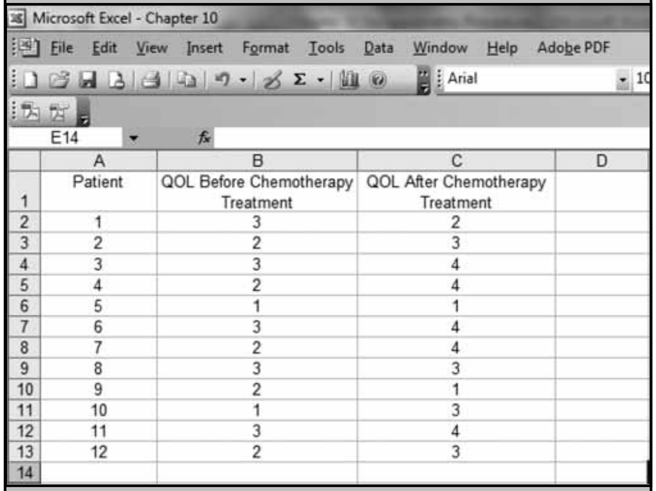
$$H_1\text{: The median difference is not zero.}$$

The data are entered into Excel as shown in Figure 10-12.

We analyze the data using the Sign test. The test statistic for the Sign test is the smaller of the number of positive or negative difference scores. We first compute difference scores by subtracting the QOL measured before treatment from that measured after (i.e., the measurement in column C minus the measurement in column B for each participant). The difference scores are shown in Figure 10-13.

With the Sign test, we only concern ourselves with the signs of the difference scores. Thus, we take each difference score in column D and retain only the sign of the difference score. This is done in Excel using the IF function. Specifically,

FIGURE 10-12 Data on Matched Pairs for Sign Test

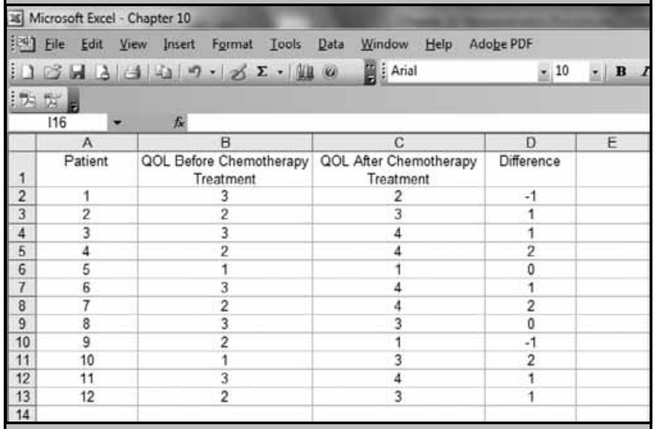

Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
E14

	A	B	C	D
1	Patient	QOL Before Chemotherapy Treatment	QOL After Chemotherapy Treatment	
2	1	3	2	
3	2	2	3	
4	3	3	4	
5	4	2	4	
6	5	1	1	
7	6	3	4	
8	7	2	4	
9	8	3	3	
10	9	2	1	
11	10	1	3	
12	11	3	4	
13	12	2	3	
14				

TABLE 10-1 QOL Before and After Chemotherapy Treatment

Patient	QOL Before Chemotherapy Treatment	QOL After Chemotherapy Treatment
1	3	2
2	2	3
3	3	4
4	2	4
5	1	1
6	3	4
7	2	4
8	3	3
9	2	1
10	1	3
11	3	4
12	2	3

FIGURE 10-13 Difference Scores


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B
I16

	A	B	C	D	E
1	Patient	QOL Before Chemotherapy Treatment	QOL After Chemotherapy Treatment	Difference	
2	1	3	2	-1	
3	2	2	3	1	
4	3	3	4	1	
5	4	2	4	2	
6	5	1	1	0	
7	6	3	4	1	
8	7	2	4	2	
9	8	3	3	0	
10	9	2	1	-1	
11	10	1	3	2	
12	11	3	4	1	
13	12	2	3	1	
14					

we evaluate each difference score in column D (in cell D2 through cell D13) and retain only the sign of the difference by using the IF function as follows: IF(*condition*, *result if condition is true*, *result if condition is false*). For example, we specify the following in cell E2: “=IF(D2<0,“−”,“ ”)”. The condition determines whether the value in cell D2 is negative (i.e., <0), and if so the IF function places a negative sign “−” in cell E2; if not, cell E2 is left blank, “ ”. We actually wish to assign a positive sign “+” to a cell with a positive difference and a negative sign “−” to a cell with a negative difference, and so to do this, we specify the following in cell E2: “=IF(D2<0,“−”, IF(D2>0,“+”,“ ”))”. This IF expression first determines whether the value in cell D2 is negative and if so, a negative sign “−” is placed in cell E2. If not, the expression then determines whether the value in cell D2 is positive, and if so, places a positive sign “+” in cell E2. If not, cell E2 is left blank. This is entered into cell E2 and is copied into cell E3 through cell E13 as shown in Figure 10-14.

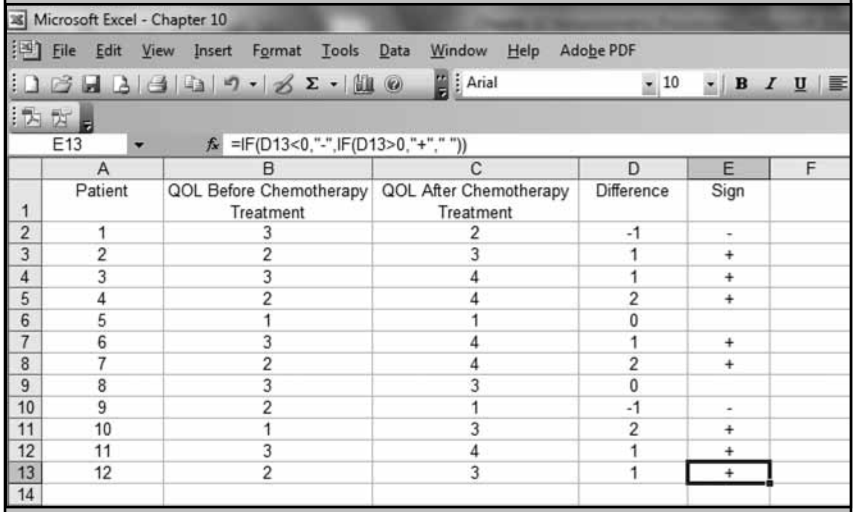
Notice that cells E6 and E9 are blank because the differences are neither positive nor negative, but are equal to zero. Recall that if there is an even number of zeros in a dataset, we randomly assign positive and negative signs to them. In this example, we assign one negative sign (i.e., “−” to patient 5) and one positive sign (i.e., “+” to patient 8). The signs are entered directly into cells E6 and E9 as shown in Figure 10-15.

Next, we count the number of positive and negative signs using the COUNTIF function. The COUNTIF function counts the number of cells in a specified range that meet a certain criterion. In this example the number of positive signs is computed as “=COUNTIF(E2:E13,“+”)” and the number of negative signs is computed as “=COUNTIF(E2:E13,“−”)” as shown in Figure 10-16.

The test statistic is the smaller of the number of positive or negative signs of the difference scores, which is equal to 3. The appropriate critical value for the Sign test is found in Table 6 in the Appendix based on the sample size (or number of matched pairs, $n = 12$), and our two-sided level of significance ($\alpha = 0.05$). The critical value for this two-sided test with $n = 12$ and $\alpha = 0.05$ is 2, and the decision rule is as follows: Reject H_0 if the smaller of the number of positive or negative signs ≤ 2. We do not reject H_0 because $3 > 2$. We do not have statistically significant evidence at $\alpha = 0.05$ to show that there is a difference in QOL after chemotherapy treatment compared to before treatment.

With the Sign test it is possible to compute a *p*-value for the test using the binomial distribution. In Chapter 7 of the Excel workbook, we computed *p*-values for the parametric tests analyzing means, proportions, differences in means and proportions, and mean differences using the *Z* and *t* probability distributions. We use the same approach here. Specifically, we use Excel to compute the test statistic (with the Sign test, the smaller of the number of positive or negative signs) and a *p*-value, and the investigator then compares the *p*-value to the predetermined level of significance to draw a conclusion about the hypotheses using the following rule: Reject H_0 if $p \leq \alpha$.

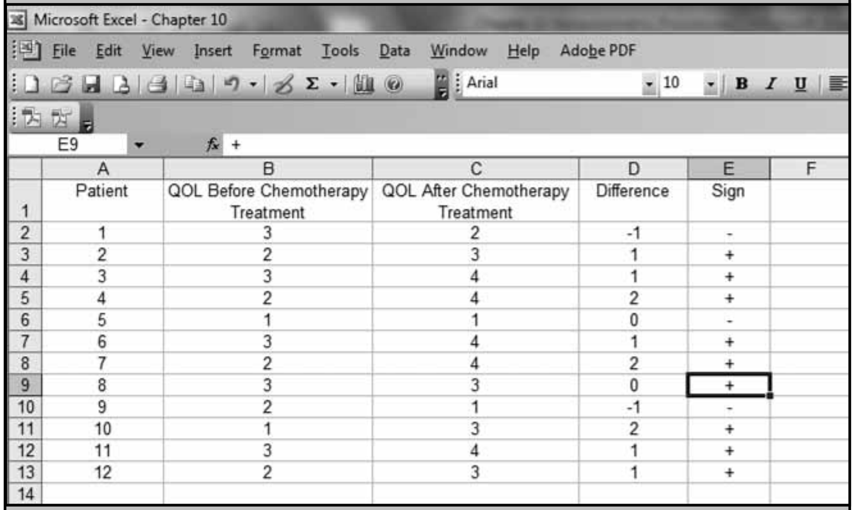
FIGURE 10-14 Signs of the Difference Scores


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B I U
E13

=IF(D13<0,"-",IF(D13>0,"+"," "))

	A	B	C	D	E	F
1	Patient	QOL Before Chemotherapy Treatment	QOL After Chemotherapy Treatment	Difference	Sign	
2	1	3	2	-1	-	
3	2	2	3	1	+	
4	3	3	4	1	+	
5	4	2	4	2	+	
6	5	1	1	0		
7	6	3	4	1	+	
8	7	2	4	2	+	
9	8	3	3	0		
10	9	2	1	-1	-	
11	10	1	3	2	+	
12	11	3	4	1	+	
13	12	2	3	1	+	
14						

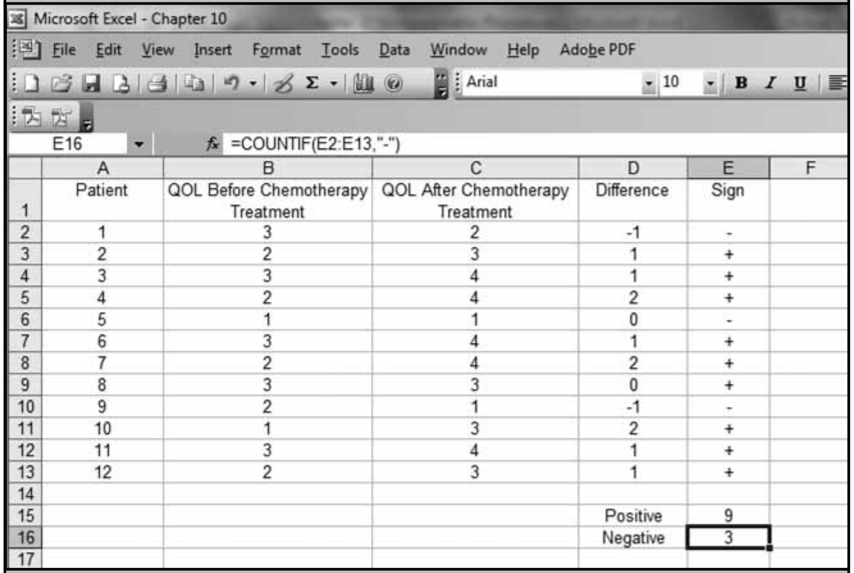
FIGURE 10-15 Assigning Signs to Differences of Zero


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B I U

E9 f_x +

	A	B	C	D	E	F
1	Patient	QOL Before Chemotherapy Treatment	QOL After Chemotherapy Treatment	Difference	Sign	
2	1	3	2	-1	-	
3	2	2	3	1	+	
4	3	3	4	1	+	
5	4	2	4	2	+	
6	5	1	1	0	-	
7	6	3	4	1	+	
8	7	2	4	2	+	
9	8	3	3	0	+	
10	9	2	1	-1	-	
11	10	1	3	2	+	
12	11	3	4	1	+	
13	12	2	3	1	+	
14						

FIGURE 10-16 Summing the Numbers of Positive and Negative Signs


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B I U

E16 f_x =COUNTIF(E2:E13,"-")

	A	B	C	D	E	F
1	Patient	QOL Before Chemotherapy Treatment	QOL After Chemotherapy Treatment	Difference	Sign	
2	1	3	2	-1	-	
3	2	2	3	1	+	
4	3	3	4	1	+	
5	4	2	4	2	+	
6	5	1	1	0	-	
7	6	3	4	1	+	
8	7	2	4	2	+	
9	8	3	3	0	+	
10	9	2	1	-1	-	
11	10	1	3	2	+	
12	11	3	4	1	+	
13	12	2	3	1	+	
14						
15				Positive	9	
16				Negative	3	
17						

The test statistic for the Sign test is the smaller of the number of positive or negative signs of the difference scores, and it follows a binomial distribution with n = the number of subjects in the study and $p = 0.5$. In this example, $n = 12$ and $p = 0.5$. The two-sided p-value for the test is $p = 2 \times P(x \leq 3)$. In Chapter 5 of the Excel workbook, we used the BINOMDIST function to compute probabilities from the binomial distribution. The BINOMDIST function is specified as follows: "=BINOMDIST(*x*,*n*,*p*,*cumulative*)". The first three inputs for the function are the same as those we use in computing probabilities by hand with the binomial distribution model (i.e., x, n, and p). Excel requires one additional input, labeled *cumulative*. This last argument in the BINOMDIST function is a logical value (i.e., one whose responses are true or false). We use the cumulative distribution function by specifying "true" or do not use the cumulative distribution function by specifying "false." The cumulative distribution function returns the probability of observing x or fewer successes. For example, if we specify "true" and indicate $x = 3$ in the function, then Excel computes $P(X \leq 3)$. In contrast, if we specify "false" and indicate $x = 3$ in the function, then Excel computes $P(X = 3)$. We wish to compute $P(X \leq 3)$ with $n = 12$ and $p = 0.5$. The computation is shown in Figure 10-17. Notice that we indicate the number of successes (x) as the smaller of E15 and E16, and n (the sample size for analysis) is determined using the COUNTIF function to count the number of positive and negative signs (see the formula in the top menu bar).

Because *p-value* = 0.1460 exceeds the level of significance ($\alpha = 0.05$) we do not have statistically significant evidence at $\alpha = 0.05$ to show that there is a difference in QOL after chemotherapy treatment compared to before treatment.

Another popular nonparametric test for matched or paired data is called the Wilcoxon Signed Rank test. Like the Sign test, it is based on difference scores, but in addition to analyzing the signs of the differences, it also takes into account the magnitude of the observed differences.

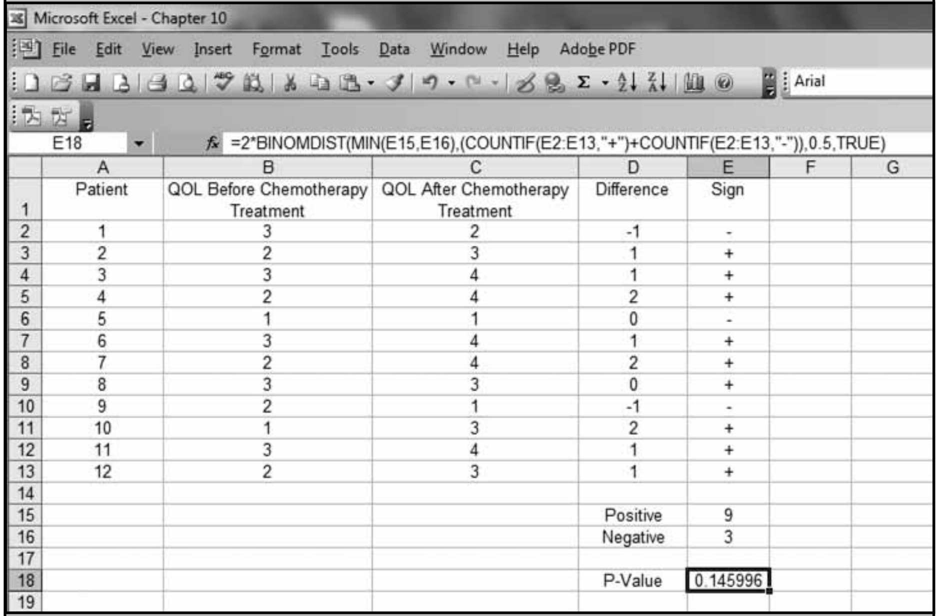
Example 10.3. In Example 10.7 in the textbook, we analyzed data from a study to evaluate the effectiveness of an exercise program in reducing systolic blood pressure in patients with prehypertension (defined as a systolic blood pressure between 120 mmHg and 139 mmHg or a diastolic blood pressure between 80 mmHg and 89 mmHg). A total of 15 patients with prehypertension enrolled in the study, and their systolic blood pressures were measured. Each patient then participated in an exercise training program where they learned proper techniques and execution of a series of exercises. Patients were instructed to do the exercise program 3 times per week for 6 weeks. After 6 weeks, systolic blood pressures were again measured and are shown in Table 10-2.

The question of interest is whether there is a difference in systolic blood pressures after participating in the exercise program compared to before the exercise program. The test is run at a 5% level of significance and the hypotheses are below:

H_0: The median difference is zero.

H_1: The median difference is not zero.

FIGURE 10-17 Computing the *p*-Value for the Sign Test



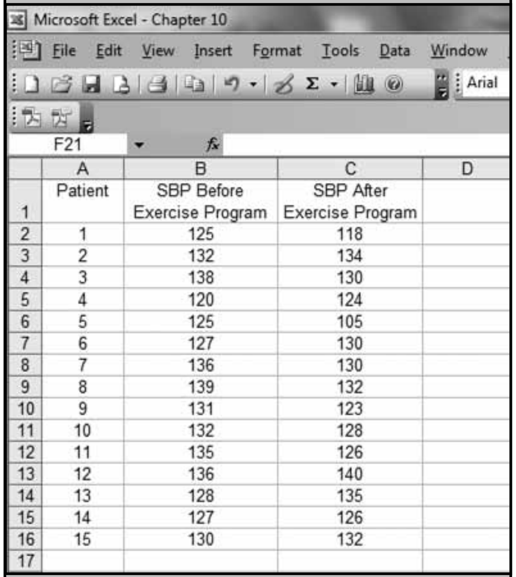
E18 =2*BINOMDIST(MIN(E15,E16),(COUNTIF(E2:E13,"+")+COUNTIF(E2:E13,"-")),0.5,TRUE)

	A	B	C	D	E	F	G
1	Patient	QOL Before Chemotherapy Treatment	QOL After Chemotherapy Treatment	Difference	Sign		
2	1	3	2	-1	-		
3	2	2	3	1	+		
4	3	3	4	1	+		
5	4	2	4	2	+		
6	5	1	1	0	-		
7	6	3	4	1	+		
8	7	2	4	2	+		
9	8	3	3	0	+		
10	9	2	1	-1	-		
11	10	1	3	2	+		
12	11	3	4	1	+		
13	12	2	3	1	+		
14							
15				Positive	9		
16				Negative	3		
17							
18				P-Value	0.145996		
19							

TABLE 10-2 Blood Pressure Before and After Exercise Program

Patient	Systolic Blood Pressure Before Exercise Program	Systolic Blood Pressure After Exercise Program
1	125	118
2	132	134
3	138	130
4	120	124
5	125	105
6	127	130
7	136	130
8	139	132
9	131	123
10	132	128
11	135	126
12	136	140
13	128	135
14	127	126
15	130	132

FIGURE 10-18 Data on Matched Pairs for Wilcoxon Signed Rank Test



	A	B	C	D
1	Patient	SBP Before Exercise Program	SBP After Exercise Program	
2	1	125	118	
3	2	132	134	
4	3	138	130	
5	4	120	124	
6	5	125	105	
7	6	127	130	
8	7	136	130	
9	8	139	132	
10	9	131	123	
11	10	132	128	
12	11	135	126	
13	12	136	140	
14	13	128	135	
15	14	127	126	
16	15	130	132	
17				

The data are entered into Excel as shown in Figure 10-18.

We analyze the data using the Wilcoxon Signed Rank test. The test statistic is W, the smaller of $W+$ and $W-$, which are the sums of the positive and negative ranks, respectively.

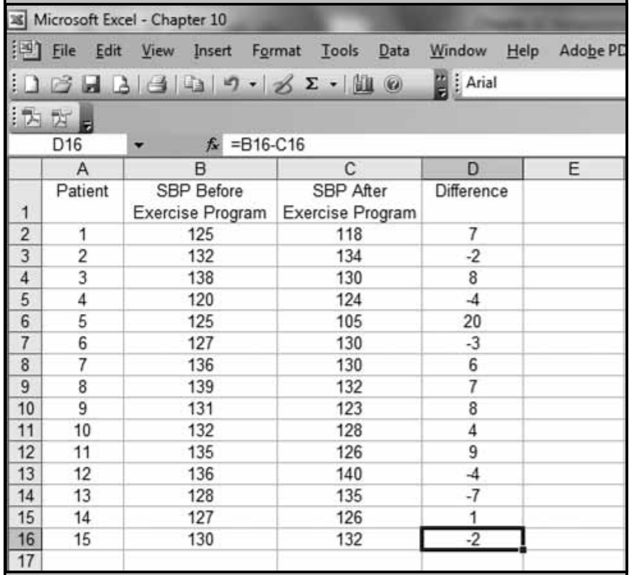
We first compute difference scores by subtracting the SBP measured after the exercise program from that measured before the exercise program. The difference scores are shown in Figure 10-19.

The next step is to rank the ordered absolute values of the difference scores. First, we generate a column of the absolute values of the difference scores using the ABS (absolute value) function. The absolute values of the difference scores are shown in column E in Figure 10-20.

Next we assign ranks from 1 through n to the smallest through the largest absolute values of the difference scores, respectively, and assign the mean rank when there are ties in the absolute values of the difference scores using the approach outlined in Section 10.1. The ranks of the absolute values of the difference scores are shown in Figure 10-21.

In the next step, we attach the signs ("+" or "−") of the observed differences to each rank using the IF function as shown in Figure 10-22. Specifically, if the difference (value in column D) is less than zero, the signed rank is equal to −1 × the rank (in column F). If not, the signed rank is equal to the rank in column F. Notice that for patient 2 (row 3) the difference is −2, and thus the signed rank is −2.5.

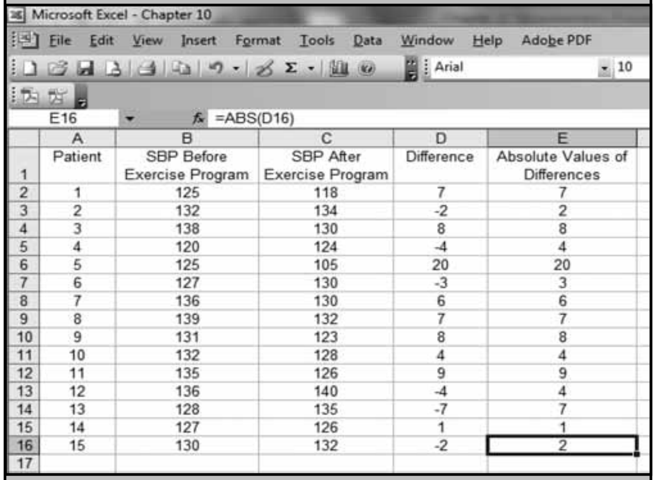
FIGURE 10-19 Difference Scores



=B16-C16

	A	B	C	D	E
1	Patient	SBP Before Exercise Program	SBP After Exercise Program	Difference	
2	1	125	118	7	
3	2	132	134	-2	
4	3	138	130	8	
5	4	120	124	-4	
6	5	125	105	20	
7	6	127	130	-3	
8	7	136	130	6	
9	8	139	132	7	
10	9	131	123	8	
11	10	132	128	4	
12	11	135	126	9	
13	12	136	140	-4	
14	13	128	135	-7	
15	14	127	126	1	
16	15	130	132	-2	
17					

FIGURE 10-20 Absolute Values of the Difference Scores


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10

E16 =ABS(D16)

	A	B	C	D	E
1	Patient	SBP Before Exercise Program	SBP After Exercise Program	Difference	Absolute Values of Differences
2	1	125	118	7	7
3	2	132	134	-2	2
4	3	138	130	8	8
5	4	120	124	-4	4
6	5	125	105	20	20
7	6	127	130	-3	3
8	7	136	130	6	6
9	8	139	132	7	7
10	9	131	123	8	8
11	10	132	128	4	4
12	11	135	126	9	9
13	12	136	140	-4	4
14	13	128	135	-7	7
15	14	127	126	1	1
16	15	130	132	-2	2
17					

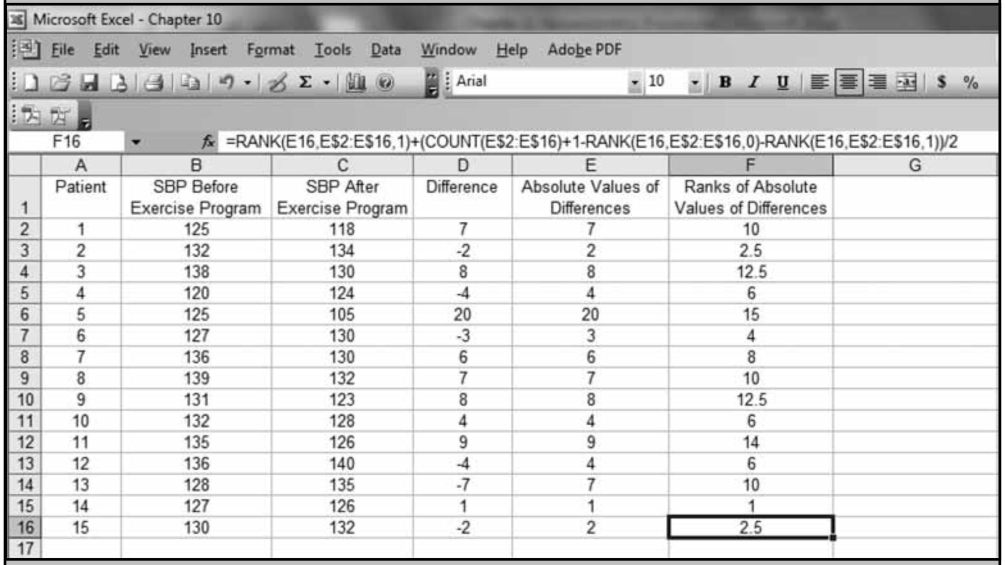
We now compute the sums of the positive and negative ranks, $W+$ and $W-$, respectively using the SUMIF function. The SUMIF function sums the values in a specified range of cells (e.g., cell G2 through cell G16) that meet a specified criterion. $W+$ is the sum of the positive ranks, computed as "=SUMIF(G2:G16,">0")". $W-$ is the sum of the negative ranks, computed as "=ABS(SUMIF(G2:G16,"<0"))". Notice that when we sum the negative ranks, we want to sum the absolute values of the negative ranks, which is done using the absolute value (ABS) function. This is shown in Figure 10-23.

The test statistic is W, the smaller of $W+$ and $W-$. $W = 31$ as shown in Figure 10-24. The critical value of W is found in Table 7 in the Appendix based on the sample size ($n = 15$) and our two-sided level of significance ($\alpha = 0.05$). The critical value for this two-sided test with $n = 15$ and $\alpha = 0.05$ is 25, and the decision rule is as follows: Reject H_0 if $W \leq 25$. We do not reject H_0 because $31 > 25$. We do not have statistically significant evidence at $\alpha = 0.05$ to show that the median difference in systolic blood pressures is not zero (i.e., that there is a significant difference in systolic blood pressures after the exercise program as compared to before).

10.4 TESTS WITH MORE THAN TWO INDEPENDENT SAMPLES

The Kruskal–Wallis test is used to compare medians of a continuous outcome in more than two independent samples with small sample sizes when the outcome is ordinal, ranked, or continuous and cannot be assumed to be normally distributed. The Kruskal–Wallis test is used to compare medians among k comparison groups ($k > 2$) and is sometimes de-

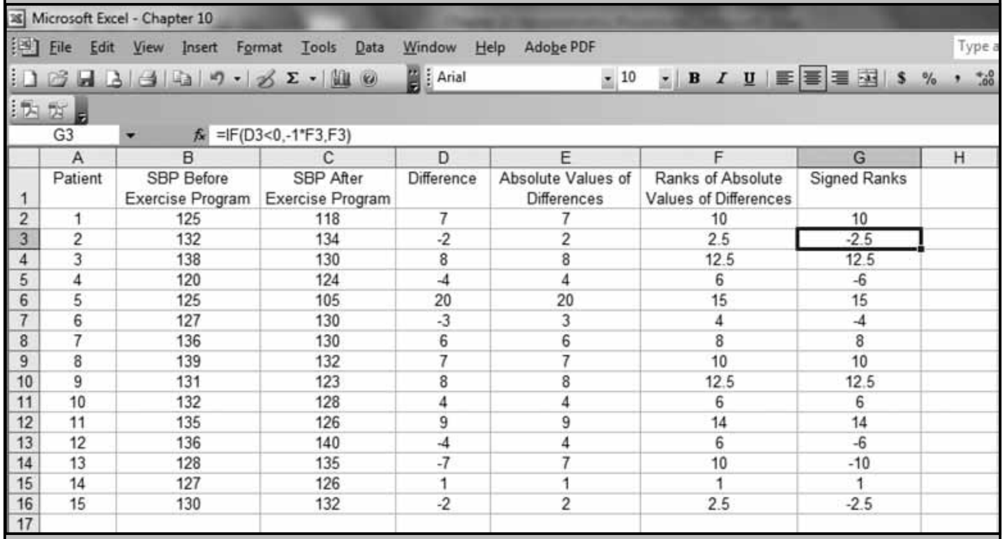
FIGURE 10-21 Ranking the Absolute Values of the Difference Scores


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B I U
$ %

F16 =RANK(E16,E$2:E$16,1)+(COUNT(E$2:E$16)+1-RANK(E16,E$2:E$16,0)-RANK(E16,E$2:E$16,1))/2

	A	B	C	D	E	F	G
1	Patient	SBP Before Exercise Program	SBP After Exercise Program	Difference	Absolute Values of Differences	Ranks of Absolute Values of Differences	
2	1	125	118	7	7	10	
3	2	132	134	-2	2	2.5	
4	3	138	130	8	8	12.5	
5	4	120	124	-4	4	6	
6	5	125	105	20	20	15	
7	6	127	130	-3	3	4	
8	7	136	130	6	6	8	
9	8	139	132	7	7	10	
10	9	131	123	8	8	12.5	
11	10	132	128	4	4	6	
12	11	135	126	9	9	14	
13	12	136	140	-4	4	6	
14	13	128	135	-7	7	10	
15	14	127	126	1	1	1	
16	15	130	132	-2	2	2.5	
17							

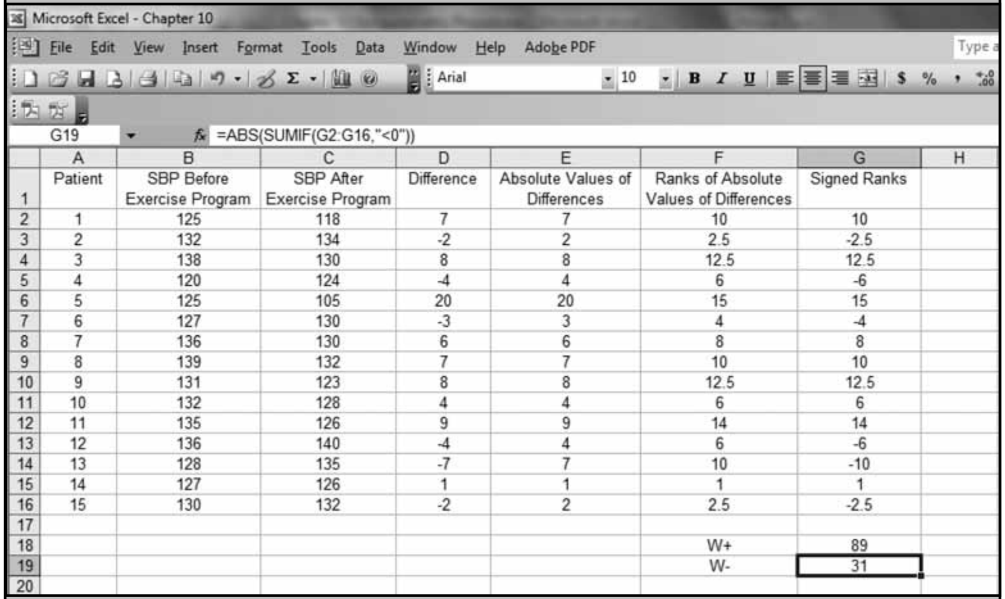
FIGURE 10-22 Signed Ranks


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Type a
Arial
10
B I U
$ %

G3 =IF(D3<0,-1*F3,F3)

	A	B	C	D	E	F	G	H
1	Patient	SBP Before Exercise Program	SBP After Exercise Program	Difference	Absolute Values of Differences	Ranks of Absolute Values of Differences	Signed Ranks	
2	1	125	118	7	7	10	10	
3	2	132	134	-2	2	2.5	-2.5	
4	3	138	130	8	8	12.5	12.5	
5	4	120	124	-4	4	6	-6	
6	5	125	105	20	20	15	15	
7	6	127	130	-3	3	4	-4	
8	7	136	130	6	6	8	8	
9	8	139	132	7	7	10	10	
10	9	131	123	8	8	12.5	12.5	
11	10	132	128	4	4	6	6	
12	11	135	126	9	9	14	14	
13	12	136	140	-4	4	6	-6	
14	13	128	135	-7	7	10	-10	
15	14	127	126	1	1	1	1	
16	15	130	132	-2	2	2.5	-2.5	
17								

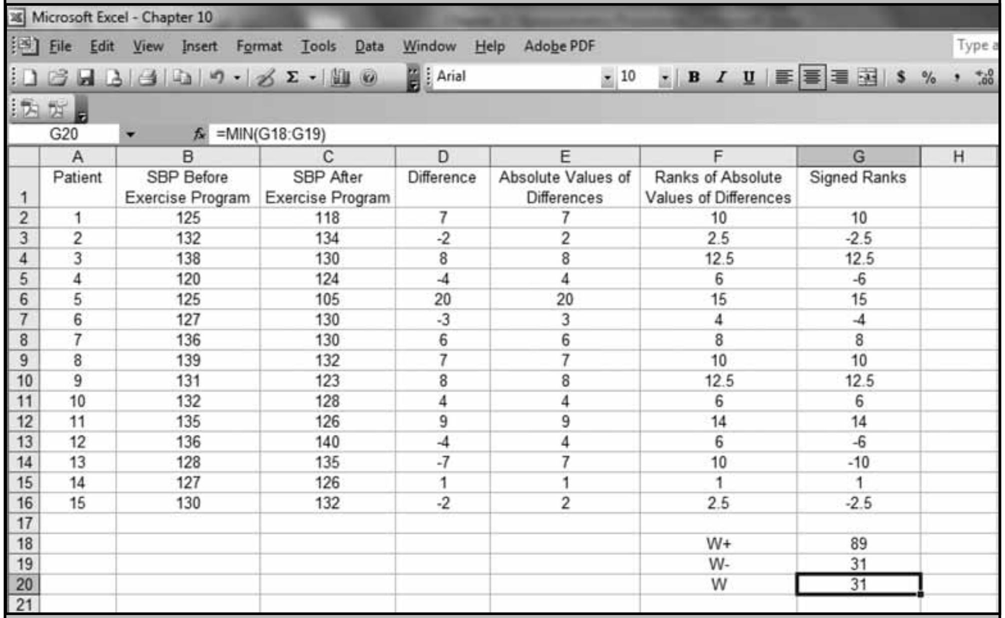
FIGURE 10-23 Computing *W*+ and *W*-


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Type a
Arial
10
B I U
$ %

G19 =ABS(SUMIF(G2:G16,"<0"))

	A	B	C	D	E	F	G	H
1	Patient	SBP Before Exercise Program	SBP After Exercise Program	Difference	Absolute Values of Differences	Ranks of Absolute Values of Differences	Signed Ranks	
2	1	125	118	7	7	10	10	
3	2	132	134	-2	2	2.5	-2.5	
4	3	138	130	8	8	12.5	12.5	
5	4	120	124	-4	4	6	-6	
6	5	125	105	20	20	15	15	
7	6	127	130	-3	3	4	-4	
8	7	136	130	6	6	8	8	
9	8	139	132	7	7	10	10	
10	9	131	123	8	8	12.5	12.5	
11	10	132	128	4	4	6	6	
12	11	135	126	9	9	14	14	
13	12	136	140	-4	4	6	-6	
14	13	128	135	-7	7	10	-10	
15	14	127	126	1	1	1	1	
16	15	130	132	-2	2	2.5	-2.5	
17								
18						W+	89	
19						W-	31	
20								

FIGURE 10-24 Computing the Test Statistic *W*


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Type a
Arial
10
B I U
$ %

G20 =MIN(G18:G19)

	A	B	C	D	E	F	G	H
1	Patient	SBP Before Exercise Program	SBP After Exercise Program	Difference	Absolute Values of Differences	Ranks of Absolute Values of Differences	Signed Ranks	
2	1	125	118	7	7	10	10	
3	2	132	134	-2	2	2.5	-2.5	
4	3	138	130	8	8	12.5	12.5	
5	4	120	124	-4	4	6	-6	
6	5	125	105	20	20	15	15	
7	6	127	130	-3	3	4	-4	
8	7	136	130	6	6	8	8	
9	8	139	132	7	7	10	10	
10	9	131	123	8	8	12.5	12.5	
11	10	132	128	4	4	6	6	
12	11	135	126	9	9	14	14	
13	12	136	140	-4	4	6	-6	
14	13	128	135	-7	7	10	-10	
15	14	127	126	1	1	1	1	
16	15	130	132	-2	2	2.5	-2.5	
17								
18						W+	89	
19						W-	31	
20						W	31	
21								

scribed as an ANOVA with the data replaced by their ranks. The null and research hypotheses for the Kruskal–Wallis nonparametric test are as follows:

H_0: The *k* population medians are equal.

H_1: The *k* population medians are not all equal.

The procedure for the test involves pooling the observations from the *k* samples into one combined sample, keeping track of which sample each observation comes from, and then ranking lowest to highest from 1 to *N*, where $N = n_1 + n_2 + \cdots + n_k$.

In Chapter 10 of the textbook, we presented the following formula for the test statistic, *H*, in the Kruskal–Wallis test: $H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1)$, where k = the number of comparison groups, N = the total sample size, n_j is the sample size in the *j*th group, and R_j is the sum of the ranks in the *j*th group.

When performing the test of hypothesis by hand, we computed the test statistic *H* and found the appropriate critical value in Table 8 in the Appendix to set up the decision rule: Reject H_0 if $H \geq$ critical value from Table 8. Excel does not have a specific analysis tool for the Kruskal–Wallis test. However, Excel can be used to assign ranks and to compute the test statistic. The conclusion of the test is based on a comparison of the test statistic to the appropriate critical value from Table 8.

Example 10.4. In Example 10.8 of the textbook, we analyzed a clinical study designed to assess differences in albumin levels in adults following different low-protein diets. Three diets were compared, ranging from 5% to 15% protein, and the 15% protein diet represents a typical American diet. The albumin levels of participants following each diet are shown in Table 10-3.

TABLE 10-3 Albumin Levels in Three Different Diets

5% Protein	10% Protein	15% Protein
3.1	3.8	4.0
2.6	4.1	5.5
2.9	2.9	5.0
	3.4	4.8
	4.2	

The question of interest is whether there is a difference in albumin levels among the three different diets. The test is run at a 5% level of significance, and the hypotheses are as follows:

H_0: The three population medians are equal.

H_1: The three population medians are not all equal.

The data are entered into Excel as shown in Figure 10-25.

To conduct the test we assign ranks using the procedures outlined in Section 10.1. This is done on the combined or total sample (i.e., pooling the data from the three comparison groups), and ranks are assigned from 1 to 12. We also need to keep track of the group assignments in the total sample (n = 12). The ranks are shown in Figure 10-26.

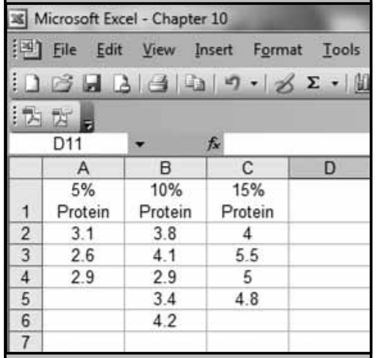
Notice that the range of the data specified in the RANK function is $A$2 through $C$6 (see formula in the top menu bar). To compute the test statistic H, we need the sum of the ranks in each group, R_j. These are denoted R_1, R_2, and R_3 and are shown in Figure 10-27. Recall that the sum of the ranks will always equal $n(n + 1)/2 = 12(13)/2 = 78$. This is shown in Figure 10-28.

The test statistic for the Kruskal–Wallis test is denoted H and is defined as follows:

$$H = \left(\frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{R_j^2}{n_j}\right) - 3(N+1)$$

To compute the test statistic in Excel, we first compute the sample sizes in each group, n_j, using the COUNT function. We

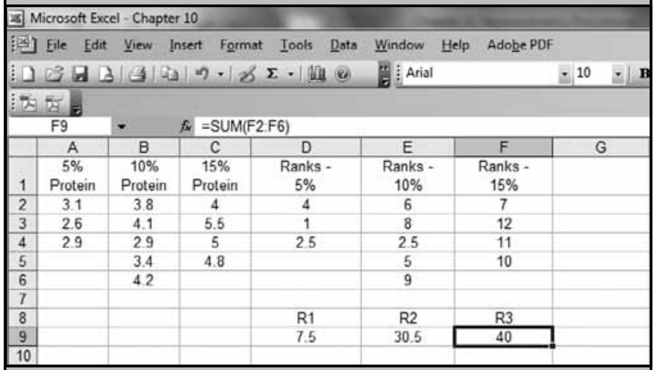
FIGURE 10-25 Data for the Kruskal–Wallis Test


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools

D11 fx

	A	B	C	D
1	5% Protein	10% Protein	15% Protein	
2	3.1	3.8	4	
3	2.6	4.1	5.5	
4	2.9	2.9	5	
5		3.4	4.8	
6		4.2		
7				

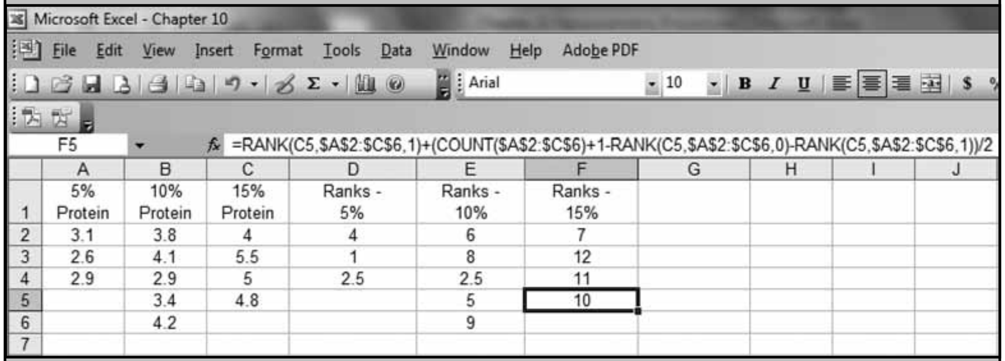
FIGURE 10-27 Summing the Ranks in Each Group


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10

F9 fx =SUM(F2:F6)

	A	B	C	D	E	F	G
1	5% Protein	10% Protein	15% Protein	Ranks - 5%	Ranks - 10%	Ranks - 15%	
2	3.1	3.8	4	4	6	7	
3	2.6	4.1	5.5	1	8	12	
4	2.9	2.9	5	2.5	2.5	11	
5		3.4	4.8		5	10	
6		4.2			9		
7							
8				R1	R2	R3	
9				7.5	30.5	40	
10							

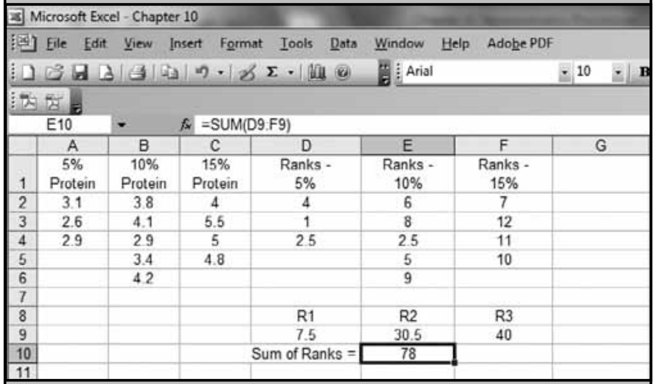
FIGURE 10-26 Ranking the Data in the Pooled Sample


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B I U
$

F5 fx =RANK(C5,$A$2:$C$6,1)+(COUNT($A$2:$C$6)+1-RANK(C5,$A$2:$C$6,0)-RANK(C5,$A$2:$C$6,1))/2

	A	B	C	D	E	F	G	H	I	J
1	5% Protein	10% Protein	15% Protein	Ranks - 5%	Ranks - 10%	Ranks - 15%				
2	3.1	3.8	4	4	6	7				
3	2.6	4.1	5.5	1	8	12				
4	2.9	2.9	5	2.5	2.5	11				
5		3.4	4.8		5	10				
6		4.2			9					
7										

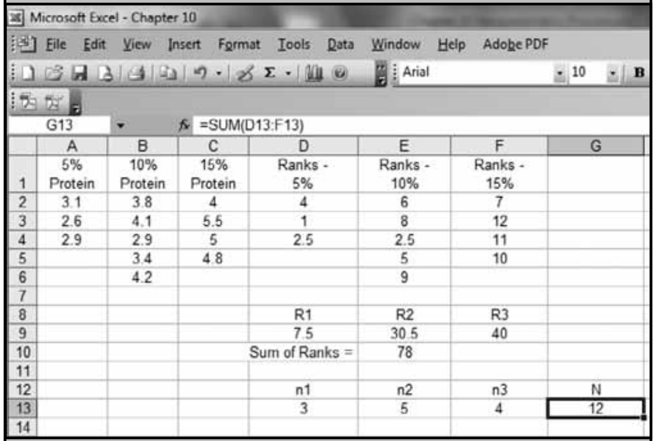
FIGURE 10-28 Checking the Sum of the Ranks


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B

E10 =SUM(D9:F9)

	A	B	C	D	E	F	G
1	5% Protein	10% Protein	15% Protein	Ranks - 5%	Ranks - 10%	Ranks - 15%	
2	3.1	3.8	4	4	6	7	
3	2.6	4.1	5.5	1	8	12	
4	2.9	2.9	5	2.5	2.5	11	
5		3.4	4.8		5	10	
6		4.2			9		
7							
8				R1	R2	R3	
9				7.5	30.5	40	
10				Sum of Ranks =	78		
11							

FIGURE 10-29 Determining Sample Sizes, n_j, and Total Sample Size N


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B

G13 =SUM(D13:F13)

	A	B	C	D	E	F	G
1	5% Protein	10% Protein	15% Protein	Ranks - 5%	Ranks - 10%	Ranks - 15%	
2	3.1	3.8	4	4	6	7	
3	2.6	4.1	5.5	1	8	12	
4	2.9	2.9	5	2.5	2.5	11	
5		3.4	4.8		5	10	
6		4.2			9		
7							
8				R1	R2	R3	
9				7.5	30.5	40	
10				Sum of Ranks =	78		
11							
12				n1	n2	n3	N
13				3	5	4	12
14							

also sum the sample sizes to compute the total sample size N. This is shown in Figure 10-29.

We now use R_j, n_j, and N to compute the test statistic H using Excel. The formula to compute the test statistic H in cell E15 is shown in the top menu bar in Figure 10-30. Thus, $H = 7.52$. To draw a final conclusion of the test, we must determine whether the observed test statistic, H, supports the null or the research hypothesis. This is done by determining a critical value of H such that if the observed value of H is greater than or equal to the critical value, we reject H_0 in favor of H_1, and if the observed value of H is less than the critical value, we do not reject H_0. The critical value of H can be found in Table 8 in the Appendix. For $n_1 = 3$, $n_2 = 5$, and $n_3 = 4$ and a level of significance of $\alpha = 0.05$, the critical value is 5.656, thus we reject H_0 because $7.52 \geq 5.656$. We have statistically significant evidence at $\alpha = 0.05$, to show that there is a difference in median albumin levels among the three different diets.

As we described in Chapter 10 of the textbook, if there are three or more comparison groups and five or more observations in each of the comparison groups, it can be shown that the test statistic H approximates a χ^2 distribution with $df = k - 1$. Thus, in a Kruskal–Wallis test with three or more comparison groups and five or more observations in each group, it is possible to compute a p-value for the test using the χ^2 distribution. Specifically, we use Excel to compute the test statistic H and a p-value, and the investigator then compares the p-value to the predetermined level of significance to draw a conclusion about the hypotheses using the rule: Reject H_0 if $p \leq \alpha$. The following example illustrates this situation.

Example 10.5. In Example 10.9 of the textbook, we compared anaerobic thresholds of elite athletes. The data in Table 10-4 are anaerobic thresholds for distance runners, distance cyclists, distance swimmers, and cross-country skiers.

The question of interest is whether there is a difference in anaerobic thresholds among the different groups of elite athletes. The test is run at a 5% level of significance and the hypotheses are as follows:

H_0: The four population medians are equal.

H_1: The four population medians are not all equal.

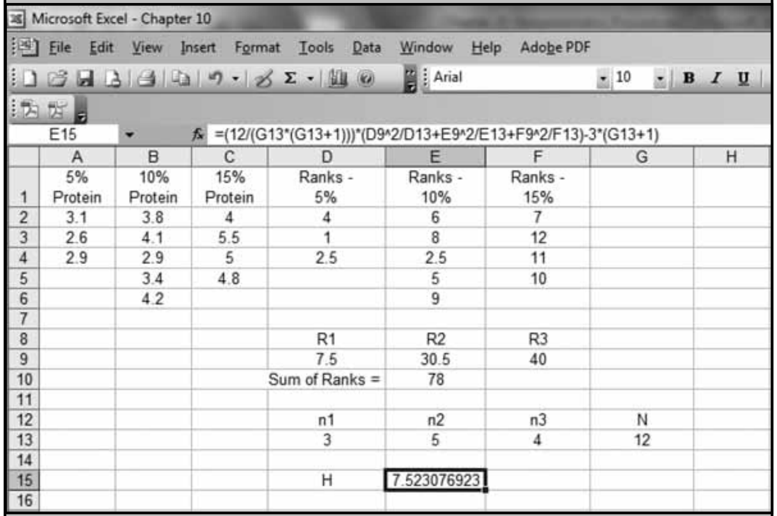
The data are entered into Excel as shown in Figure 10-31.

To conduct the test we assign ranks using the procedures outlined in Section 10.1. This is done on the combined or total sample (i.e., pooling the data from the four comparison

TABLE 10-4 Anaerobic Thresholds

Distance Runners	Distance Cyclists	Distance Swimmers	Cross-Country Skiers
185	190	166	201
179	209	159	195
192	182	170	180
165	178	183	187
174	181	160	215

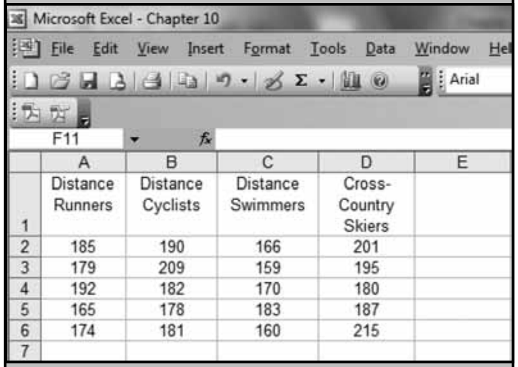
FIGURE 10-30 Computing the Test Statistic *H*


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10

E15 =(12/(G13*(G13+1)))*(D9^2/D13+E9^2/E13+F9^2/F13)-3*(G13+1)

	A	B	C	D	E	F	G	H
1	5% Protein	10% Protein	15% Protein	Ranks - 5%	Ranks - 10%	Ranks - 15%		
2	3.1	3.8	4	4	6	7		
3	2.6	4.1	5.5	1	8	12		
4	2.9	2.9	5	2.5	2.5	11		
5		3.4	4.8		5	10		
6		4.2			9			
7								
8				R1	R2	R3		
9				7.5	30.5	40		
10				Sum of Ranks =	78			
11								
12				n1	n2	n3	N	
13				3	5	4	12	
14								
15				H	7.523076923			
16								

FIGURE 10-31 Data for the Kruskal–Wallis Test


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Hel
Arial

F11

	A	B	C	D	E
1	Distance Runners	Distance Cyclists	Distance Swimmers	Cross-Country Skiers	
2	185	190	166	201	
3	179	209	159	195	
4	192	182	170	180	
5	165	178	183	187	
6	174	181	160	215	
7					

groups), and ranks are assigned from 1 to 20. We also need to keep track of the group assignments in the total sample (n = 20). The ranks are shown in Figure 10-32.

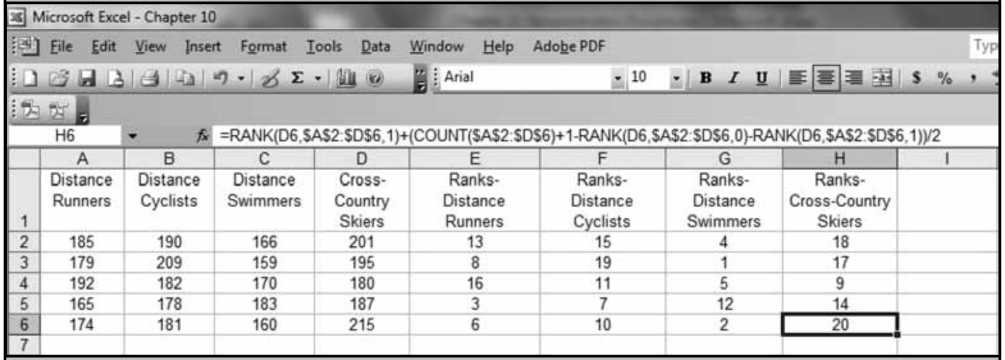
Notice that the range of the data specified in the RANK function is $A$2 through $D$6 (see formula in the top menu bar). To compute the test statistic H, we need the sum of the ranks in each group, R_j. These are denoted R_1, R_2, R_3, and R_4 and are shown in Figure 10-33. Recall that the sum of the ranks will always equal $n(n + 1)/2 = 20(21)/2 = 210$. This is shown in Figure 10-34.

The test statistic for the Kruskal–Wallis test is denoted H and is defined as follows:

$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1)$$

To compute the test statistic in Excel, we first compute the sample sizes in each group, n_j, using the COUNT function. We also sum the sample sizes to compute the total sample size N. This is shown in Figure 10-35.

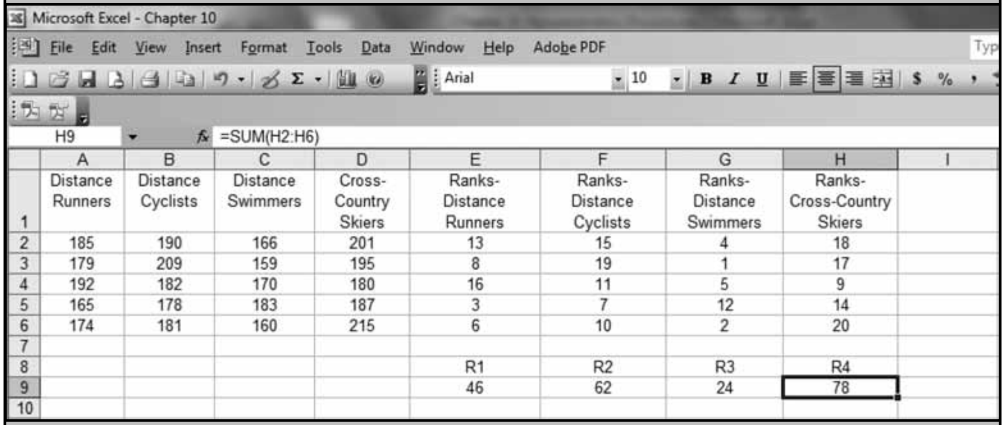
FIGURE 10-32 Ranking the Data in the Pooled Sample


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Typ
Arial
10
$ %

H6 =RANK(D6,$A$2:$D$6,1)+(COUNT($A$2:$D$6)+1-RANK(D6,$A$2:$D$6,0)-RANK(D6,$A$2:$D$6,1))/2

	A	B	C	D	E	F	G	H	I
1	Distance Runners	Distance Cyclists	Distance Swimmers	Cross-Country Skiers	Ranks-Distance Runners	Ranks-Distance Cyclists	Ranks-Distance Swimmers	Ranks-Cross-Country Skiers	
2	185	190	166	201	13	15	4	18	
3	179	209	159	195	8	19	1	17	
4	192	182	170	180	16	11	5	9	
5	165	178	183	187	3	7	12	14	
6	174	181	160	215	6	10	2	20	
7									

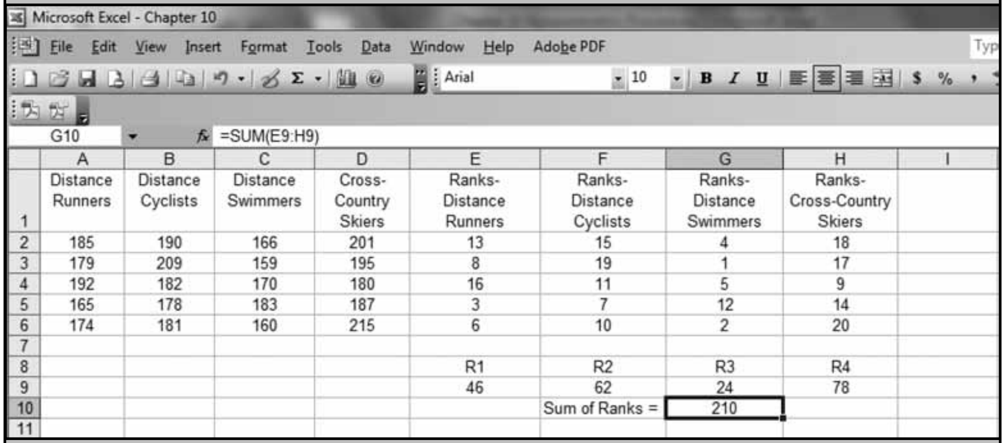
FIGURE 10-33 Summing the Ranks in Each Group


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B I U
$ %

H9 =SUM(H2:H6)

	A	B	C	D	E	F	G	H	I
1	Distance Runners	Distance Cyclists	Distance Swimmers	Cross-Country Skiers	Ranks-Distance Runners	Ranks-Distance Cyclists	Ranks-Distance Swimmers	Ranks-Cross-Country Skiers	
2	185	190	166	201	13	15	4	18	
3	179	209	159	195	8	19	1	17	
4	192	182	170	180	16	11	5	9	
5	165	178	183	187	3	7	12	14	
6	174	181	160	215	6	10	2	20	
7									
8					R1	R2	R3	R4	
9					46	62	24	78	
10									

FIGURE 10-34 Checking the Sum of the Ranks


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B I U
$ %

G10 =SUM(E9:H9)

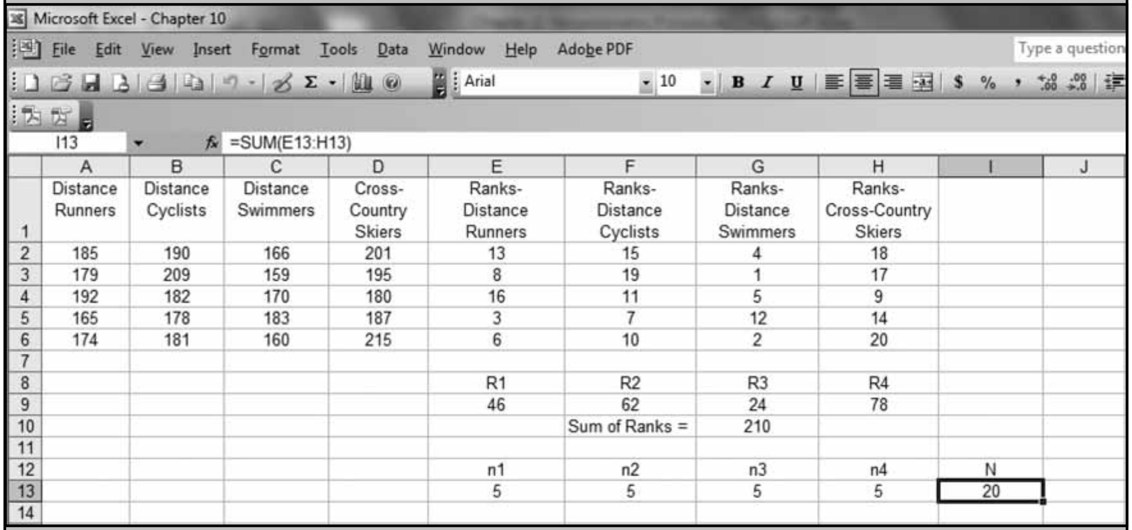
	A	B	C	D	E	F	G	H	I
1	Distance Runners	Distance Cyclists	Distance Swimmers	Cross-Country Skiers	Ranks-Distance Runners	Ranks-Distance Cyclists	Ranks-Distance Swimmers	Ranks-Cross-Country Skiers	
2	185	190	166	201	13	15	4	18	
3	179	209	159	195	8	19	1	17	
4	192	182	170	180	16	11	5	9	
5	165	178	183	187	3	7	12	14	
6	174	181	160	215	6	10	2	20	
7									
8					R1	R2	R3	R4	
9					46	62	24	78	
10						Sum of Ranks =	210		
11									

We now use R_j, n_j, and N to compute the test statistic H using Excel. The formula to compute the test statistic H in cell G15 is shown in the top menu bar in Figure 10-36.

Thus, $H = 9.11$. We can now compute a p-value for the test using the χ^2 distribution and the CHIDIST function in Excel. The CHIDIST function requires specification of the test statistic (H in this case) and the degrees of freedom. The degrees of freedom is defined as $k - 1$, where k is the number of comparison groups. Computation of the p-value for the test is shown in Figure 10-37.

Because the p-value = 0.0278 is less than the level of significance ($\alpha = 0.05$), we reject H_0. We have statistically

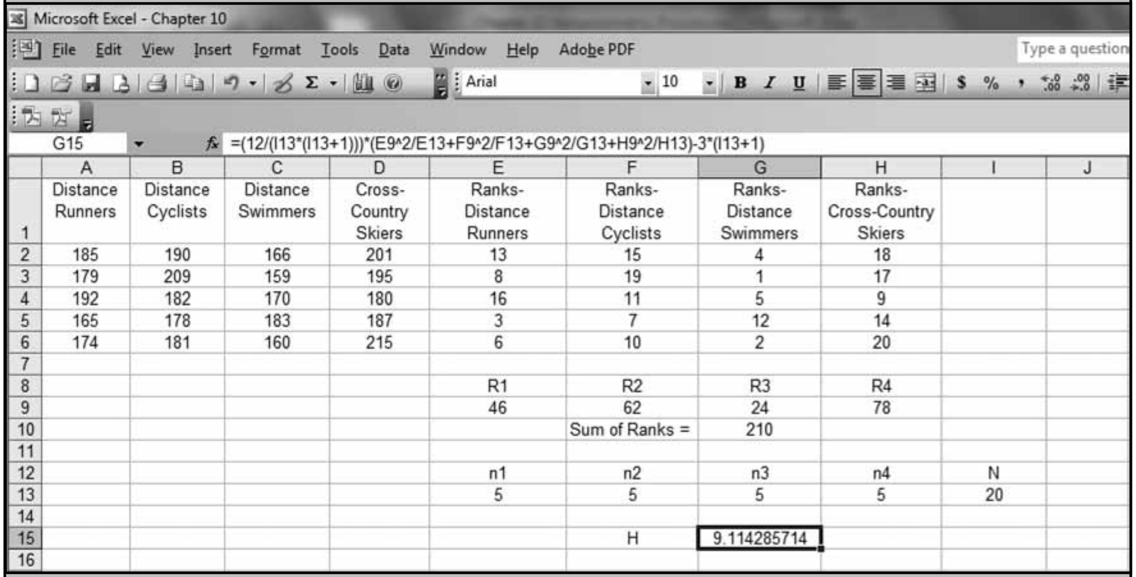
FIGURE 10-35 Determining the Sample Sizes, n_j, and Total Sample Size N


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Type a question
Arial
10

I13 =SUM(E13:H13)

	A	B	C	D	E	F	G	H	I	J
1	Distance Runners	Distance Cyclists	Distance Swimmers	Cross-Country Skiers	Ranks-Distance Runners	Ranks-Distance Cyclists	Ranks-Distance Swimmers	Ranks-Cross-Country Skiers		
2	185	190	166	201	13	15	4	18		
3	179	209	159	195	8	19	1	17		
4	192	182	170	180	16	11	5	9		
5	165	178	183	187	3	7	12	14		
6	174	181	160	215	6	10	2	20		
7										
8					R1	R2	R3	R4		
9					46	62	24	78		
10						Sum of Ranks =	210			
11										
12					n1	n2	n3	n4	N	
13					5	5	5	5	20	
14										

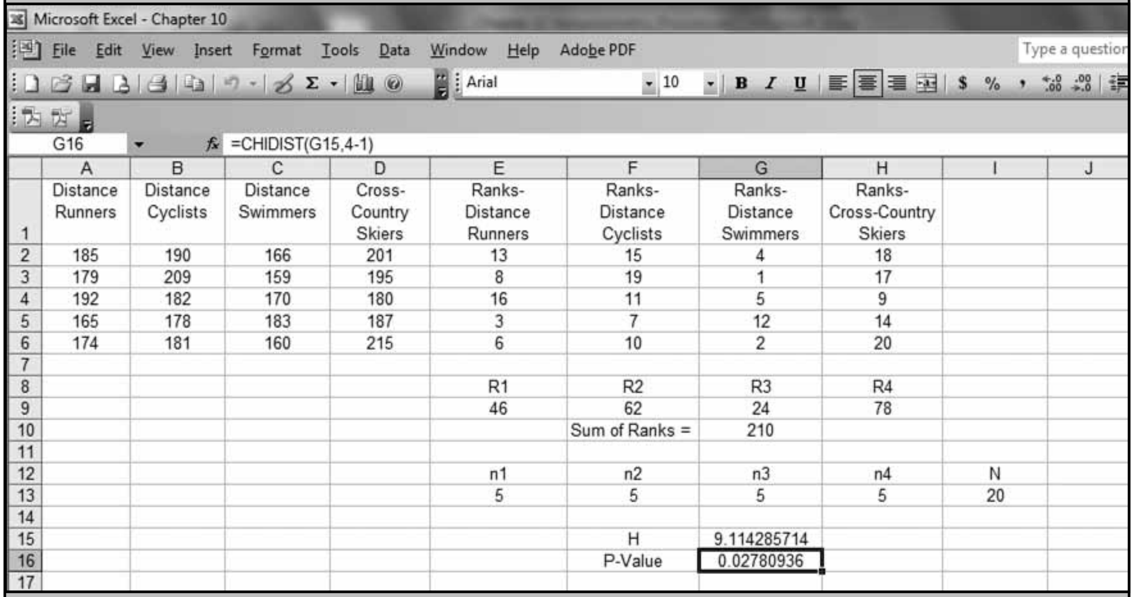
FIGURE 10-36 Computing the Test Statistic H


Microsoft Excel - Chapter 10
File Edit View Insert Format Tools Data Window Help Adobe PDF
Type a question
Arial
10

G15 =(12/(I13*(I13+1)))*(E9^2/E13+F9^2/F13+G9^2/G13+H9^2/H13)-3*(I13+1)

	A	B	C	D	E	F	G	H	I	J
1	Distance Runners	Distance Cyclists	Distance Swimmers	Cross-Country Skiers	Ranks-Distance Runners	Ranks-Distance Cyclists	Ranks-Distance Swimmers	Ranks-Cross-Country Skiers		
2	185	190	166	201	13	15	4	18		
3	179	209	159	195	8	19	1	17		
4	192	182	170	180	16	11	5	9		
5	165	178	183	187	3	7	12	14		
6	174	181	160	215	6	10	2	20		
7										
8					R1	R2	R3	R4		
9					46	62	24	78		
10						Sum of Ranks =	210			
11										
12					n1	n2	n3	n4	N	
13					5	5	5	5	20	
14										
15						H	9.114285714			
16										

FIGURE 10-37 Computing the *p*-Value Using the Chi-Square Distribution



G16 =CHIDIST(G15,4-1)

	A	B	C	D	E	F	G	H	I	J
1	Distance Runners	Distance Cyclists	Distance Swimmers	Cross-Country Skiers	Ranks-Distance Runners	Ranks-Distance Cyclists	Ranks-Distance Swimmers	Ranks-Cross-Country Skiers		
2	185	190	166	201	13	15	4	18		
3	179	209	159	195	8	19	1	17		
4	192	182	170	180	16	11	5	9		
5	165	178	183	187	3	7	12	14		
6	174	181	160	215	6	10	2	20		
7										
8					R1	R2	R3	R4		
9					46	62	24	78		
10						Sum of Ranks =	210			
11										
12					n1	n2	n3	n4	N	
13					5	5	5	5	20	
14										
15						H	9.114285714			
16						P-Value	0.02780936			
17										

significant evidence at $\alpha = 0.05$ to show that there is a difference in median anaerobic thresholds among the four different groups of elite athletes.

10.5 PRACTICE PROBLEMS

1. A company is evaluating the impact of a wellness program offered on-site as a means of reducing employee sick days. A total of 8 employees agree to participate in the evaluation, which lasts 12 weeks. Their sick days in the 12 months prior to the start of the wellness program and again over the 12 months after the completion of the program are recorded and are shown in Table 10-5. Is there a significant reduction in the number of sick days taken after completing the wellness program? Use the Sign test at a 5% level of significance.

TABLE 10-5 Data for Practice Problems 1 and 2

Employee	Sick Days Taken in 12 Months Prior to Program	Sick Days Taken in 12 Months Following program
1	8	7
2	6	6
3	4	5
4	12	11
5	10	7
6	8	4
7	6	3
8	2	1

2. Using the data in Problem 1, assess whether there is there a significant reduction in the number of sick days taken after completing the wellness program using the Wilcoxon Signed Rank test at a 5% level of significance.
3. A small study ($n = 10$) is designed to assess whether there is an association between smoking in pregnancy and low birth weight. Low-birth-weight babies are those born less than 5.5 pounds. The following data represent the birth weights (in pounds) of babies born to mothers who reported smoking in pregnancy and to those who did not.

Mother smoked in pregnancy	5.0	4.2	4.8	3.3	3.9
Mother did not smoke during pregnancy	5.1	4.9	5.3	5.4	4.6

Is there a significant difference in birth weights between mothers who smoked during pregnancy and those who did not? Run the appropriate test at a 5% level of significance.

4. The following data represent the number of playground injuries occurring among children aged 5 to 9 years over a 3-month period in 12 playgrounds in and around the neighborhoods of Boston. Playground injuries include fractures, internal injuries, lacerations, and dislocations. The question of interest is whether there are differences in the numbers of injuries at playgrounds in various locations. The data below represent the numbers of injuries recorded at four randomly selected playgrounds located on school properties, at day-care centers, and in residential neighborhoods.

School properties	39	51	42	29
Day-care centers	28	25	30	15
Residential neighborhoods	28	16	25	22

Run the appropriate test at a 5% level of significance.

5. The recommended daily allowance of Vitamin A for children between 1 and 3 years of age is 400 micrograms (mcg). Vitamin A deficiency is linked to a number of adverse health outcomes including poor eyesight, susceptibility to infection, and dry skin. The following are Vitamin A concentrations in children with and without poor eyesight, a history of infection, and dry skin.

With poor eyesight, a history of infection, and dry skin	270	420	180	345	390	430
Free of poor eyesight, a history of infection, and dry skin	450	500	395	380	430	

Is there a significant difference in Vitamin A concentrations between children with and without poor eyesight, a history of infection, and dry skin? Run the appropriate test at a 5% level of significance.

TABLE 10-6 Data for Practice Problem 6

Participant	Total Cholesterol Before Treatment	Total Cholesterol After Treatment
1	250	241
2	265	260
3	240	253
4	233	230
5	255	224
6	275	227
7	241	232

6. A study is conducted to assess the potential benefits of an ayurvedic treatment to reduce high cholesterol. Seven patients agree to participate in the study. Each has their cholesterol measured at the start of the study and then again after 4 weeks taking a popular herb called arjuna (see Table 10-6). Is there a significant difference in total cholesterol after taking the herb? Use the Sign test at a 5% level of significance.
7. Using the data in Problem 6, assess whether there is there a difference in total cholesterol after taking the herb using the Wilcoxon Signed Rank test at a 5% level of significance.
8. An investigator wants to test if there is a difference in endotoxin levels in children who are exposed to endotoxin as a function of their proximity to operating farms. The following are endotoxin levels in units per milligram of dust sampled from children's mattresses, organized by children's proximity to farms.

Within 5 miles	54	62	78	90	70
5–24.9 miles	28	42	39	81	65
25–49.9 miles	37	29	30	50	53
50 miles or more	36	19	22	28	27

Run the appropriate test at a 5% level of significance.

CHAPTER 11

Survival Analysis

In Chapter 11 of the textbook we presented techniques to analyze time-to-event data, or survival data. Because of the unique features of survival data, most specifically the presence of censoring, special statistical procedures are necessary to analyze these data. In survival analysis applications, it is often of interest to estimate the survival function, or survival probabilities over time. We presented two popular nonparametric techniques called the life table or actuarial table approach and the Kaplan–Meier approach to constructing cohort life tables or follow-up life tables. Both approaches generate estimates of the survival function which can be used to estimate the probability that a participant survives to a specific time (e.g., 5 or 10 years) or the median survival time. Microsoft Office® Excel® 2003 does not have a built-in analysis tool to construct a follow-up life table. Here we use Excel to program the formulas to estimate survival probabilities and to generate graphical displays of survival functions.

It is also often of interest to assess whether there are statistically significant differences in survival between groups, i.e., between competing treatment groups in a clinical trial, or between men and women, or between patients with and without a specific risk factor in an observational study. There are many statistical tests available; in Chapter 11 of the textbook we presented the log-rank test, which is a popular nonparametric test to compare survival between two independent groups. It makes no assumptions about the survival distributions and can be conducted relatively easily using life tables based on the Kaplan–Meier approach. Again, Excel does not have a built-in analysis tool to conduct the log-rank test. However, we use Excel to program the computation of the test statistic and to determine a *p*-value which can be used to assess the statistical significance of the difference in survival between groups.

In Chapter 11 of the textbook we also discussed Cox proportional hazards regression analysis, which is a popular multivariable technique to estimate the effect of several risk factors, considered simultaneously, on survival. Excel does not have the capability to estimate the parameters of the Cox proportional hazards model. Interested readers should see Allison[1] for details regarding the estimation of parameters in a Cox proportional hazards regression model using SAS®.

11.1 ESTIMATING THE SURVIVAL FUNCTION

There are several different ways to estimate a survival function or a survival curve. A number of popular parametric methods are used to model survival data, and they differ in terms of the assumptions that are made about the distribution of survival times in the population. In Chapter 11 of the textbook, we focused on two nonparametric methods, the actuarial or life table approach and the Kaplan–Meier approach, which made no assumptions about how the probability that a person develops the event changes over time. The approaches are summarized here and their implementation is illustrated using Excel.

With the actuarial or life table approach, we first organize the observed follow-up times into equally spaced intervals. We might, for example, consider 1-, 2-, or 5-year intervals depending on the duration of the follow-up. We then sum the number of participants who are at risk at the beginning of each interval, the number who suffer the event of interest, and the number who are censored or lost to follow-up in each interval. We compute the proportions who suffer the event of interest and who do not in each interval, and then we compute the survival probability. The notation we presented in Chapter 11 of the textbook is summarized as follows:

N_t = number of participants who are event-free and considered at risk during interval t

D_t = number of participants who suffer the event of interest during interval t

C_t = number of participants who are censored during interval t

N_{t^*} = average number of participants at risk during interval t, $N_{t^*} = N_t - C_t/2$

q_t = proportion suffering the event of interest during interval t, $q_t = D_t/N_{t^*}$

p_t = proportion remaining event-free during interval t, $p_t = 1 - q_t$

S_t = proportion remaining event-free past interval t, $S_{t+1} = p_{t+1} \times S_t$, where $S_0 = 1$.

With the Kaplan–Meier approach, we do not consider equally spaced intervals; instead, we re-estimate the survival probability at each observed event time. At each observed time (event time or censored time), we compute the number of participants at risk at that time (N_t), the number of deaths at that time (D_t), the number censored (C_t), and the survival probability (S_t). The survival probabilities are computed using $S_{t+1} = S_t \times ((N_{t+1} - D_{t+1})/N_{t+1})$. It is important to note that the calculations using the Kaplan–Meier approach are similar to those using the actuarial life table approach. The main difference is the time intervals. With the actuarial life table approach we consider equally spaced intervals, while with the Kaplan–Meier approach we use observed event times and censoring times.

Example 11.1. In Example 11.2 in the textbook, we analyzed a small prospective cohort study with death as the primary outcome. The study involved participants who were 65 years of age and older who were followed for up to 24 years. The study involved 20 participants ($n = 20$) who were enrolled over a period of 5 years and followed until death, until the study ended, or until they dropped out of the study (lost to follow-up). The data are shown in Table 11-1. We use Excel to construct a life table using the actuarial approach. The data are entered into Excel as shown in Figure 11-1.

To construct the life table, we create two new variables. The first is the observed time (either year of death or year of last contact). We create the time variable in column D, and we use the IF function to copy either the year of death or the year of last contact from column A or B, respectively, depending on which was measured. Specifically, in cell D2 we enter "=IF(A2>0, A2,B2)". Recall that the IF function checks the specified criterion (in this case, whether A2>0). Here, if the criterion is met, then the value of A2 is placed in cell D2, otherwise the value of B2 is placed in cell D2. The formula is then copied from cell D2 into cell D3 through cell D21 (see the formula in the top menu bar for the entry in cell D21 in Figure 11-2).

TABLE 11-1 Year of Death or Year of Last Contact

Participant Identification Number	Year of Death	Year of Last Contact
1		24
2	3	
3		11
4		19
5		24
6		13
7	14	
8		2
9		18
10		17
11		24
12		21
13		12
14	1	
15		10
16	23	
17		6
18	5	
19		9
20	17	

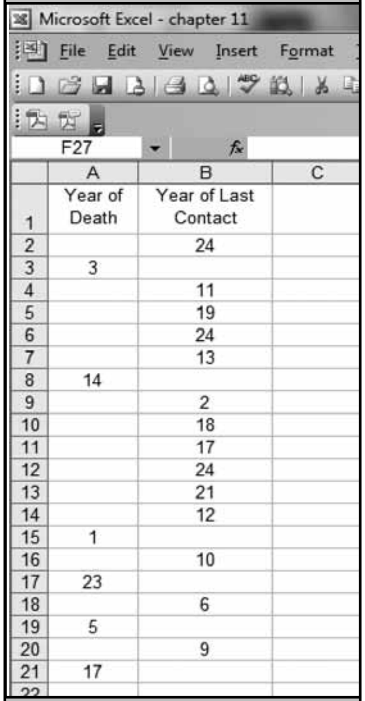
We now create an indicator variable (coded 0 or 1) to indicate whether the participant suffered the event of interest (death in this example) or not. We again use the IF function, and if a year of death is recorded in column A we assign a 1; otherwise, we assign a 0 to indicate that the observed time is censored. The event variable is created in column E as shown in Figure 11-3.

Using Excel, we now construct the life table using the time variable and the event indicator. (It is not necessary to sort the data to construct the table, although the data sorted by time does facilitate interpretation.)

To construct the life table, we first organize the follow-up times into equally spaced intervals. In this example we have a maximum follow-up of 24 years, and we consider 5-year intervals (0–4 years, 5–9 years, 10–14 years, 15–19 years, and 20–24 years). In Excel we now create two new variables that indicate the start and end of each interval. The start and end of the desired intervals are entered into columns G and H, respectively, as shown in Figure 11-4.

Next, we sum the number of participants who are alive at the beginning of each interval (N_t), and the number who die (D_t) and the number who are censored (C_t) in each interval. To compute these sums, we use the COUNTIF function. For example, to compute the number of participants alive at the beginning of each interval we specify "=COUNTIF(D2: D21,">="*start of interval*)", where D2:D21 refers to the range

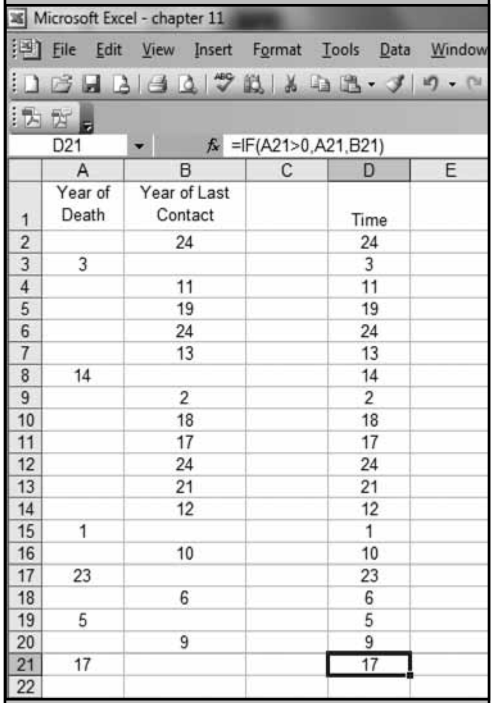
FIGURE 11-1 Data for Life Table Using Actuarial Approach



F27 fx

	A	B	C
1	Year of Death	Year of Last Contact	
2		24	
3	3		
4		11	
5		19	
6		24	
7		13	
8	14		
9		2	
10		18	
11		17	
12		24	
13		21	
14		12	
15	1		
16		10	
17	23		
18		6	
19	5		
20		9	
21	17		

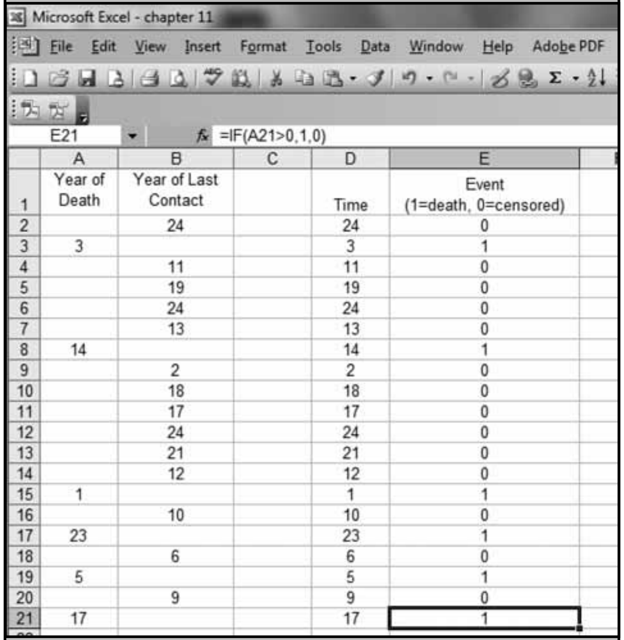
FIGURE 11-2 Creating the Time Variable


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window

D21 fx =IF(A21>0,A21,B21)

	A	B	C	D	E
1	Year of Death	Year of Last Contact		Time	
2		24		24	
3	3			3	
4		11		11	
5		19		19	
6		24		24	
7		13		13	
8	14			14	
9		2		2	
10		18		18	
11		17		17	
12		24		24	
13		21		21	
14		12		12	
15	1			1	
16		10		10	
17	23			23	
18		6		6	
19	5			5	
20		9		9	
21	17			17	
22					

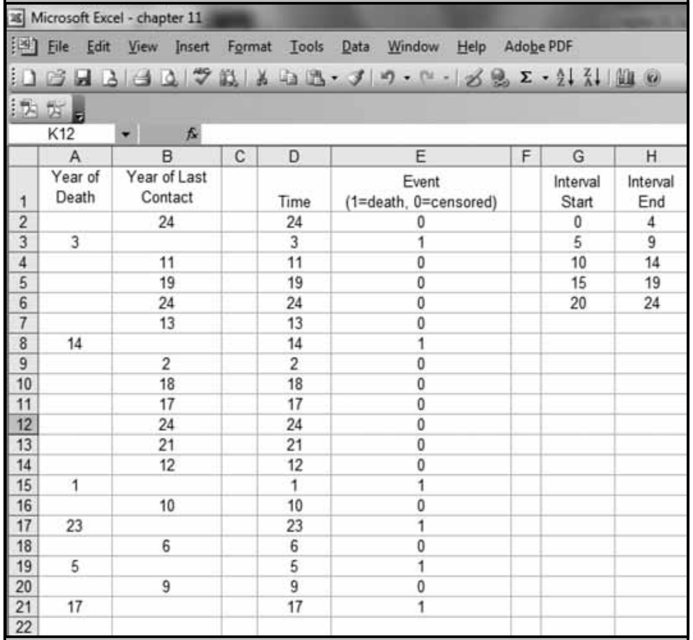
FIGURE 11-3 Creating the Event Indicator


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF

E21 fx =IF(A21>0,1,0)

	A	B	C	D	E
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)
2		24		24	0
3	3			3	1
4		11		11	0
5		19		19	0
6		24		24	0
7		13		13	0
8	14			14	1
9		2		2	0
10		18		18	0
11		17		17	0
12		24		24	0
13		21		21	0
14		12		12	0
15	1			1	1
16		10		10	0
17	23			23	1
18		6		6	0
19	5			5	1
20		9		9	0
21	17			17	1

FIGURE 11-4 Entering the Start and End of Intervals for the Life Table


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF

K12 fx

	A	B	C	D	E	F	G	H
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Interval Start	Interval End
2		24		24	0		0	4
3	3			3	1		5	9
4		11		11	0		10	14
5		19		19	0		15	19
6		24		24	0		20	24
7		13		13	0			
8	14			14	1			
9		2		2	0			
10		18		18	0			
11		17		17	0			
12		24		24	0			
13		21		21	0			
14		12		12	0			
15	1			1	1			
16		10		10	0			
17	23			23	1			
18		6		6	0			
19	5			5	1			
20		9		9	0			
21	17			17	1			
22								

of the observed times, and *start of interval* refers to the value in cell G2 through cell G6. We illustrate the computation of N_t for each interval in Figure 11-5.

Note in cell I6, we specify “=COUNTIF(D$2:D$21,“>=” &G6)”. We use absolute cell references to indicate the data range (i.e., D$2:D$21) so that when the formula is copied from cell I2 to cell I6 we continue to count the same data. The second argument to the COUNTIF function is the criterion we wish Excel to consider. Specifically, for each interval we want to count all times that are greater than or equal to the start of the interval. In row 6 this is indicated by “>=” &G6, for example. To consider a cell reference in the criterion of the COUNTIF function, we include “&” before the cell address.

We now sum the number of participants who die (D_t) in each interval. We again use the COUNTIF function. However, we sum the number of participants with year of death in each interval as follows. For interval 1 (0–4 years), we sum the number of deaths that occur between 0 and 4 years of follow-up (i.e., if 0 ≤ year of death ≤ 4). To implement the two criteria (year of death greater than or equal to 0 years and year of death less than or equal to 4 years), we first sum the number of participants with year of death greater than or equal to 0 years and then subtract the number of participants with year of death greater than 4 years. Specifically, in cell J2 we enter “=COUNTIF(A$2:A$21,“>=”&G2) – COUNTIF(A$2:A$21, “>”&H2)”. We then copy this formula from cell J2 into cell J3 through cell J6, as shown in Figure 11-6.

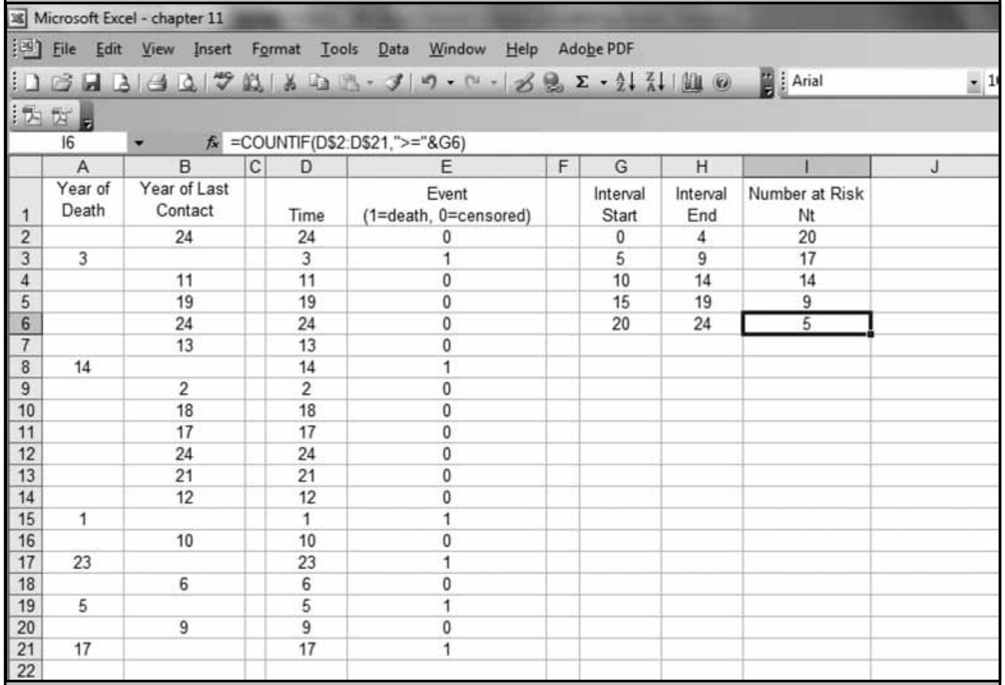
We use the same approach to sum the number of participants who are censored (C_t) in each interval. We again use the COUNTIF function. However, we sum the number of participants with year of last contact in each interval as follows. For interval 1 (0–4 years), we sum the number of participants with year of last contact between 0 and 4 years of follow-up (i.e., if 0 ≤ year of last contact ≤ 4). Specifically, in cell K2 we enter “=COUNTIF(B$2:B$21“>=”&G2) – COUNTIF(B$2:B$21, “>”&H2)”. We then copy this formula from cell K2 into cell K3 through cell K6, as shown in Figure 11-7.

We now compute N_{t^*}, the average number of participants at risk in each interval using $N_{t^*} = N_t - C_t/2$. This is shown in Figure 11-8.

Next we compute the proportion who suffer the event of interest in each interval (in this example, the proportion who die), $q_t = D_t/N_{t^*}$. This is shown in Figure 11-9.

Next we compute the proportion who remain event-free in each interval, $p_t = 1 - q_t$. This is shown in Figure 11-10.

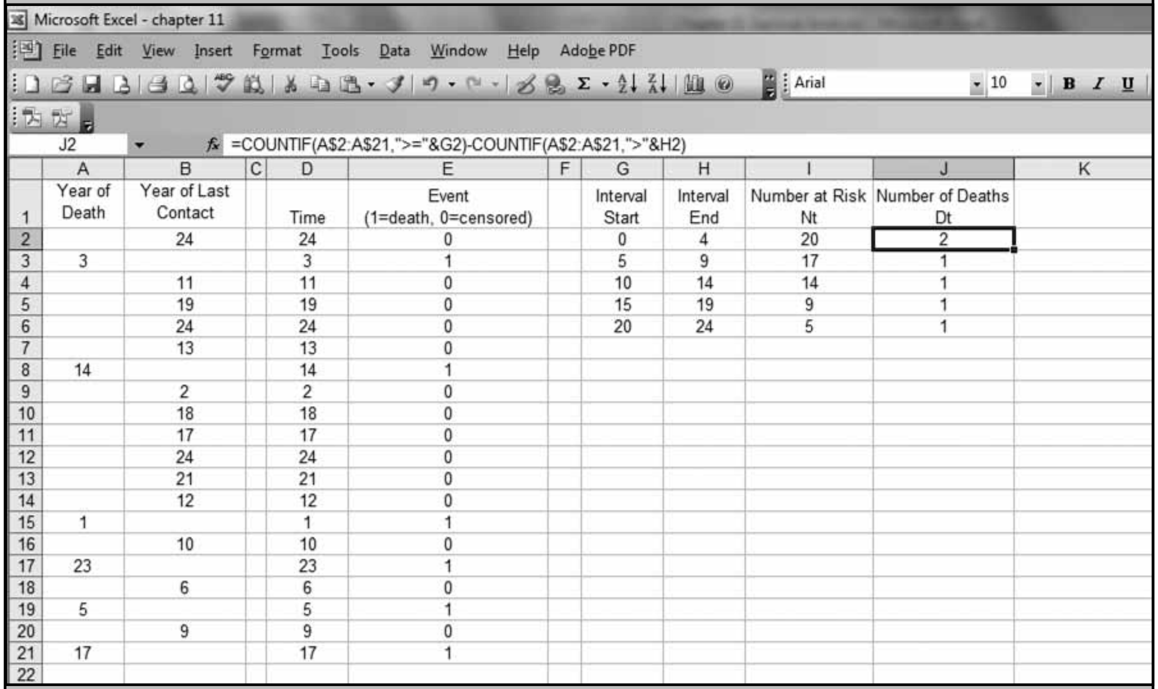
FIGURE 11-5 Computing the Number at Risk, N_t



I6 =COUNTIF(D$2:D$21,">="&G6)

	A	B	C	D	E	F	G	H	I	J
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Interval Start	Interval End	Number at Risk Nt	
2		24		24	0		0	4	20	
3	3			3	1		5	9	17	
4		11		11	0		10	14	14	
5		19		19	0		15	19	9	
6		24		24	0		20	24	5	
7		13		13	0					
8	14			14	1					
9		2		2	0					
10		18		18	0					
11		17		17	0					
12		24		24	0					
13		21		21	0					
14		12		12	0					
15	1			1	1					
16		10		10	0					
17	23			23	1					
18		6		6	0					
19	5			5	1					
20		9		9	0					
21	17			17	1					
22										

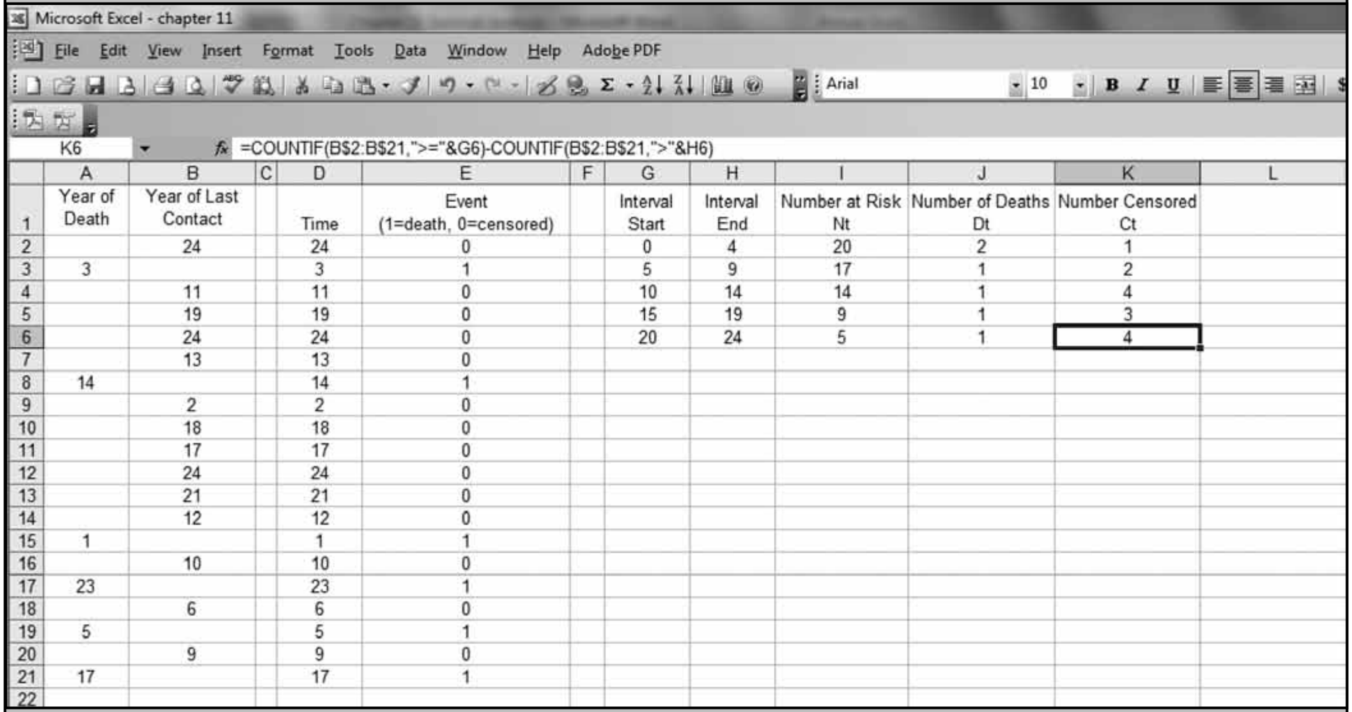
FIGURE 11-6 Computing Number of Events, D_t


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B I U

J2 =COUNTIF(A$2:A$21,">="&G2)-COUNTIF(A$2:A$21,">"&H2)

	A	B	C	D	E	F	G	H	I	J	K
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Interval Start	Interval End	Number at Risk Nt	Number of Deaths Dt	
2		24		24	0		0	4	20	2	
3	3			3	1		5	9	17	1	
4		11		11	0		10	14	14	1	
5		19		19	0		15	19	9	1	
6		24		24	0		20	24	5	1	
7		13		13	0						
8	14			14	1						
9		2		2	0						
10		18		18	0						
11		17		17	0						
12		24		24	0						
13		21		21	0						
14		12		12	0						
15	1			1	1						
16		10		10	0						
17	23			23	1						
18		6		6	0						
19	5			5	1						
20		9		9	0						
21	17			17	1						
22											

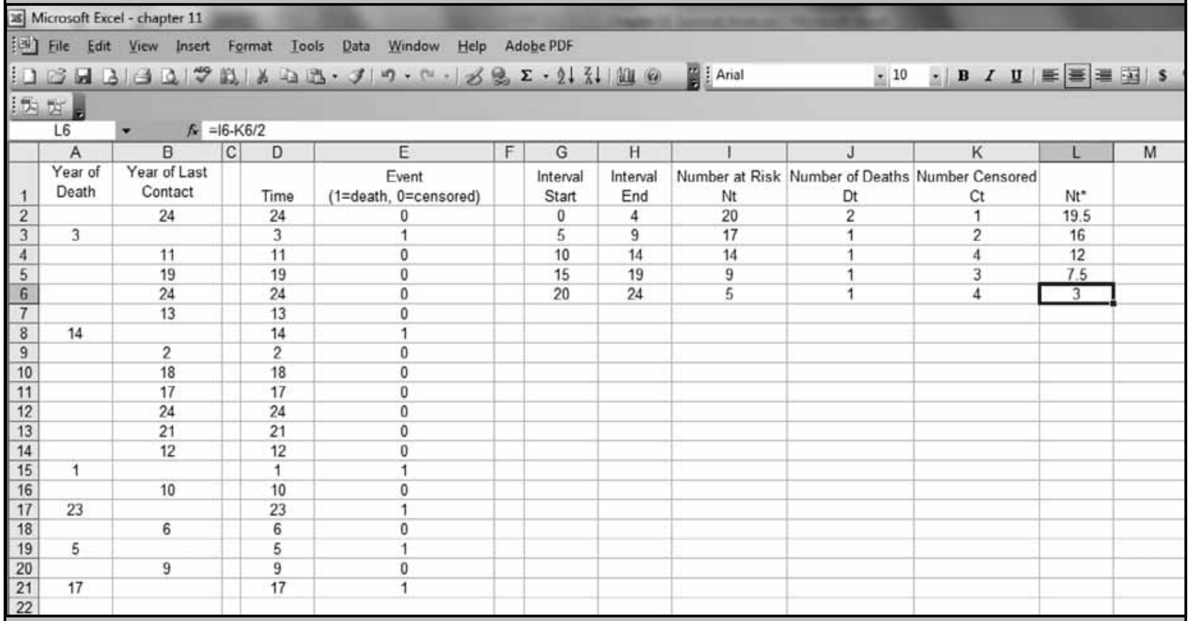
FIGURE 11-7 Computing the Number Censrored, C_t


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B I U

K6 =COUNTIF(B$2:B$21,">="&G6)-COUNTIF(B$2:B$21,">"&H6)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Interval Start	Interval End	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	
2		24		24	0		0	4	20	2	1	
3	3			3	1		5	9	17	1	2	
4		11		11	0		10	14	14	1	4	
5		19		19	0		15	19	9	1	3	
6		24		24	0		20	24	5	1	4	
7		13		13	0							
8	14			14	1							
9		2		2	0							
10		18		18	0							
11		17		17	0							
12		24		24	0							
13		21		21	0							
14		12		12	0							
15	1			1	1							
16		10		10	0							
17	23			23	1							
18		6		6	0							
19	5			5	1							
20		9		9	0							
21	17			17	1							
22												

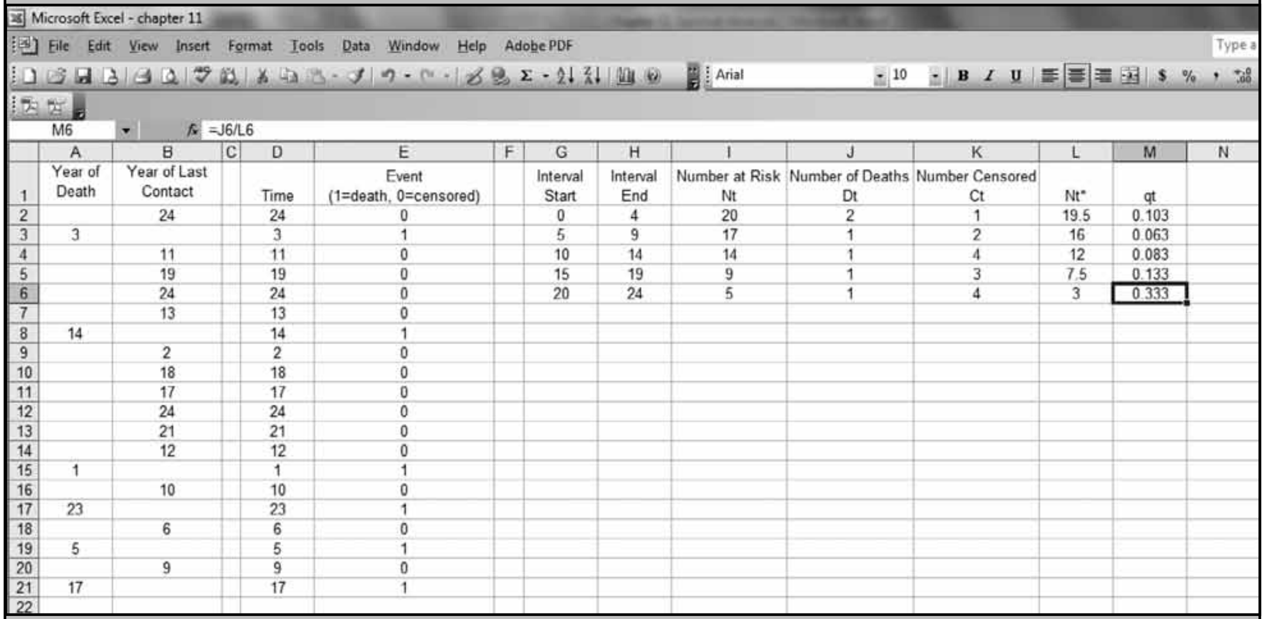
FIGURE 11-8 Computing the Average Number at Risk, N_{t*}


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10

L6 =I6-K6/2

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Interval Start	Interval End	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Nt*	
2		24		24	0		0	4	20	2	1	19.5	
3	3			3	1		5	9	17	1	2	16	
4		11		11	0		10	14	14	1	4	12	
5		19		19	0		15	19	9	1	3	7.5	
6		24		24	0		20	24	5	1	4	3	
7		13		13	0								
8	14			14	1								
9		2		2	0								
10		18		18	0								
11		17		17	0								
12		24		24	0								
13		21		21	0								
14		12		12	0								
15	1			1	1								
16		10		10	0								
17	23			23	1								
18		6		6	0								
19	5			5	1								
20		9		9	0								
21	17			17	1								
22													

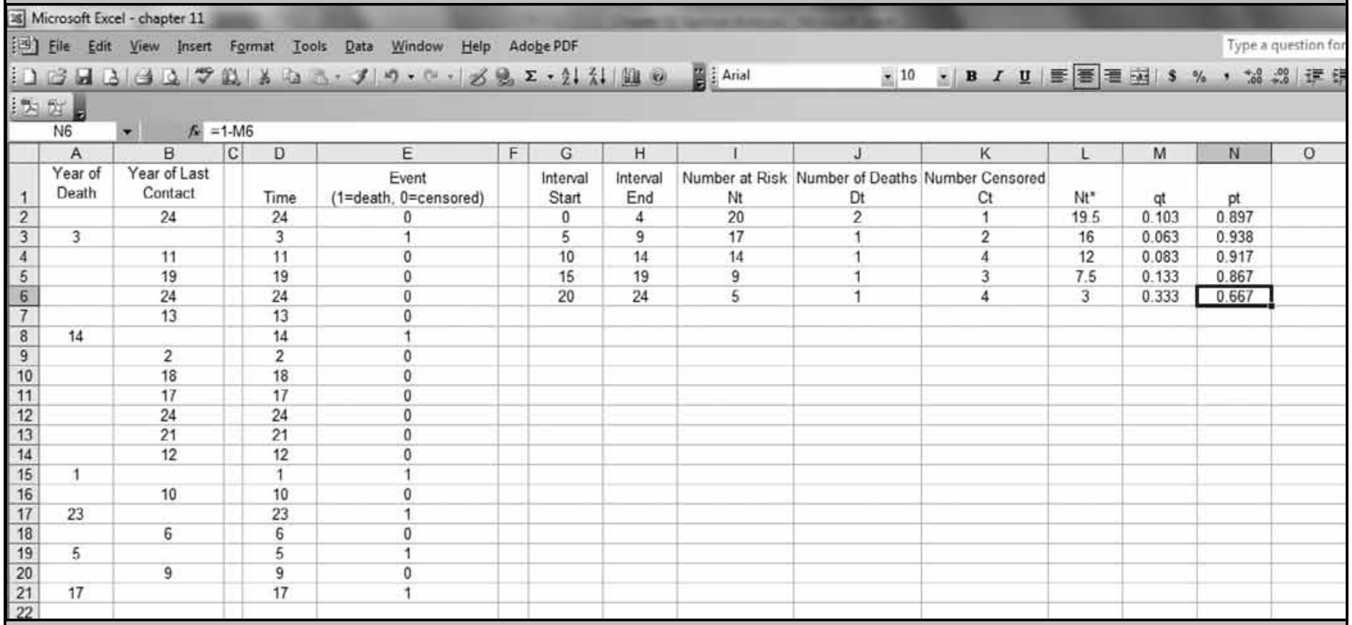
FIGURE 11-9 Computing the Proportion who Suffer Event, q_t


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Type a
Arial
10

M6 =J6/L6

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Interval Start	Interval End	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Nt*	qt	
2		24		24	0		0	4	20	2	1	19.5	0.103	
3	3			3	1		5	9	17	1	2	16	0.063	
4		11		11	0		10	14	14	1	4	12	0.083	
5		19		19	0		15	19	9	1	3	7.5	0.133	
6		24		24	0		20	24	5	1	4	3	0.333	
7		13		13	0									
8	14			14	1									
9		2		2	0									
10		18		18	0									
11		17		17	0									
12		24		24	0									
13		21		21	0									
14		12		12	0									
15	1			1	1									
16		10		10	0									
17	23			23	1									
18		6		6	0									
19	5			5	1									
20		9		9	0									
21	17			17	1									
22														

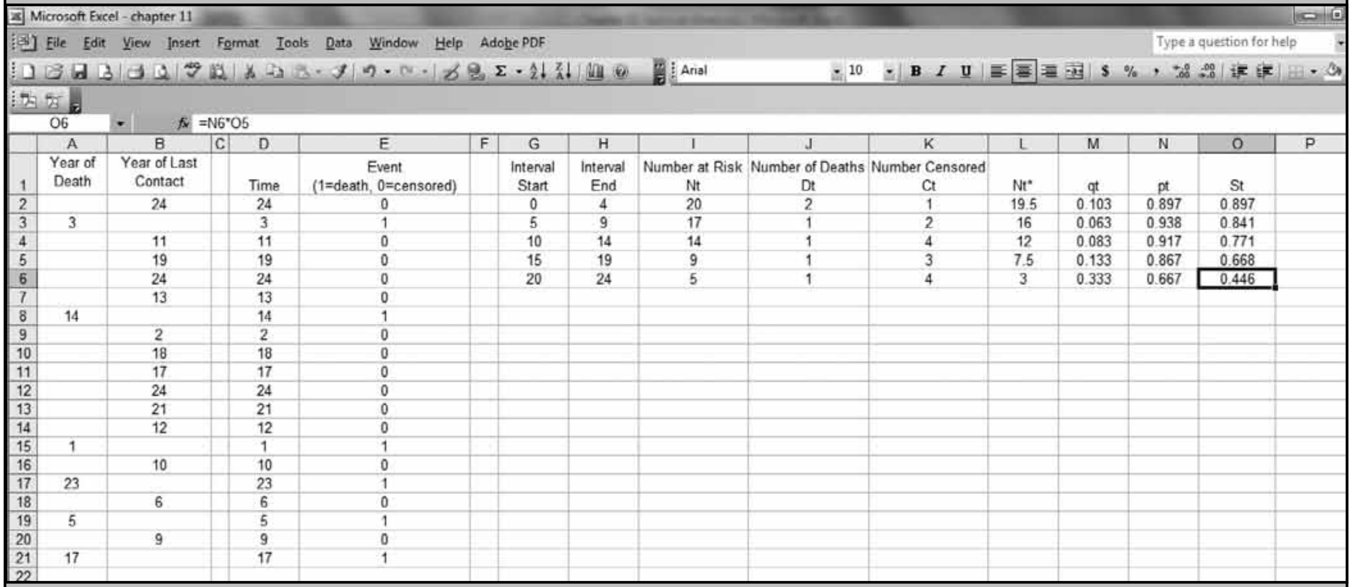
FIGURE 11-10 Computing Proportion who Remain Event-Free, p_t



N6 =1-M6

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Interval Start	Interval End	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Nt*	qt	pt	
2		24		24	0		0	4	20	2	1	19.5	0.103	0.897	
3	3			3	1		5	9	17	1	2	16	0.063	0.938	
4		11		11	0		10	14	14	1	4	12	0.083	0.917	
5		19		19	0		15	19	9	1	3	7.5	0.133	0.867	
6		24		24	0		20	24	5	1	4	3	0.333	0.667	
7		13		13	0										
8	14			14	1										
9		2		2	0										
10		18		18	0										
11		17		17	0										
12		24		24	0										
13		21		21	0										
14		12		12	0										
15	1			1	1										
16		10		10	0										
17	23			23	1										
18		6		6	0										
19	5			5	1										
20		9		9	0										
21	17			17	1										
22															

FIGURE 11-11 Computing the Survival Probabilities, S_t



O6 =N6*O5

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Interval Start	Interval End	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Nt*	qt	pt	St	
2		24		24	0		0	4	20	2	1	19.5	0.103	0.897	0.897	
3	3			3	1		5	9	17	1	2	16	0.063	0.938	0.841	
4		11		11	0		10	14	14	1	4	12	0.083	0.917	0.771	
5		19		19	0		15	19	9	1	3	7.5	0.133	0.867	0.668	
6		24		24	0		20	24	5	1	4	3	0.333	0.667	0.446	
7		13		13	0											
8	14			14	1											
9		2		2	0											
10		18		18	0											
11		17		17	0											
12		24		24	0											
13		21		21	0											
14		12		12	0											
15	1			1	1											
16		10		10	0											
17	23			23	1											
18		6		6	0											
19	5			5	1											
20		9		9	0											
21	17			17	1											
22																

The final step is to compute the survival probabilities, S_t, using $S_{t+1} = p_{t+1} \times S_t$. Recall, $S_0 = 1$ and thus we compute the survival probability for the first interval by specifying "=N2*1" in cell O2. The survival probabilities for the subsequent intervals use the formula $S_{t+1} = p_{t+1} \times S_t$, as shown in Figure 11-11.

While it is tedious to construct the life table the first time, once the structure is entered into Excel, construction of life tables for new datasets is relatively easy. Only the data range in the formulas must be modified.

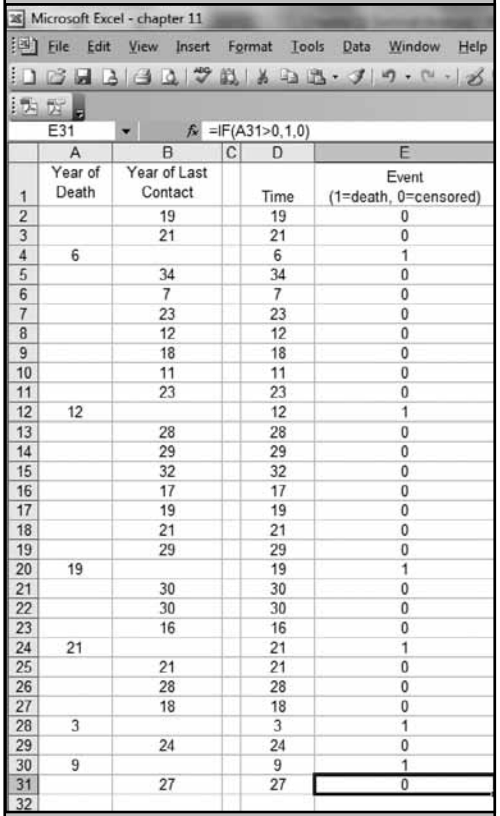
Example 11.2. Consider a prospective cohort study where 30 participants ($n = 30$) are enrolled and followed until time of death, until the study ends, or until they drop out of the study (i.e., are lost to follow-up). The data are shown in Table 11-2, and follow-up times are measured in years.

We use Excel to construct a life table using the actuarial approach and the template we developed for Example 11.1. We first copy the worksheet with all of the formulas we developed for Example 11.1 and enter the data above into the worksheet as shown in Figure 11-12.

TABLE 11-2 Year of Death or Year of Last Contact

Participant Identification Number	Year of Death	Year of Last Contact
1		19
2		21
3	6	
4		34
5		7
6		23
7		12
8		18
9		11
10		23
11	12	
12		28
13		29
14		32
15		17
16		19
17		21
18		29
19	19	
20		30
21		30
22		16
23	21	
24		21
25		28
26		18
27	3	
28		24
29	9	
30		27

FIGURE 11-12 Data for Life Table Using Actuarial Approach



E31 =IF(A31>0,1,0)

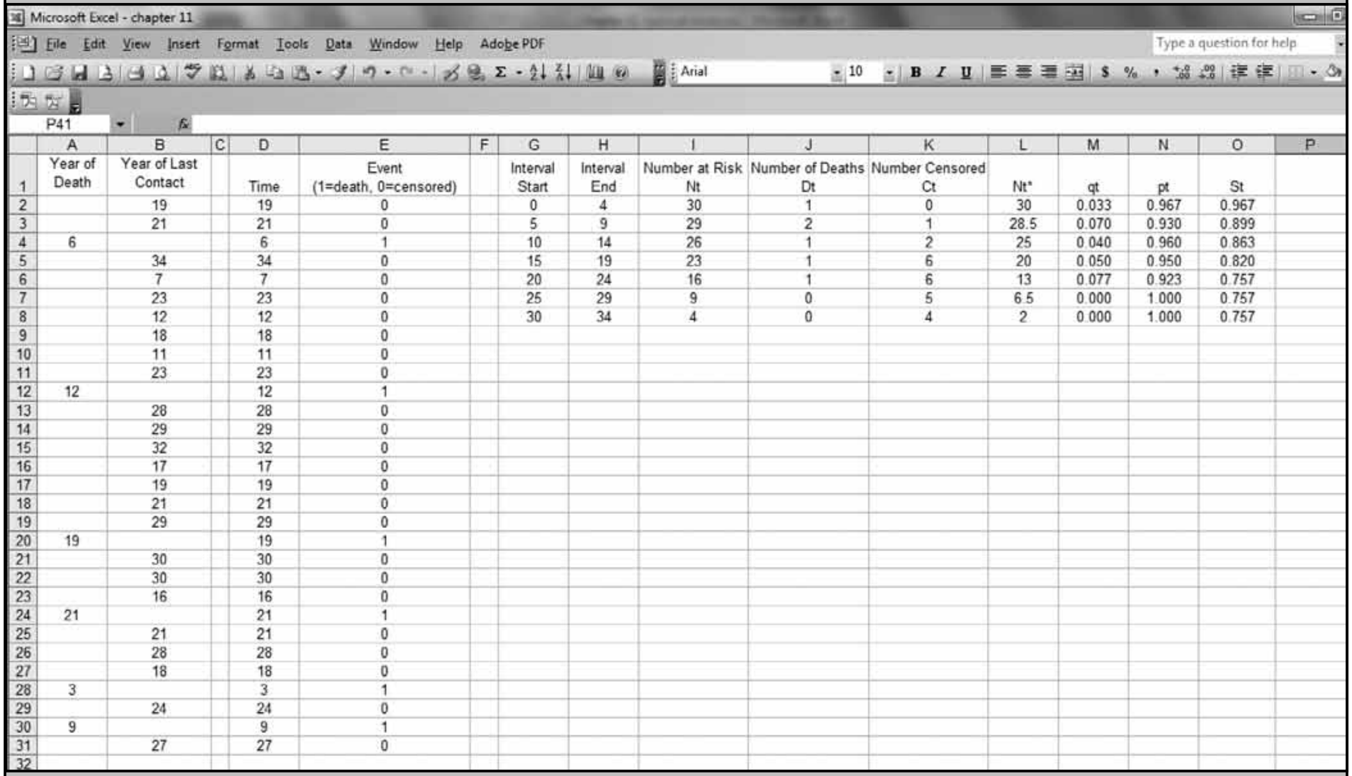
	A	B	C	D	E
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)
2		19		19	0
3		21		21	0
4	6			6	1
5		34		34	0
6		7		7	0
7		23		23	0
8		12		12	0
9		18		18	0
10		11		11	0
11		23		23	0
12	12			12	1
13		28		28	0
14		29		29	0
15		32		32	0
16		17		17	0
17		19		19	0
18		21		21	0
19		29		29	0
20	19			19	1
21		30		30	0
22		30		30	0
23		16		16	0
24	21			21	1
25		21		21	0
26		28		28	0
27		18		18	0
28	3			3	1
29		24		24	0
30	9			9	1
31		27		27	0
32					

Notice that we copied the formulas from cells D21 and E21 into cells D22 and E22 through cells D31 and E31 (see Figure 11-3) to accommodate the larger sample size. Because the follow-up times extend to 34 years, we need to add two additional time intervals to our life table. Specifically, we add 25–29 and 30–34 years to cells G7–H7 and G8–H8, respectively. In addition, we must update the data range from 21 rows to 31 rows for the computations of N_t, D_t, and C_t in columns I, J, and K, respectively (see Figure 11-5 through Figure 11-7). Finally, we copy the formulas to compute N_{t^*}, q_t, p_t, and S_t in cell L6 through cell O6 into cell L7 through cell O8. The life table for the data in Example 11.2 is shown in Figure 11-13.

We now illustrate the Kaplan–Meier approach to estimate survival probabilities using Excel.

Example 11.3. Consider again Example 11.1 (Example 11.2 in the textbook), where we analyzed a small prospective cohort study with death as the primary outcome. The study involved 20 participants ($n = 20$) who were 65 years of age and older, who were followed for up to 24 years until they died,

FIGURE 11-13 The Life Table Using Actuarial Approach


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Type a question for help
Arial
10
P41
fx

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Interval Start	Interval End	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Nt*	qt	pt	St	
2		19		19	0		0	4	30	1	0	30	0.033	0.967	0.967	
3		21		21	0		5	9	29	2	1	28.5	0.070	0.930	0.899	
4	6			6	1		10	14	26	1	2	25	0.040	0.960	0.863	
5		34		34	0		15	19	23	1	6	20	0.050	0.950	0.820	
6		7		7	0		20	24	16	1	6	13	0.077	0.923	0.757	
7		23		23	0		25	29	9	0	5	6.5	0.000	1.000	0.757	
8		12		12	0		30	34	4	0	4	2	0.000	1.000	0.757	
9		18		18	0											
10		11		11	0											
11		23		23	0											
12	12			12	1											
13		28		28	0											
14		29		29	0											
15		32		32	0											
16		17		17	0											
17		19		19	0											
18		21		21	0											
19		29		29	0											
20	19			19	1											
21		30		30	0											
22		30		30	0											
23		16		16	0											
24	21			21	1											
25		21		21	0											
26		28		28	0											
27		18		18	0											
28	3			3	1											
29		24		24	0											
30	9			9	1											
31		27		27	0											
32																

until the study ended, or until they dropped out of the study (were lost to follow-up). The data are shown in Table 11-1.

We now use Excel to construct a life table using the Kaplan–Meier approach. The data were entered into Excel as shown in Figure 11-1 and the two new variables, time and an indicator of event status (event) were created as shown in Figure 11-2 and Figure 11-3, respectively.

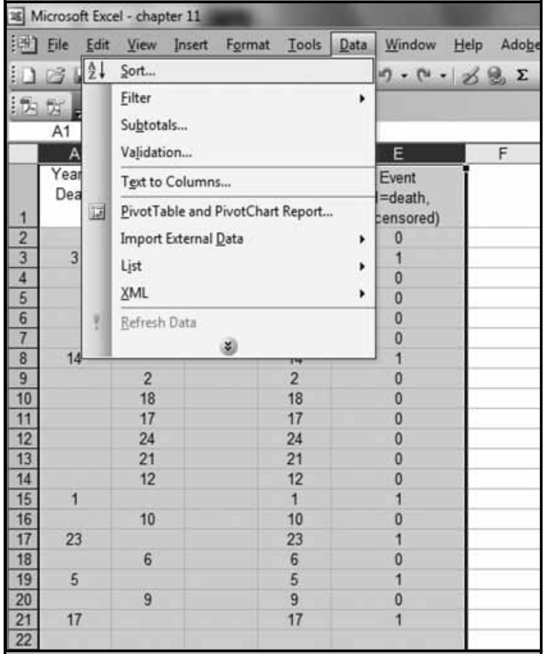
The first step is to sort the data by the time variable (column D). This is done by highlighting the data (column A through column E) and choosing **Sort** under the **Data option** from the top menu bar, as shown in Figure 11-14.

Once we select the **Sort** option, Excel presents the dialog box shown in Figure 11-15. We specify that the primary sort variable is time and that we want the data sorted by time in ascending order. We specify that we then wish to sort the data by the variable event and that we want the data also sorted by event in descending order. This is shown in Figure 11-15. Notice that we also indicate that there is a Header row in the dataset (option at bottom of dialog box). When we indicate that there is a Header row, Excel unselects the first row. When we click OK, Excel produces the data shown in Figure 11-16.

Using the sorted data, in particular the time and event variables, we now construct the life table using the Kaplan–Meier approach. Specifically, we compute the number at risk, N_t; the number of deaths, D_t; the number censored, C_t; and the survival probability, S_t, for each time. First, we compute the number at risk, N_t, using the COUNTIF function. Specifically, for each time, we count the number of participants whose observed time is greater than or equal to the current time. For example, at time 1 (1 year), there are 20 participants at risk (i.e., 20 participants who survive one year or more). To compute the number at risk at the first observed time, we enter the following into cell G2: "=COUNTIF(D$2:D$21">="&D2)". We then copy this formula from cell G2 into cell G3 through cell G21 as shown in Figure 11-17.

Next we compute the number of events (deaths) at each time, D_t. This is done using the IF function. Specifically, at each time, if the event (column E) occurs (event = 1), then we count that event. If the event does not occur (event = 0), we do not. In cell H2 we enter "=IF(E2=1,1,0)". Recall that the IF function checks the specified criterion (in this case whether E2 = 1). Here, if the criterion is met then a "1" is placed in cell

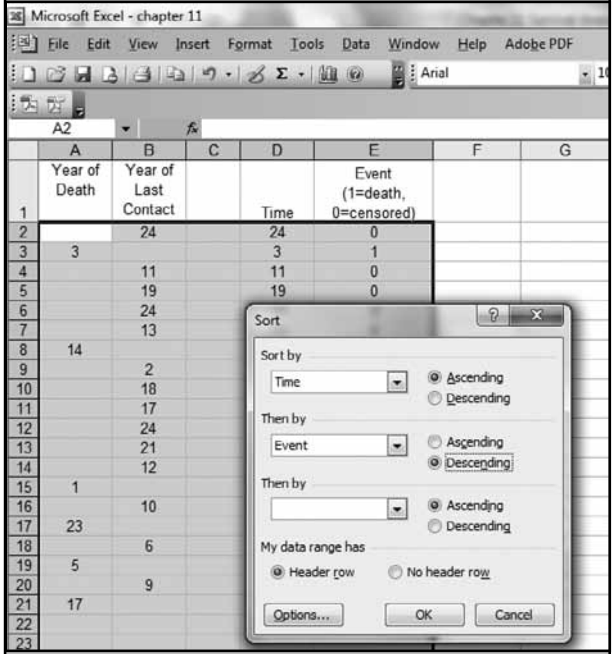
FIGURE 11-14 Sorting the Data by Time and Event


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe
Sort...
Filter
Subtotals...
Validation...
Text to Columns...
PivotTable and PivotChart Report...
Import External Data
List
XML
Refresh Data

A1

	A	B	C	D	E	F
1	Year Dea				Event =death, censored)	
2					0	
3	3				1	
4					0	
5					0	
6					0	
7					0	
8	14			14	1	
9		2		2	0	
10		18		18	0	
11		17		17	0	
12		24		24	0	
13		21		21	0	
14		12		12	0	
15	1			1	1	
16		10		10	0	
17	23			23	1	
18		6		6	0	
19	5			5	1	
20		9		9	0	
21	17			17	1	
22						

FIGURE 11-15 Sorting by Time in Ascending Order


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
Sort
Options...
OK
Cancel

A2

	A	B	C	D	E	F	G
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		
2		24		24	0		
3	3			3	1		
4		11		11	0		
5		19		19	0		
6		24					
7		13					
8	14						
9		2					
10		18					
11		17					
12		24					
13		21					
14		12					
15	1						
16		10					
17	23						
18		6					
19	5						
20		9					
21	17						
22							
23							

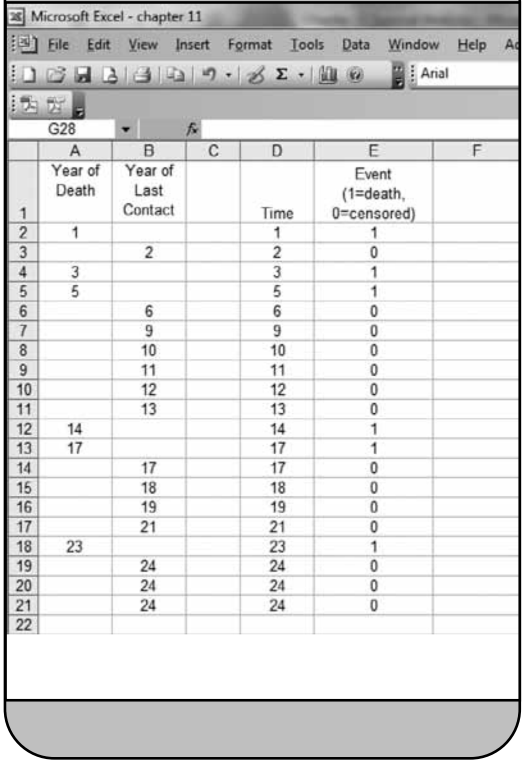
Sort by: Time ● Ascending ○ Descending

Then by: Event ○ Ascending ● Descending

Then by: ● Ascending ○ Descending

My data range has: ● Header row ○ No header row

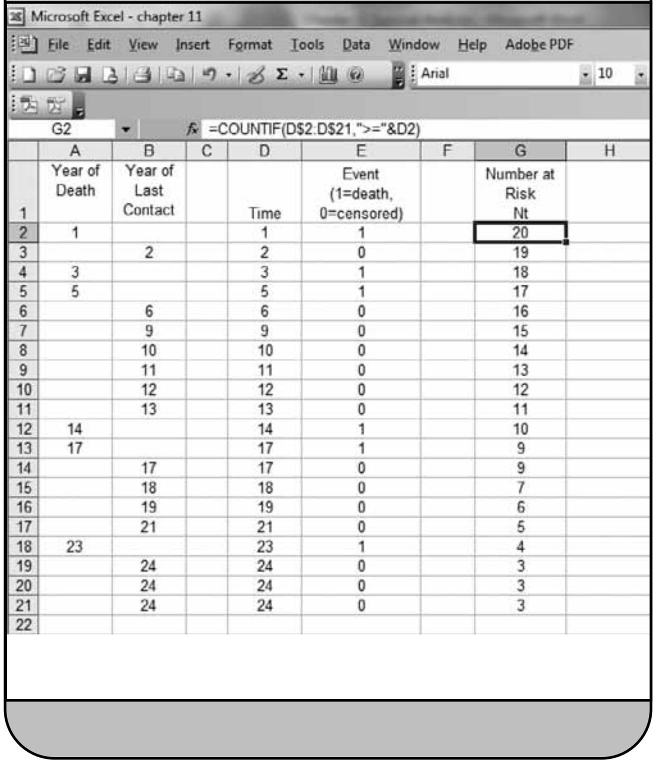
FIGURE 11-16 Sorted Data


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Ac
Arial

G28

	A	B	C	D	E	F
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)	
2	1			1	1	
3		2		2	0	
4	3			3	1	
5	5			5	1	
6		6		6	0	
7		9		9	0	
8		10		10	0	
9		11		11	0	
10		12		12	0	
11		13		13	0	
12	14			14	1	
13	17			17	1	
14		17		17	0	
15		18		18	0	
16		19		19	0	
17		21		21	0	
18	23			23	1	
19		24		24	0	
20		24		24	0	
21		24		24	0	
22						

FIGURE 11-17 Computing the Number at Risk, N_t


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10

G2 =COUNTIF(D$2:D$21,">="&D2)

	A	B	C	D	E	F	G	H
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Number at Risk Nt	
2	1			1	1		20	
3		2		2	0		19	
4	3			3	1		18	
5	5			5	1		17	
6		6		6	0		16	
7		9		9	0		15	
8		10		10	0		14	
9		11		11	0		13	
10		12		12	0		12	
11		13		13	0		11	
12	14			14	1		10	
13	17			17	1		9	
14		17		17	0		9	
15		18		18	0		7	
16		19		19	0		6	
17		21		21	0		5	
18	23			23	1		4	
19		24		24	0		3	
20		24		24	0		3	
21		24		24	0		3	
22								

H2, otherwise a "0" is placed in cell H2. This is shown in Figure 11-18, and the formula is copied from cell H2 into cell H3 through cell H21.

Next we compute the number of participants who are censored at each time, C_t. This is again done using the IF function. Specifically, at each time, if the event (column E) is coded 0 (censored) then we count that event as censored. For example, in cell I2 we enter "= IF(E2=0,1,0)". This is shown in Figure 11-19, and the formula is copied from cell I2 into cell I3 through cell I21.

The final step is to compute the survival probabilities using $S_{t+1} = S_t \times ((N_{t+1} - D_{t+1})/N_{t+1})$. Before computing the survival probabilities at each time, we first insert a row to represent the start of the study (i.e., baseline or time = 0) using the Insert Row option on the top menu bar. In cell D2 we enter 0 (i.e., time = 0), and in cell J2 we enter 1 (i.e., survival probability = 1, $S_0 = 1$). The survival probabilities for each subsequent time are then computing using the formula $S_{t+1} = S_t \times ((N_{t+1} - D_{t+1})/N_{t+1})$, as shown in Figure 11-20 (see formula in the top menu bar). Notice that the survival probabilities only change when there are observed events (in this example, deaths). Censored times do not affect the estimates of the survival probabilities.

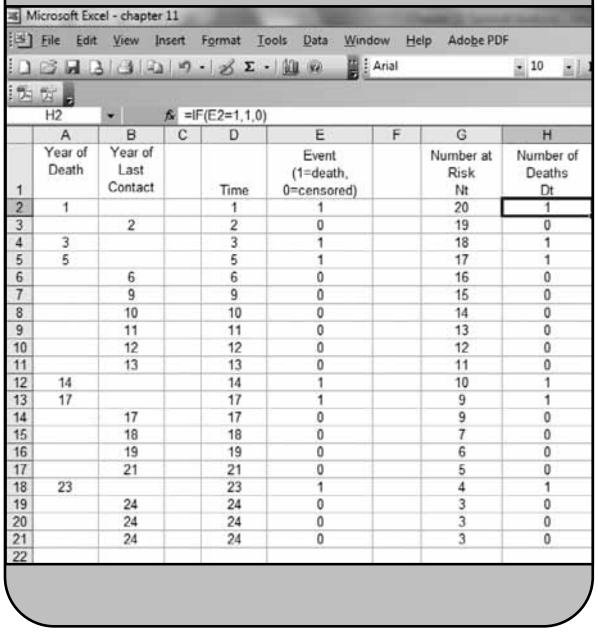
Again, while it is tedious to construct the life table using the Kaplan–Meier approach the first time, once the structure is entered into Excel, construction of life tables for new datasets is relatively easy. The new data need to be entered and sorted, and only the data range in the formulas needs to be modified.

In Chapter 11 of the textbook, we also computed standard errors of the survival estimates. Several formulas are used to produce standard errors; we discussed Greenwood's formula, $\text{SE}(S_t) = S_t\sqrt{\sum \frac{D_t}{N_t(N_t - D_t)}}$, where the quantity where the quantity $\frac{D_t}{N_t(N_t - D_t)}$ is summed for numbers at risk (N_t) and numbers of deaths (D_t) occurring through the time of interest (i.e., cumulative, across all times before the time of interest). In Example 11.4 we use Excel to compute standard errors for the survival estimates for the data in Example 11.3. We also produce 95% confidence intervals for the survival probabilities using $S_t \pm 1.96 \times \text{SE}(S_t)$.

Example 11.4. Consider again Example 11.3, where we analyzed a small prospective cohort study with death as the primary outcome. The study involved 20 participants ($n = 20$) who were 65 years of age and older who were followed for up to 24 years until they died, until the study ended, or until they dropped out of the study (i.e., were lost to follow-up). Using Excel, we estimated the survival function as shown in Figure 11-20. We now use Excel to compute standard errors and 95% confidence limits for the estimates of the survival probabilities.

Before we begin to add the computations, we first hide column A through column D of the spreadsheet to allow for better visualization of the additional columns needed for the computations of the standard errors and confidence limits.

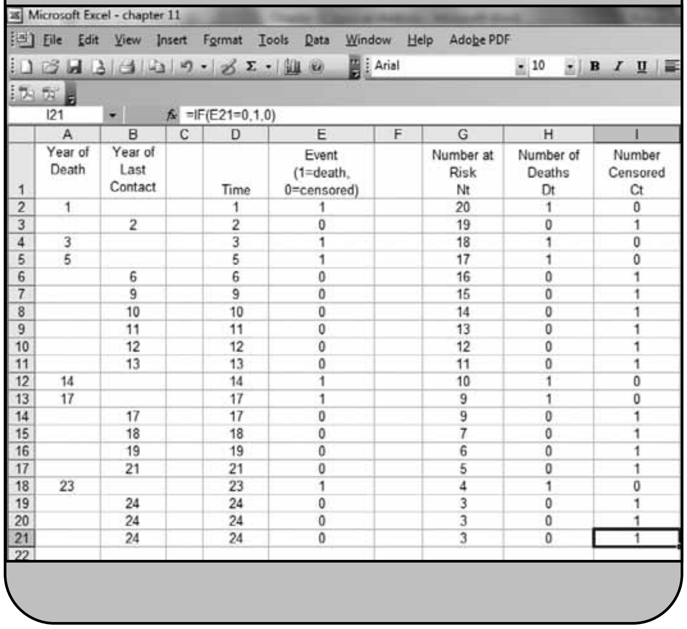
FIGURE 11-18 Computing the Number of Events (Deaths), D_t


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10

H2 =IF(E2=1,1,0)

	A	B	C	D	E	F	G	H
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Number at Risk Nt	Number of Deaths Dt
2	1			1	1		20	1
3		2		2	0		19	0
4	3			3	1		18	1
5	5			5	1		17	1
6		6		6	0		16	0
7		9		9	0		15	0
8		10		10	0		14	0
9		11		11	0		13	0
10		12		12	0		12	0
11		13		13	0		11	0
12	14			14	1		10	1
13	17			17	1		9	1
14		17		17	0		9	0
15		18		18	0		7	0
16		19		19	0		6	0
17		21		21	0		5	0
18	23			23	1		4	1
19		24		24	0		3	0
20		24		24	0		3	0
21		24		24	0		3	0
22								

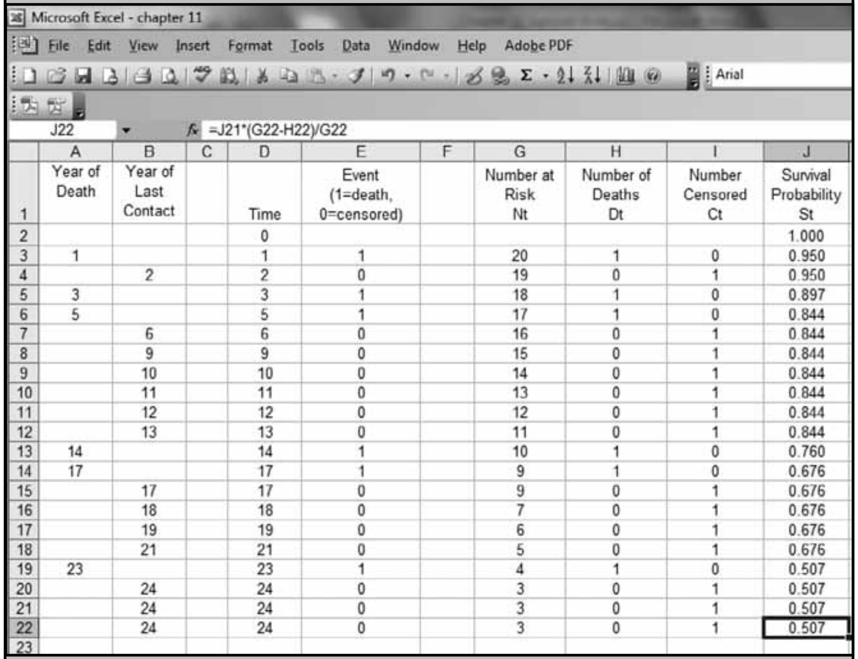
FIGURE 11-19 Computing the Number Censored, C_t


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B
I
U

I21 =IF(E21=0,1,0)

	A	B	C	D	E	F	G	H	I
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Number at Risk Nt	Number of Deaths Dt	Number Censored Ct
2	1			1	1		20	1	0
3		2		2	0		19	0	1
4	3			3	1		18	1	0
5	5			5	1		17	1	0
6		6		6	0		16	0	1
7		9		9	0		15	0	1
8		10		10	0		14	0	1
9		11		11	0		13	0	1
10		12		12	0		12	0	1
11		13		13	0		11	0	1
12	14			14	1		10	1	0
13	17			17	1		9	1	0
14		17		17	0		9	0	1
15		18		18	0		7	0	1
16		19		19	0		6	0	1
17		21		21	0		5	0	1
18	23			23	1		4	1	0
19		24		24	0		3	0	1
20		24		24	0		3	0	1
21		24		24	0		3	0	1
22									

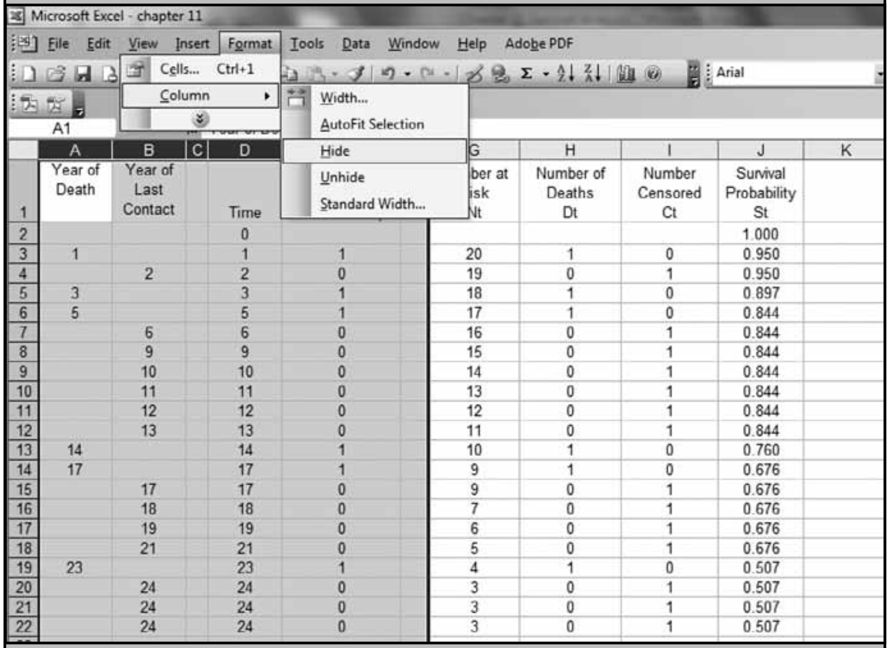
FIGURE 11-20 Computing the Survival Probabilities, S_t


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial

J22 =J21*(G22-H22)/G22

	A	B	C	D	E	F	G	H	I	J
1	Year of Death	Year of Last Contact		Time	Event (1=death, 0=censored)		Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St
2				0						1.000
3	1			1	1		20	1	0	0.950
4		2		2	0		19	0	1	0.950
5	3			3	1		18	1	0	0.897
6	5			5	1		17	1	0	0.844
7		6		6	0		16	0	1	0.844
8		9		9	0		15	0	1	0.844
9		10		10	0		14	0	1	0.844
10		11		11	0		13	0	1	0.844
11		12		12	0		12	0	1	0.844
12		13		13	0		11	0	1	0.844
13	14			14	1		10	1	0	0.760
14	17			17	1		9	1	0	0.676
15		17		17	0		9	0	1	0.676
16		18		18	0		7	0	1	0.676
17		19		19	0		6	0	1	0.676
18		21		21	0		5	0	1	0.676
19	23			23	1		4	1	0	0.507
20		24		24	0		3	0	1	0.507
21		24		24	0		3	0	1	0.507
22		24		24	0		3	0	1	0.507
23										

FIGURE 11-21 Hiding Columns


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Cells... Ctrl+1
Column
Width...
AutoFit Selection
Hide
Unhide
Standard Width...
Arial
A1

	A	B	C	D		G	H	I	J	K
1	Year of Death	Year of Last Contact		Time		ber at isk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St	
2				0					1.000	
3	1			1	1	20	1	0	0.950	
4		2		2	0	19	0	1	0.950	
5	3			3	1	18	1	0	0.897	
6	5			5	1	17	1	0	0.844	
7		6		6	0	16	0	1	0.844	
8		9		9	0	15	0	1	0.844	
9		10		10	0	14	0	1	0.844	
10		11		11	0	13	0	1	0.844	
11		12		12	0	12	0	1	0.844	
12		13		13	0	11	0	1	0.844	
13	14			14	1	10	1	0	0.760	
14	17			17	1	9	1	0	0.676	
15		17		17	0	9	0	1	0.676	
16		18		18	0	7	0	1	0.676	
17		19		19	0	6	0	1	0.676	
18		21		21	0	5	0	1	0.676	
19	23			23	1	4	1	0	0.507	
20		24		24	0	3	0	1	0.507	
21		24		24	0	3	0	1	0.507	
22		24		24	0	3	0	1	0.507	

This is done using the **Hide Column** option under **Format** on the top menu bar as shown in Figure 11-21.

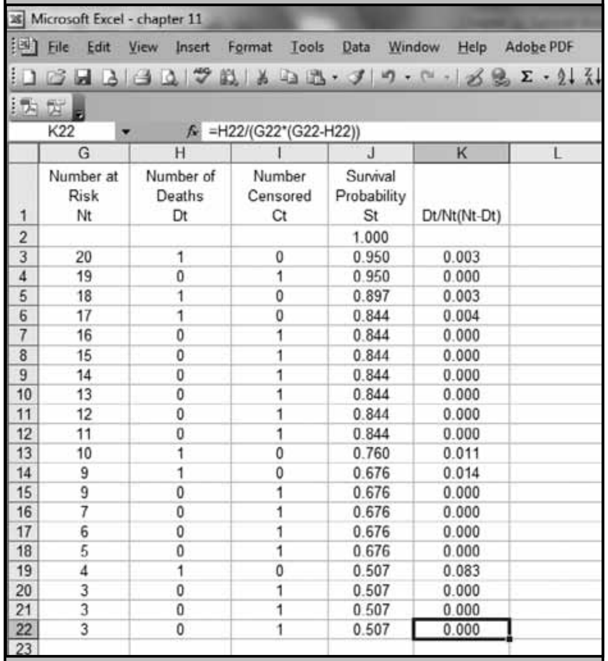
We now create the quantity $D_t/N_t(N_t - D_t)$ at each time as shown in Figure 11-22 (see formula in cell K22 showing in the top menu bar).

Next we sum the quantities $\frac{D_t}{N_t(N_t - D_t)}$ in column K through each time of interest (i.e., cumulative, across all times before the time of interest). This is done using the SUM function. For example, in cell L3 we enter "=K3". In cell L4, we enter "=K4+L3". We then copy the formula in cell L4 into cell L5 through cell L22, as shown in Figure 11-23.

The next step is to compute the standard error for each time $\text{SE}(S_t) = S_t\sqrt{\sum \frac{D_t}{N_t(N_t - D_t)}}$. Specifically, we multiply the survival probability (column J) by the square root of the sum of the quantities $\sum \frac{D_t}{N_t(N_t - D_t)}$ (column L) to produce the standard errors at each time. This is shown in Figure 11-24.

The next step is to compute the margins of error for the 95% confidence intervals using 1.96 × $\text{SE}(S_t)$. This is shown in Figure 11-25.

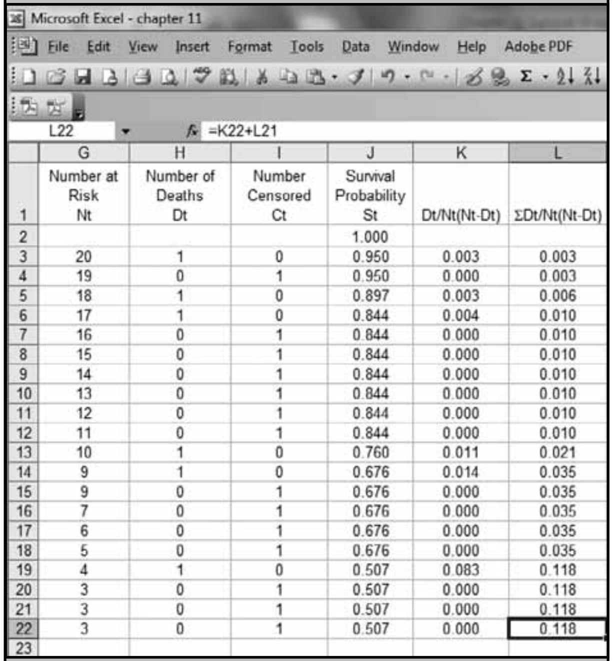
FIGURE 11-22 Computing the Standard Errors


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF

K22 =H22/(G22*(G22-H22))

	G	H	I	J	K	L
1	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St	Dt/Nt(Nt-Dt)	
2				1.000		
3	20	1	0	0.950	0.003	
4	19	0	1	0.950	0.000	
5	18	1	0	0.897	0.003	
6	17	1	0	0.844	0.004	
7	16	0	1	0.844	0.000	
8	15	0	1	0.844	0.000	
9	14	0	1	0.844	0.000	
10	13	0	1	0.844	0.000	
11	12	0	1	0.844	0.000	
12	11	0	1	0.844	0.000	
13	10	1	0	0.760	0.011	
14	9	1	0	0.676	0.014	
15	9	0	1	0.676	0.000	
16	7	0	1	0.676	0.000	
17	6	0	1	0.676	0.000	
18	5	0	1	0.676	0.000	
19	4	1	0	0.507	0.083	
20	3	0	1	0.507	0.000	
21	3	0	1	0.507	0.000	
22	3	0	1	0.507	0.000	
23						

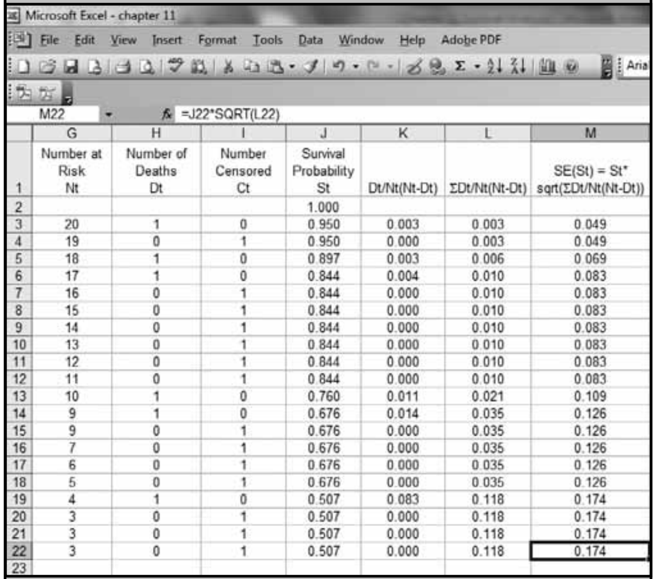
FIGURE 11-23 Computing the Standard Errors


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF

L22 =K22+L21

	G	H	I	J	K	L
1	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St	Dt/Nt(Nt-Dt)	ΣDt/Nt(Nt-Dt)
2				1.000		
3	20	1	0	0.950	0.003	0.003
4	19	0	1	0.950	0.000	0.003
5	18	1	0	0.897	0.003	0.006
6	17	1	0	0.844	0.004	0.010
7	16	0	1	0.844	0.000	0.010
8	15	0	1	0.844	0.000	0.010
9	14	0	1	0.844	0.000	0.010
10	13	0	1	0.844	0.000	0.010
11	12	0	1	0.844	0.000	0.010
12	11	0	1	0.844	0.000	0.010
13	10	1	0	0.760	0.011	0.021
14	9	1	0	0.676	0.014	0.035
15	9	0	1	0.676	0.000	0.035
16	7	0	1	0.676	0.000	0.035
17	6	0	1	0.676	0.000	0.035
18	5	0	1	0.676	0.000	0.035
19	4	1	0	0.507	0.083	0.118
20	3	0	1	0.507	0.000	0.118
21	3	0	1	0.507	0.000	0.118
22	3	0	1	0.507	0.000	0.118
23						

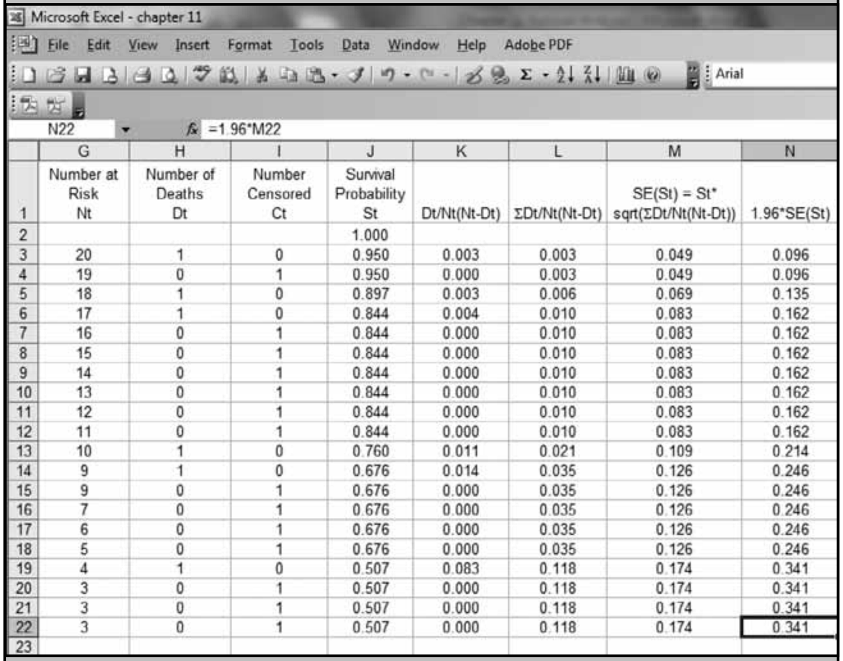
FIGURE 11-24 Computing the Standard Errors


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial

M22 =J22*SQRT(L22)

	G	H	I	J	K	L	M
1	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St	Dt/Nt(Nt-Dt)	ΣDt/Nt(Nt-Dt)	SE(St) = St* sqrt(ΣDt/Nt(Nt-Dt))
2				1.000			
3	20	1	0	0.950	0.003	0.003	0.049
4	19	0	1	0.950	0.000	0.003	0.049
5	18	1	0	0.897	0.003	0.006	0.069
6	17	1	0	0.844	0.004	0.010	0.083
7	16	0	1	0.844	0.000	0.010	0.083
8	15	0	1	0.844	0.000	0.010	0.083
9	14	0	1	0.844	0.000	0.010	0.083
10	13	0	1	0.844	0.000	0.010	0.083
11	12	0	1	0.844	0.000	0.010	0.083
12	11	0	1	0.844	0.000	0.010	0.083
13	10	1	0	0.760	0.011	0.021	0.109
14	9	1	0	0.676	0.014	0.035	0.126
15	9	0	1	0.676	0.000	0.035	0.126
16	7	0	1	0.676	0.000	0.035	0.126
17	6	0	1	0.676	0.000	0.035	0.126
18	5	0	1	0.676	0.000	0.035	0.126
19	4	1	0	0.507	0.083	0.118	0.174
20	3	0	1	0.507	0.000	0.118	0.174
21	3	0	1	0.507	0.000	0.118	0.174
22	3	0	1	0.507	0.000	0.118	0.174
23							

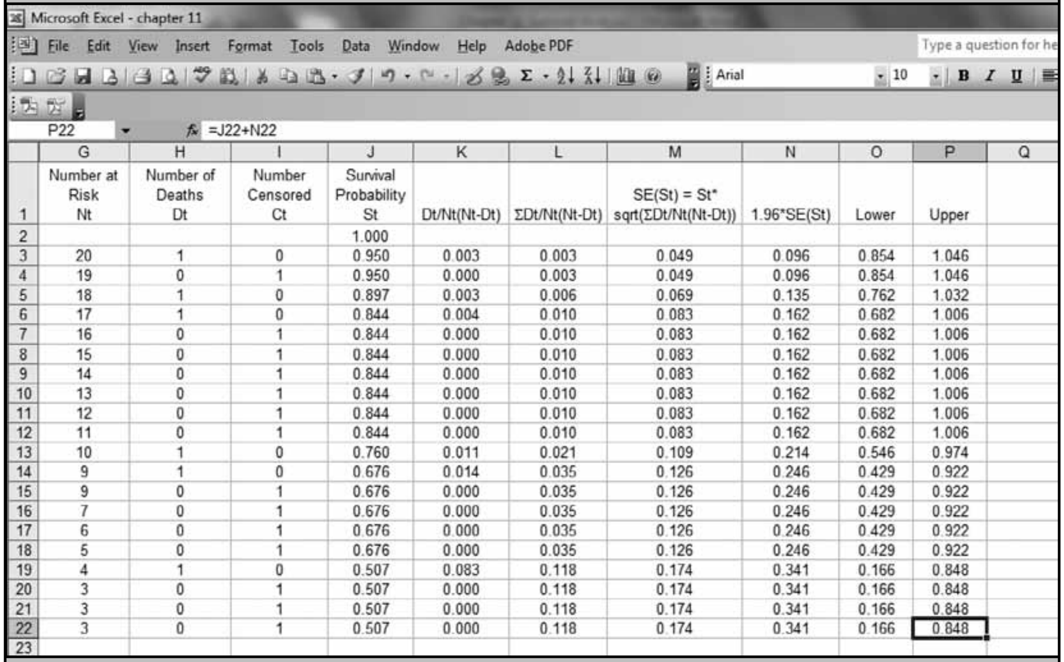
FIGURE 11-25 Computing the Margins of Error


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial

N22 =1.96*M22

	G	H	I	J	K	L	M	N
1	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St	Dt/Nt(Nt-Dt)	ΣDt/Nt(Nt-Dt)	SE(St) = St* sqrt(ΣDt/Nt(Nt-Dt))	1.96*SE(St)
2				1.000				
3	20	1	0	0.950	0.003	0.003	0.049	0.096
4	19	0	1	0.950	0.000	0.003	0.049	0.096
5	18	1	0	0.897	0.003	0.006	0.069	0.135
6	17	1	0	0.844	0.004	0.010	0.083	0.162
7	16	0	1	0.844	0.000	0.010	0.083	0.162
8	15	0	1	0.844	0.000	0.010	0.083	0.162
9	14	0	1	0.844	0.000	0.010	0.083	0.162
10	13	0	1	0.844	0.000	0.010	0.083	0.162
11	12	0	1	0.844	0.000	0.010	0.083	0.162
12	11	0	1	0.844	0.000	0.010	0.083	0.162
13	10	1	0	0.760	0.011	0.021	0.109	0.214
14	9	1	0	0.676	0.014	0.035	0.126	0.246
15	9	0	1	0.676	0.000	0.035	0.126	0.246
16	7	0	1	0.676	0.000	0.035	0.126	0.246
17	6	0	1	0.676	0.000	0.035	0.126	0.246
18	5	0	1	0.676	0.000	0.035	0.126	0.246
19	4	1	0	0.507	0.083	0.118	0.174	0.341
20	3	0	1	0.507	0.000	0.118	0.174	0.341
21	3	0	1	0.507	0.000	0.118	0.174	0.341
22	3	0	1	0.507	0.000	0.118	0.174	0.341
23								

FIGURE 11-26 95% Confidence Limits for the Survival Probabilities


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Type a question for he
Arial
10
B I U

P22 =J22+N22

	G	H	I	J	K	L	M	N	O	P	Q
1	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St	Dt/Nt(Nt-Dt)	ΣDt/Nt(Nt-Dt)	SE(St) = St* sqrt(ΣDt/Nt(Nt-Dt))	1.96*SE(St)	Lower	Upper	
2				1.000							
3	20	1	0	0.950	0.003	0.003	0.049	0.096	0.854	1.046	
4	19	0	1	0.950	0.000	0.003	0.049	0.096	0.854	1.046	
5	18	1	0	0.897	0.003	0.006	0.069	0.135	0.762	1.032	
6	17	1	0	0.844	0.004	0.010	0.083	0.162	0.682	1.006	
7	16	0	1	0.844	0.000	0.010	0.083	0.162	0.682	1.006	
8	15	0	1	0.844	0.000	0.010	0.083	0.162	0.682	1.006	
9	14	0	1	0.844	0.000	0.010	0.083	0.162	0.682	1.006	
10	13	0	1	0.844	0.000	0.010	0.083	0.162	0.682	1.006	
11	12	0	1	0.844	0.000	0.010	0.083	0.162	0.682	1.006	
12	11	0	1	0.844	0.000	0.010	0.083	0.162	0.682	1.006	
13	10	1	0	0.760	0.011	0.021	0.109	0.214	0.546	0.974	
14	9	1	0	0.676	0.014	0.035	0.126	0.246	0.429	0.922	
15	9	0	1	0.676	0.000	0.035	0.126	0.246	0.429	0.922	
16	7	0	1	0.676	0.000	0.035	0.126	0.246	0.429	0.922	
17	6	0	1	0.676	0.000	0.035	0.126	0.246	0.429	0.922	
18	5	0	1	0.676	0.000	0.035	0.126	0.246	0.429	0.922	
19	4	1	0	0.507	0.083	0.118	0.174	0.341	0.166	0.848	
20	3	0	1	0.507	0.000	0.118	0.174	0.341	0.166	0.848	
21	3	0	1	0.507	0.000	0.118	0.174	0.341	0.166	0.848	
22	3	0	1	0.507	0.000	0.118	0.174	0.341	0.166	0.848	
23											

The last step is to produce the 95% confidence limits using the formula *point estimate* ± *margin of error*. In this case, we use $S_t \pm 1.96 \times \text{SE}(S_t)$. In Figure 11-26 we create two new variables representing the lower and upper limits of the 95% confidence interval, respectively. The lower limit is computed by subtracting the margin of error (column N) from the survival probability (column J), and the upper limit is computed by adding the margin of error (column N) to the survival probability (column J). The limits are computed in Figure 11-26.

11.2 PLOTTING A SURVIVAL FUNCTION

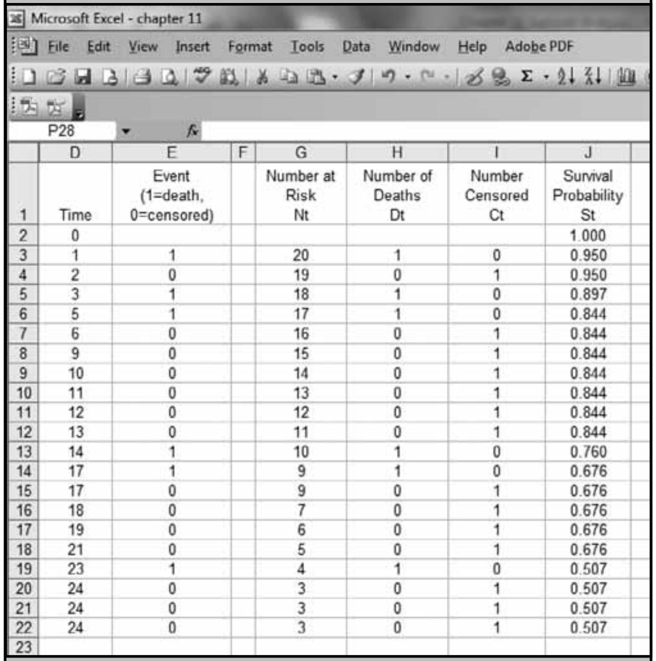
A graphical display of the survival function is a very useful means of reporting or presenting survival information. A graphical display of the Kaplan–Meier survival curve can be produced from the life table we created using Excel.

Example 11.5. Consider again Example 11.3 where we analyzed a small prospective cohort study with death as the primary outcome. The study involved 20 participants ($n = 20$) who were 65 years of age and older who were followed for up to 24 years until they died, until the study ended, or until they dropped out of the study (i.e., were lost to follow-up). Using Excel we estimated the survival function as shown in Figure 11-20. We now use Excel to create a graphical display of the survival function.

In Chapter 11 of the textbook, we presented graphical displays of survival functions which we produced in Excel and showed time along the *x*-axis and the survival probabilities along the *y*-axis. In order to produce the graphical displays, which take the form of step functions, some manipulation of the data is required. Here we detail the steps needed to produce the displays.

The data for the graphical display are shown in Figure 11-27. The key data elements are the time (column D) and the survival probability (column J). Notice that the observed data (entered in column A and column B) as well as a blank column C are hidden.

FIGURE 11-27 Data for the Graphical Display

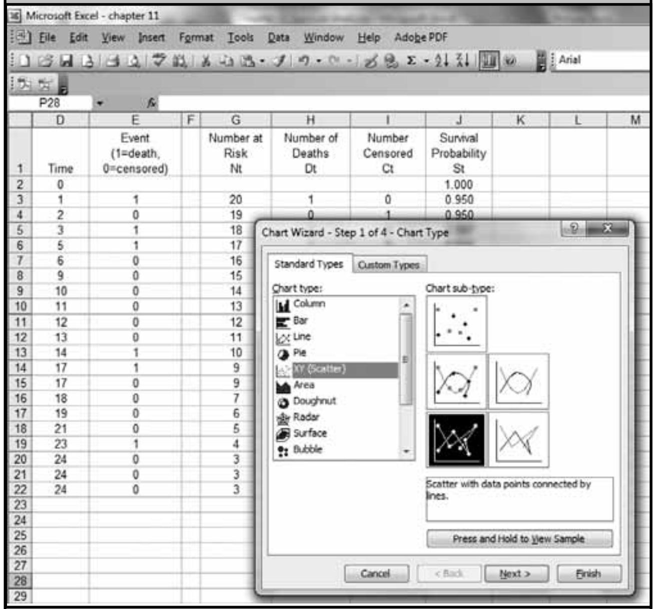


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
P28

	D	E	F	G	H	I	J
1	Time	Event (1=death, 0=censored)		Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St
2	0						1.000
3	1	1		20	1	0	0.950
4	2	0		19	0	1	0.950
5	3	1		18	1	0	0.897
6	5	1		17	1	0	0.844
7	6	0		16	0	1	0.844
8	9	0		15	0	1	0.844
9	10	0		14	0	1	0.844
10	11	0		13	0	1	0.844
11	12	0		12	0	1	0.844
12	13	0		11	0	1	0.844
13	14	1		10	1	0	0.760
14	17	1		9	1	0	0.676
15	17	0		9	0	1	0.676
16	18	0		7	0	1	0.676
17	19	0		6	0	1	0.676
18	21	0		5	0	1	0.676
19	23	1		4	1	0	0.507
20	24	0		3	0	1	0.507
21	24	0		3	0	1	0.507
22	24	0		3	0	1	0.507
23							

Excel has a number of built-in graphical displays which are available using the Chart Wizard (see Chapter 3 of the workbook for examples using the Chart Wizard). We use the Chart Wizard here to produce the displays of the survival function. However, we first need to do some formatting of the data. To illustrate why the formatting is needed, consider the following. Suppose we open the **Chart Wizard** by clicking the chart icon in the top menu bar. This is shown in Figure 11-28.

FIGURE 11-28 Using the Chart Wizard



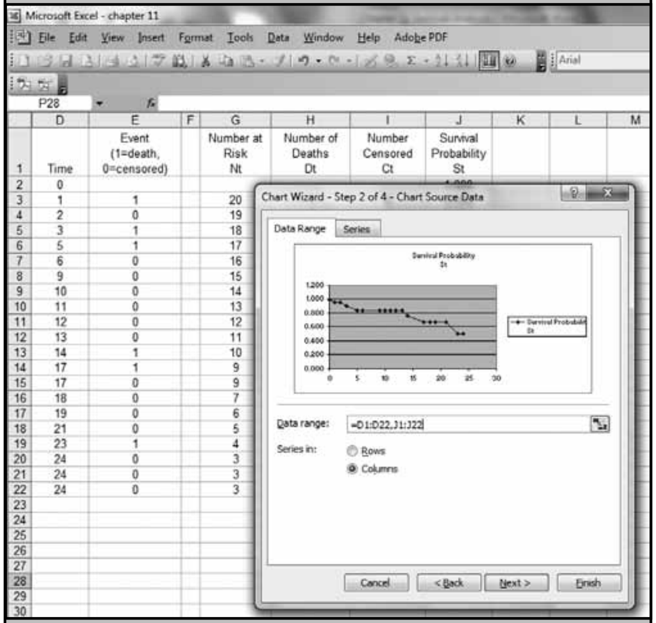
Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
P28
Chart Wizard - Step 1 of 4 - Chart Type
Standard Types
Custom Types
Chart type:
Column
Bar
Line
Pie
XY (Scatter)
Area
Doughnut
Radar
Surface
Bubble
Chart sub-type:
Press and Hold to View Sample
Cancel
< Back
Next >
Finish

Scatter with data points connected by lines.

	D	E	F	G	H	I	J	K	L	M
1	Time	Event (1=death, 0=censored)		Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St			
2	0						1.000			
3	1	1		20	1	0	0.950			
4	2	0		19	0	1	0.950			
5	3	1		18						
6	5	1		17						
7	6	0		16						
8	9	0		15						
9	10	0		14						
10	11	0		13						
11	12	0		12						
12	13	0		11						
13	14	1		10						
14	17	1		9						
15	17	0		9						
16	18	0		7						
17	19	0		6						
18	21	0		5						
19	23	1		4						
20	24	0		3						
21	24	0		3						
22	24	0		3						
23										
24										
25										
26										
27										
28										
29										

Once we select the Chart Wizard, Excel opens a dialog box with various options. We first select the chart type. In this case, we select the **XY (Scatter)** type from the list on the left and then the bottom left option from the chart sub-types. Once we click **Next**, Excel then asks for the range of the data for the display (see Figure 11-29). In the **Data Range** input field, we

FIGURE 11-29 Data for the Display



	D	E	G	H	I	J
1	Time	Event (1=death, 0=censored)	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St
2	0					
3	1	1	20			
4	2	0	19			
5	3	1	18			
6	5	1	17			
7	6	0	16			
8	9	0	15			
9	10	0	14			
10	11	0	13			
11	12	0	12			
12	13	0	11			
13	14	1	10			
14	17	1	9			
15	17	0	9			
16	18	0	7			
17	19	0	6			
18	21	0	5			
19	23	1	4			
20	24	0	3			
21	24	0	3			
22	24	0	3			

Chart Wizard - Step 2 of 4 - Chart Source Data

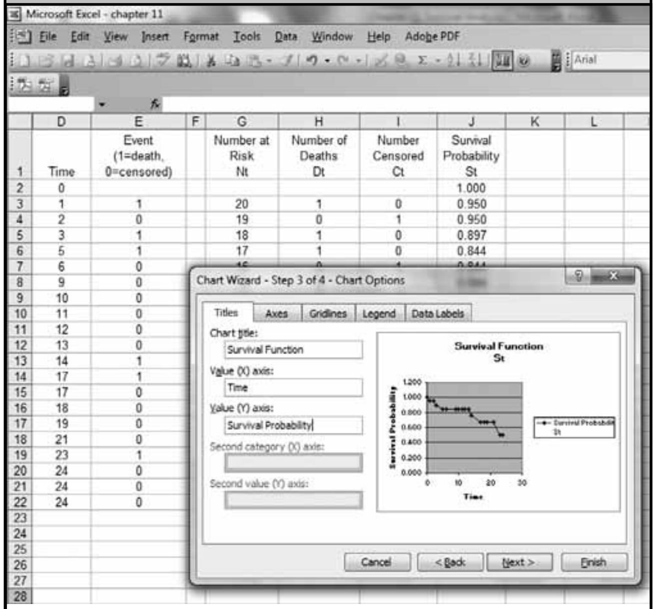
Data Range | Series

Data range: =D1:D22,J1:J22

Series in: ○ Rows ● Columns

Cancel | < Back | Next > | Finish

FIGURE 11-30 Specifying a Title and Labels for the Axes



	D	E	G	H	I	J
1	Time	Event (1=death, 0=censored)	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St
2	0					1.000
3	1	1	20	1	0	0.950
4	2	0	19	0	1	0.950
5	3	1	18	1	0	0.897
6	5	1	17	1	0	0.844
7	6	0				
8	9	0				
9	10	0				
10	11	0				
11	12	0				
12	13	0				
13	14	1				
14	17	1				
15	17	0				
16	18	0				
17	19	0				
18	21	0				
19	23	1				
20	24	0				
21	24	0				
22	24	0				

Chart Wizard - Step 3 of 4 - Chart Options

Titles | Axes | Gridlines | Legend | Data Labels

Chart title: Survival Function

Value (X) axis: Time

Value (Y) axis: Survival Probability

Second category (X) axis:

Second value (Y) axis:

Cancel | < Back | Next > | Finish

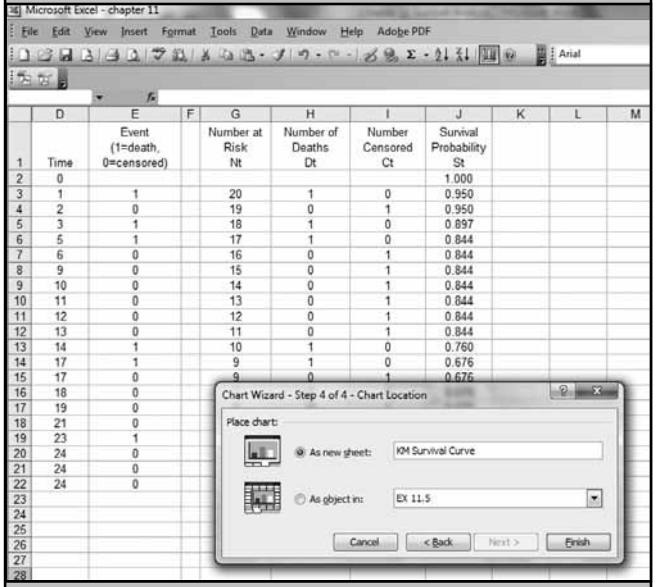
specify "D1:D22,J1:J22". The data for the *x*-axis (time) is contained in column D, and the data for the *y*-axis (survival probability) is contained in column J.

As soon as the data range is entered, Excel shows a preliminary version of the graph in the upper portion of the dialog box. Once we click **Next**, Excel then asks for a chart title as well as labels for the *x*- and *y*-axes (see Figure 11-30).

We enter the chart title and the labels for the *x*- and *y*-axes. When we click **Next**, Excel then asks where we would like to place the display. The options are a new worksheet in the current Excel workbook or an object (an overlay) on the current worksheet. In Figure 11-31 we select a new worksheet and provide a name for the new worksheet. We click **Finish**, and Excel generates the new worksheet containing the display shown in Figure 11-32.

We will make a number of formatting changes in the display prior to presenting it. However, notice that the display is not a step function. The survival curve connects observed changes in survival over time using diagonal lines (i.e., using interpolation). We do not wish to interpolate; instead, we want to show steps, i.e., vertical lines from an observed survival probability connecting to a horizontal line at the next observed survival probability, taking a shape like an "L" at each transition point.

FIGURE 11-31 Specifying a Location for the Display



	D	E	G	H	I	J
1	Time	Event (1=death, 0=censored)	Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St
2	0					1.000
3	1	1	20	1	0	0.950
4	2	0	19	0	1	0.950
5	3	1	18	1	0	0.897
6	5	1	17	1	0	0.844
7	6	0	16	0	1	0.844
8	9	0	15	0	1	0.844
9	10	0	14	0	1	0.844
10	11	0	13	0	1	0.844
11	12	0	12	0	1	0.844
12	13	0	11	0	1	0.844
13	14	1	10	1	0	0.760
14	17	1	9	1	0	0.676
15	17	0	9	0	1	0.676
16	18	0				
17	19	0				
18	21	0				
19	23	1				
20	24	0				
21	24	0				
22	24	0				

Chart Wizard - Step 4 of 4 - Chart Location

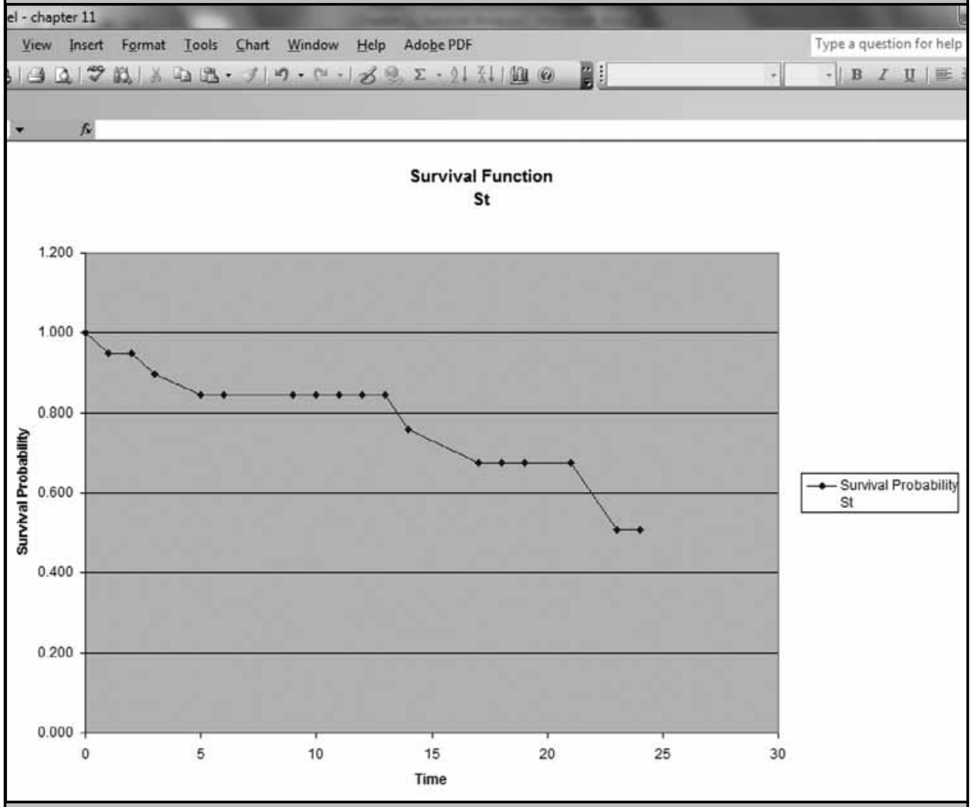
Place chart:

● As new sheet: KM Survival Curve

○ As object in: EX 11.5

Cancel | < Back | Next > | Finish

FIGURE 11-32 The Display



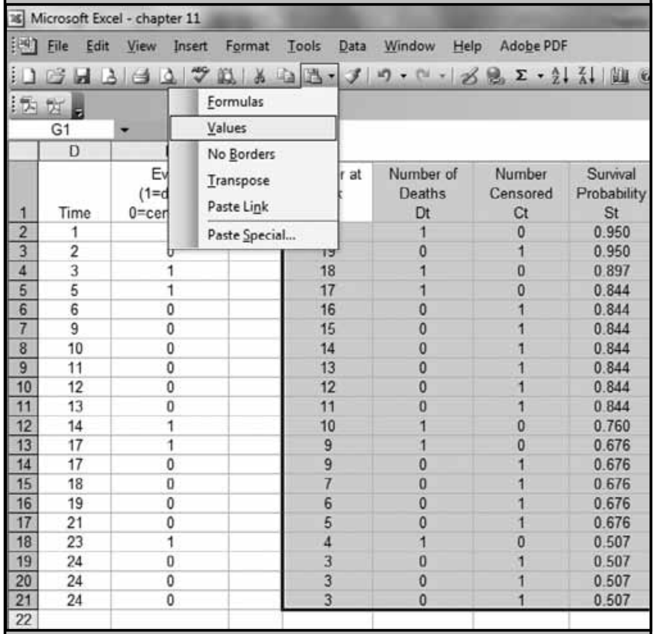
In order to produce a step function, we must format the data before invoking the Chart Wizard as follows. Using the data shown in Figure 11-27, we first must replace the formulas that we used to compute the survival probabilities with their calculated values. We do this is so that, in the process of formatting the data to generate the plot, we do not inadvertently change the estimates of the survival probabilities. Specifically, we want to replace the formulas currently in cell G2 through cell J21 with their computed values. This is done by highlighting these cells and clicking on the copy icon along the top menu bar. This generates a flashing dashed line around the highlighted cells. We then click on the arrow next to the paste icon along the top menu bar and select the **Values** option as shown in Figure 11-33.

The next step is to format the data so that, at each transition point (i.e., change in survival probability), the graphical display shows a step. This is done by inserting a new row in between each transition point (which is equivalent to inserting a row before each observed event (death) in the dataset). This is done by placing the cursor on a specific row and selecting the **Insert Row** option from the top menu bar as shown in Figure 11-34. The resulting worksheet after inserting rows at each transition point is shown in Figure 11-35.

The next step involves inserting data (times and survival probabilities) into the newly inserted rows. Specifically, in each newly inserted (blank) row, we copy the time from the cell above and the survival probability from the cell below. For example, in cell D3 we enter "=D2" (the time from cell D2) and in cell J3 we enter "=J4" (the survival probability from cell J4). The same procedure is followed for rows 6, 8, 16, 18, and 24. For example, in cell D24 we enter "=D23" and in cell J24 we enter "=J25". The updated worksheet is shown in Figure 11-36.

Using the updated worksheet, we follow the same steps illustrated in Figure 11-28 through Figure 11-32. Note that the

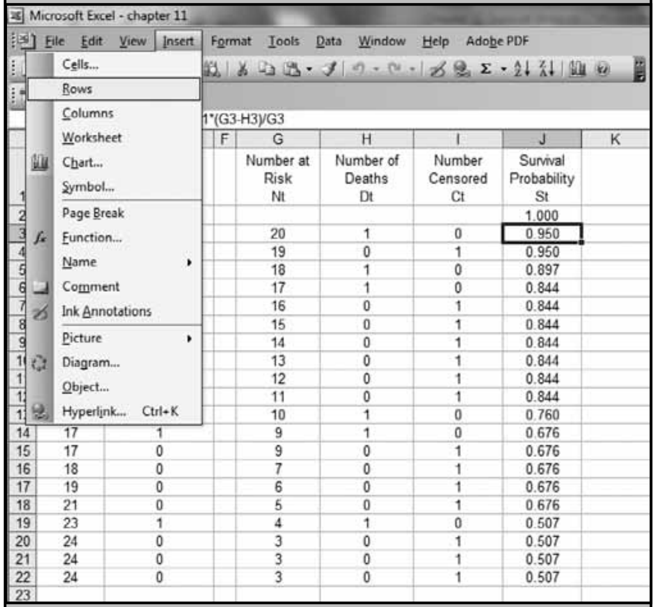
FIGURE 11-33 Replacing Formulas with Their Calculated Values


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Formulas
Values
No Borders
Transpose
Paste Link
Paste Special...

G1

	D	E		G			
1	Time	Ev (1=d 0=cen		r at k	Number of Deaths Dt	Number Censored Ct	Survival Probability St
2	1				1	0	0.950
3	2	0		19	0	1	0.950
4	3	1		18	1	0	0.897
5	5	1		17	1	0	0.844
6	6	0		16	0	1	0.844
7	9	0		15	0	1	0.844
8	10	0		14	0	1	0.844
9	11	0		13	0	1	0.844
10	12	0		12	0	1	0.844
11	13	0		11	0	1	0.844
12	14	1		10	1	0	0.760
13	17	1		9	1	0	0.676
14	17	0		9	0	1	0.676
15	18	0		7	0	1	0.676
16	19	0		6	0	1	0.676
17	21	0		5	0	1	0.676
18	23	1		4	1	0	0.507
19	24	0		3	0	1	0.507
20	24	0		3	0	1	0.507
21	24	0		3	0	1	0.507
22							

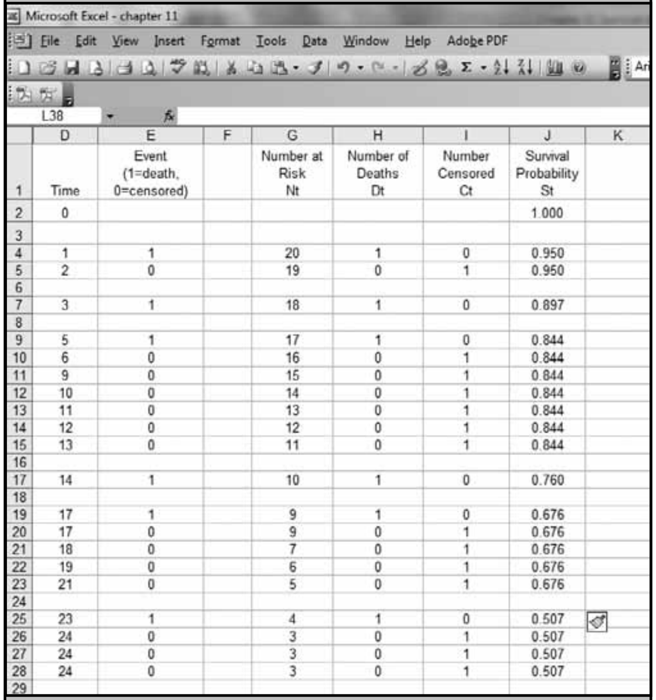
FIGURE 11-34 Inserting a Row



1*(G3-H3)/G3

			F	G	H	I	J	K
1				Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St	
2							1.000	
3				20	1	0	0.950	
4				19	0	1	0.950	
5				18	1	0	0.897	
6				17	1	0	0.844	
7				16	0	1	0.844	
8				15	0	1	0.844	
9				14	0	1	0.844	
1				13	0	1	0.844	
1				12	0	1	0.844	
1				11	0	1	0.844	
1				10	1	0	0.760	
14	17	1		9	1	0	0.676	
15	17	0		9	0	1	0.676	
16	18	0		7	0	1	0.676	
17	19	0		6	0	1	0.676	
18	21	0		5	0	1	0.676	
19	23	1		4	1	0	0.507	
20	24	0		3	0	1	0.507	
21	24	0		3	0	1	0.507	
22	24	0		3	0	1	0.507	
23								

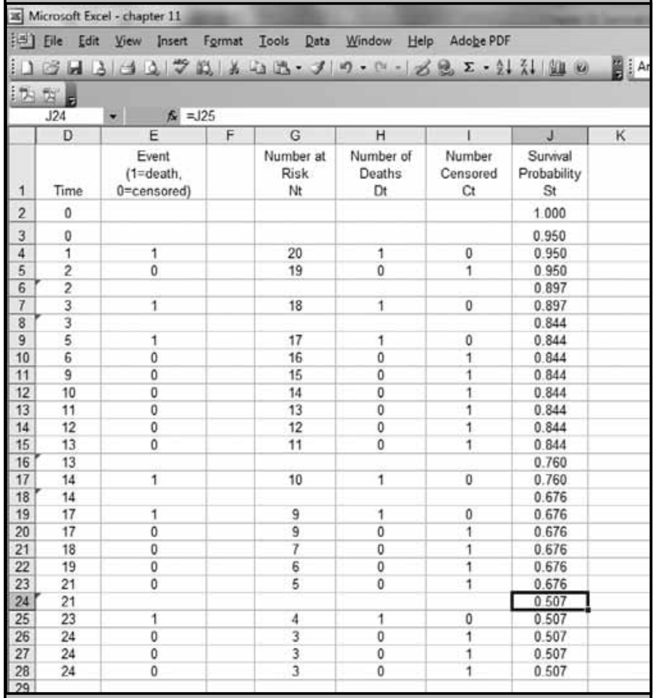
FIGURE 11-35 Inserting Rows at Each Transition Point


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF

L38

	D	E	F	G	H	I	J	K
1	Time	Event (1=death, 0=censored)		Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St	
2	0						1.000	
3								
4	1	1		20	1	0	0.950	
5	2	0		19	0	1	0.950	
6								
7	3	1		18	1	0	0.897	
8								
9	5	1		17	1	0	0.844	
10	6	0		16	0	1	0.844	
11	9	0		15	0	1	0.844	
12	10	0		14	0	1	0.844	
13	11	0		13	0	1	0.844	
14	12	0		12	0	1	0.844	
15	13	0		11	0	1	0.844	
16								
17	14	1		10	1	0	0.760	
18								
19	17	1		9	1	0	0.676	
20	17	0		9	0	1	0.676	
21	18	0		7	0	1	0.676	
22	19	0		6	0	1	0.676	
23	21	0		5	0	1	0.676	
24								
25	23	1		4	1	0	0.507	
26	24	0		3	0	1	0.507	
27	24	0		3	0	1	0.507	
28	24	0		3	0	1	0.507	
29								

FIGURE 11-36 Data to Create a Step Function


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF

J24 =J25

	D	E	F	G	H	I	J	K
1	Time	Event (1=death, 0=censored)		Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St	
2	0						1.000	
3	0						0.950	
4	1	1		20	1	0	0.950	
5	2	0		19	0	1	0.950	
6	2						0.897	
7	3	1		18	1	0	0.897	
8	3						0.844	
9	5	1		17	1	0	0.844	
10	6	0		16	0	1	0.844	
11	9	0		15	0	1	0.844	
12	10	0		14	0	1	0.844	
13	11	0		13	0	1	0.844	
14	12	0		12	0	1	0.844	
15	13	0		11	0	1	0.844	
16	13						0.760	
17	14	1		10	1	0	0.760	
18	14						0.676	
19	17	1		9	1	0	0.676	
20	17	0		9	0	1	0.676	
21	18	0		7	0	1	0.676	
22	19	0		6	0	1	0.676	
23	21	0		5	0	1	0.676	
24	21						0.507	
25	23	1		4	1	0	0.507	
26	24	0		3	0	1	0.507	
27	24	0		3	0	1	0.507	
28	24	0		3	0	1	0.507	
29								

data range is now larger to include the newly inserted rows. The data for the *x*-axis now reside in cell D1 through cell D28, and the data for the *y*-axis now reside in cell J1 through cell J28. Thus, when specifying the data range (see Figure 11-29), we indicate "D1:D28,J1:J28". Following the steps outlined above, we produce the display shown in Figure 11-37.

We now format the display for presentation. First, we change the background from grey to white (or no background color). This is done by double-clicking on any part of the background, which opens up the dialog box shown in Figure 11-38. Under the Area section, we select None.

Next, we remove the legend by clicking on the legend box and clicking **Delete**. We do the same to the horizontal lines in the display. If we click on any of the lines and click **Delete**, all of the horizontal lines are removed. Finally, we format the *y*-axis from 0 to 1. This is done by double-clicking anywhere along the *y*-axis, which brings up the dialog box shown in Figure 11-39. Under the **Scale** tab, we specify the maximum as 1.0 (instead of 1.2, which was the default). Clicking **OK** updates the *y*-axis as shown in Figure 11-40.

The display can now be easily copied into papers, reports, or other presentations. In the survival curve shown in Figure

FIGURE 11-37 The Display

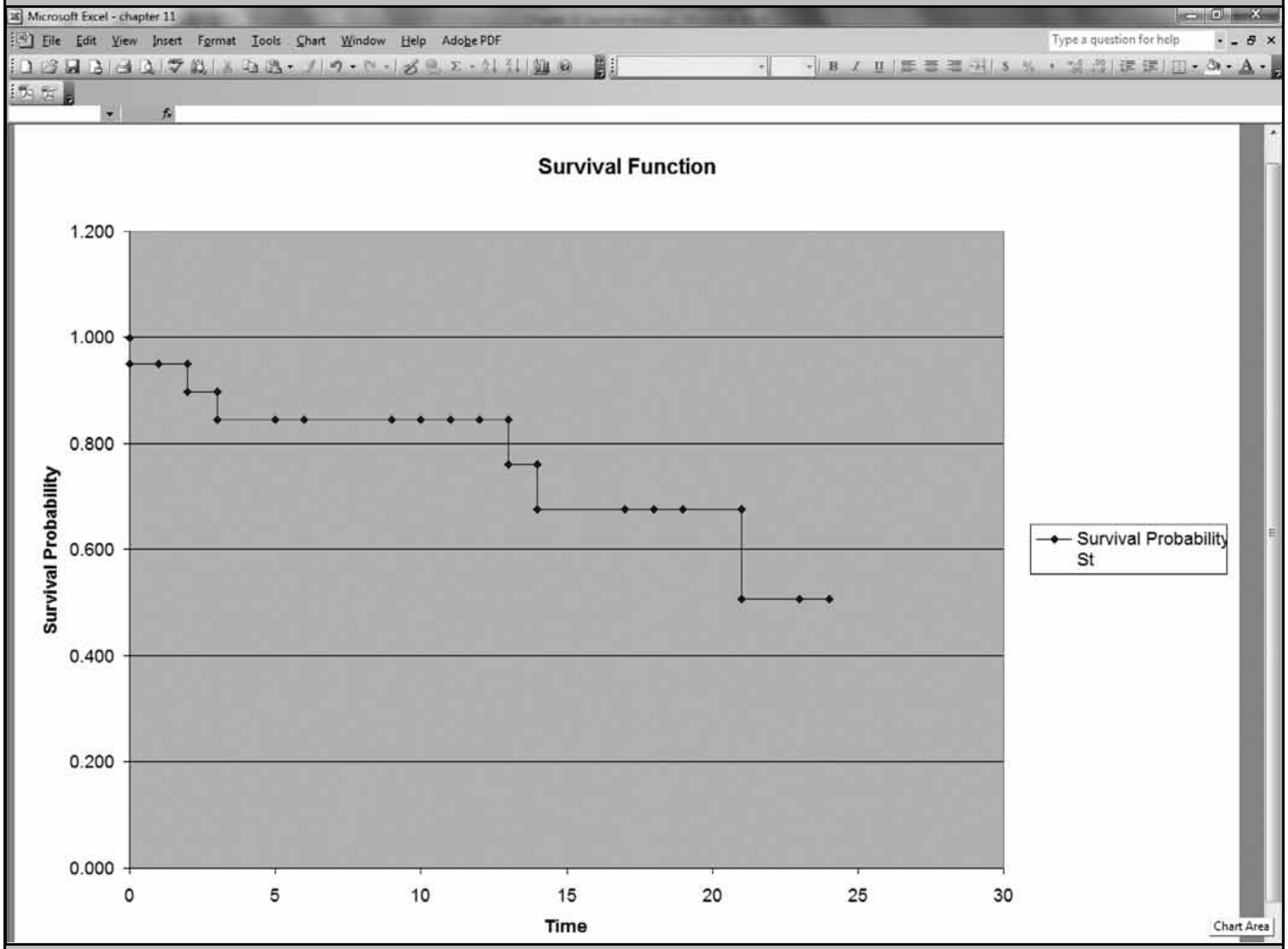
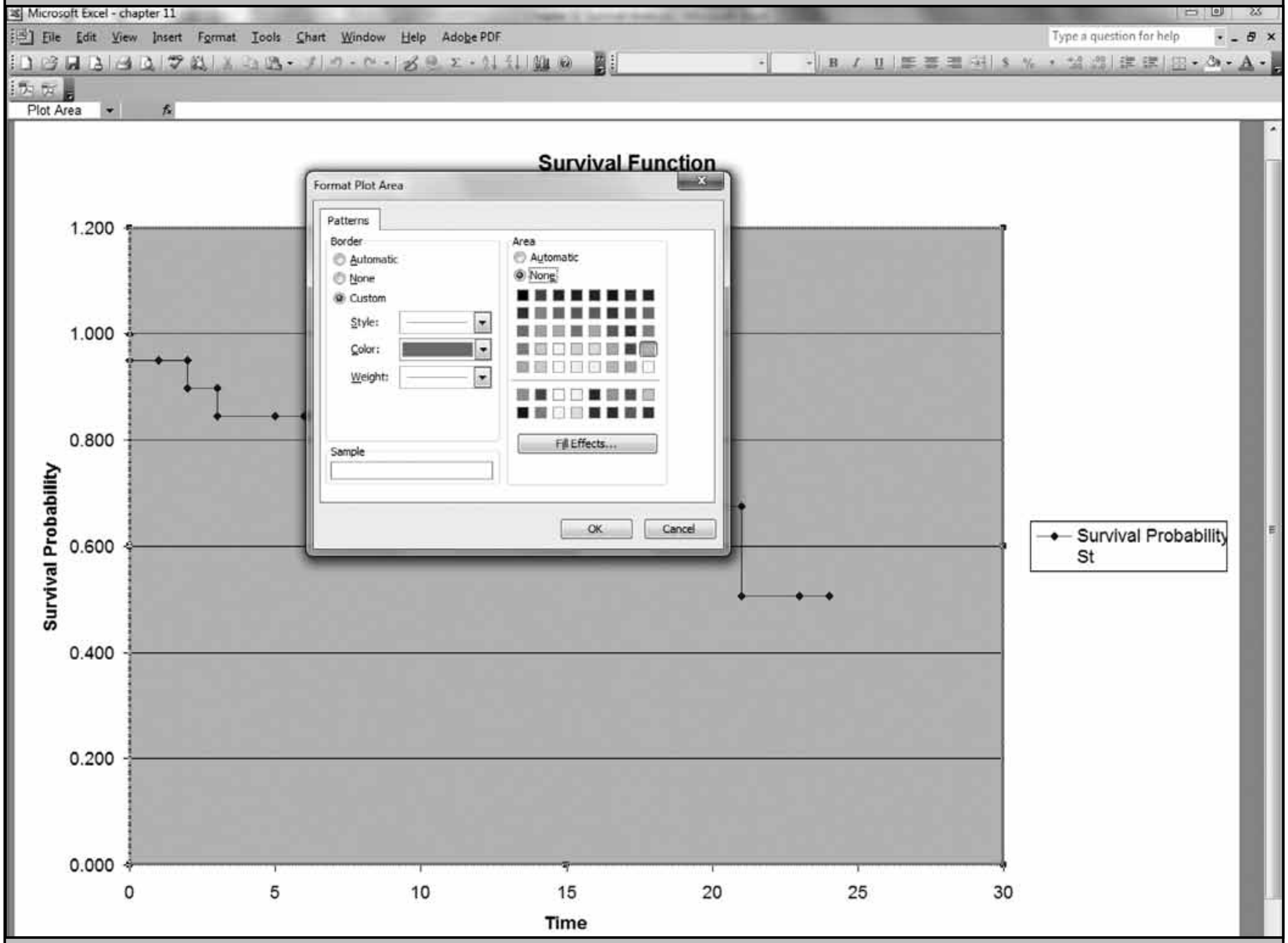


FIGURE 11-38 Formatting the Display



Border
○ Automatic
○ None
● Custom
Style:
Color:
Weight:
Sample

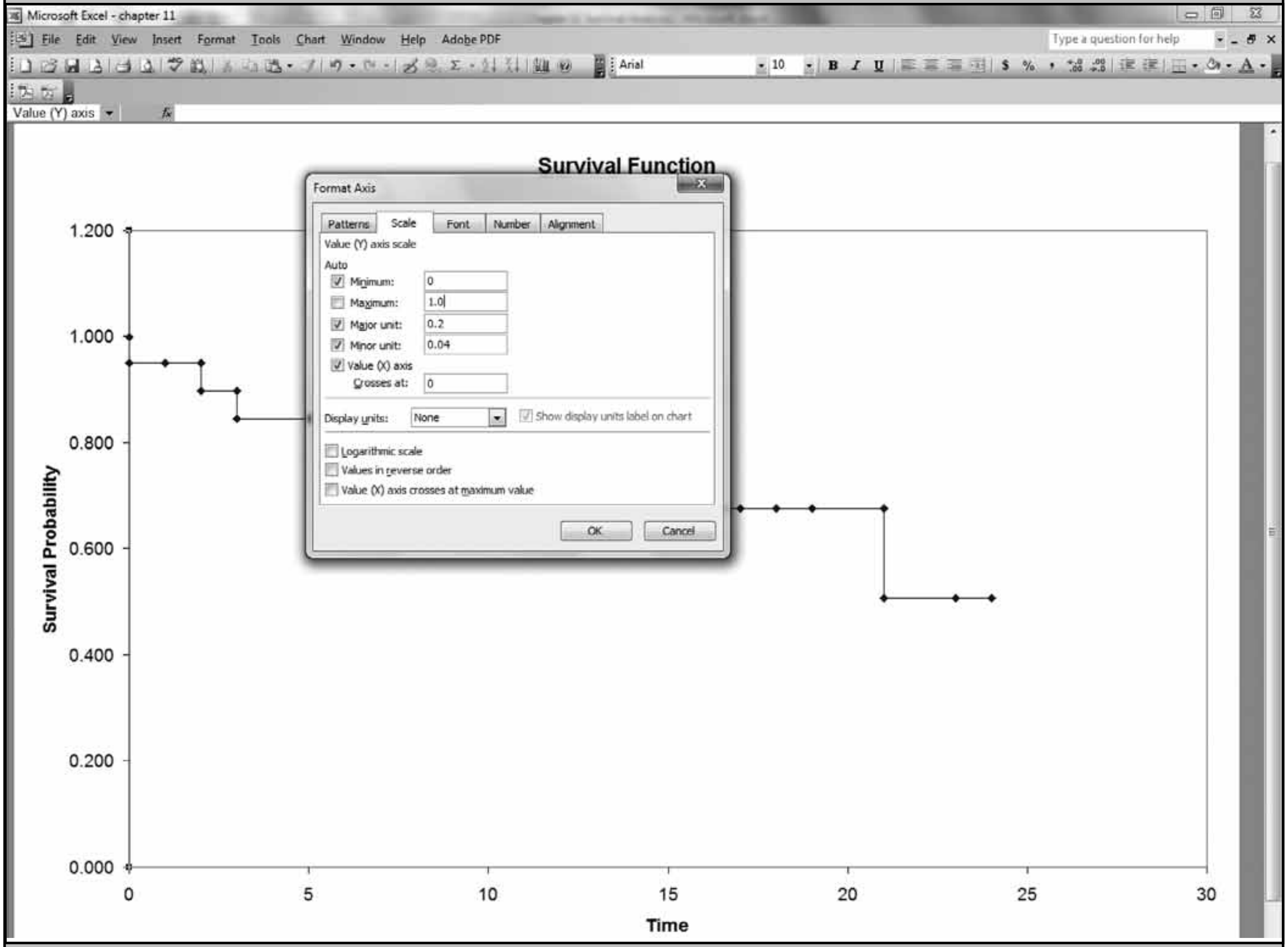
Area
○ Automatic
● None

11-40, the symbols represent each event time, either a death or a censored time.

Sometimes it is of interest to plot the estimates of the survival probabilities (as shown in Figure 11-40) along with 95% confidence limits. This is done using exactly the same approach used to plot the survival function. Specifically, we first generate the estimates of the survival probabilities and the 95% confidence limits as shown in Example 11.4 and in Figure 11-26. We then format the data to produce the step functions (in this case for the estimates as well as the upper and lower limits of the confidence interval). We illustrate the approach in the next example.

Example 11.6. Consider again the data in Example 11.4, where we generated estimates of survival probabilities and 95%

FIGURE 11-39 Rescaling the *y*-Axis



Value (Y) axis scale

Auto

☑ Minimum: 0

□ Maximum: 1.0

☑ Major unit: 0.2

☑ Minor unit: 0.04

☑ Value (X) axis Crosses at: 0

Display units: None ☑ Show display units label on chart

□ Logarithmic scale

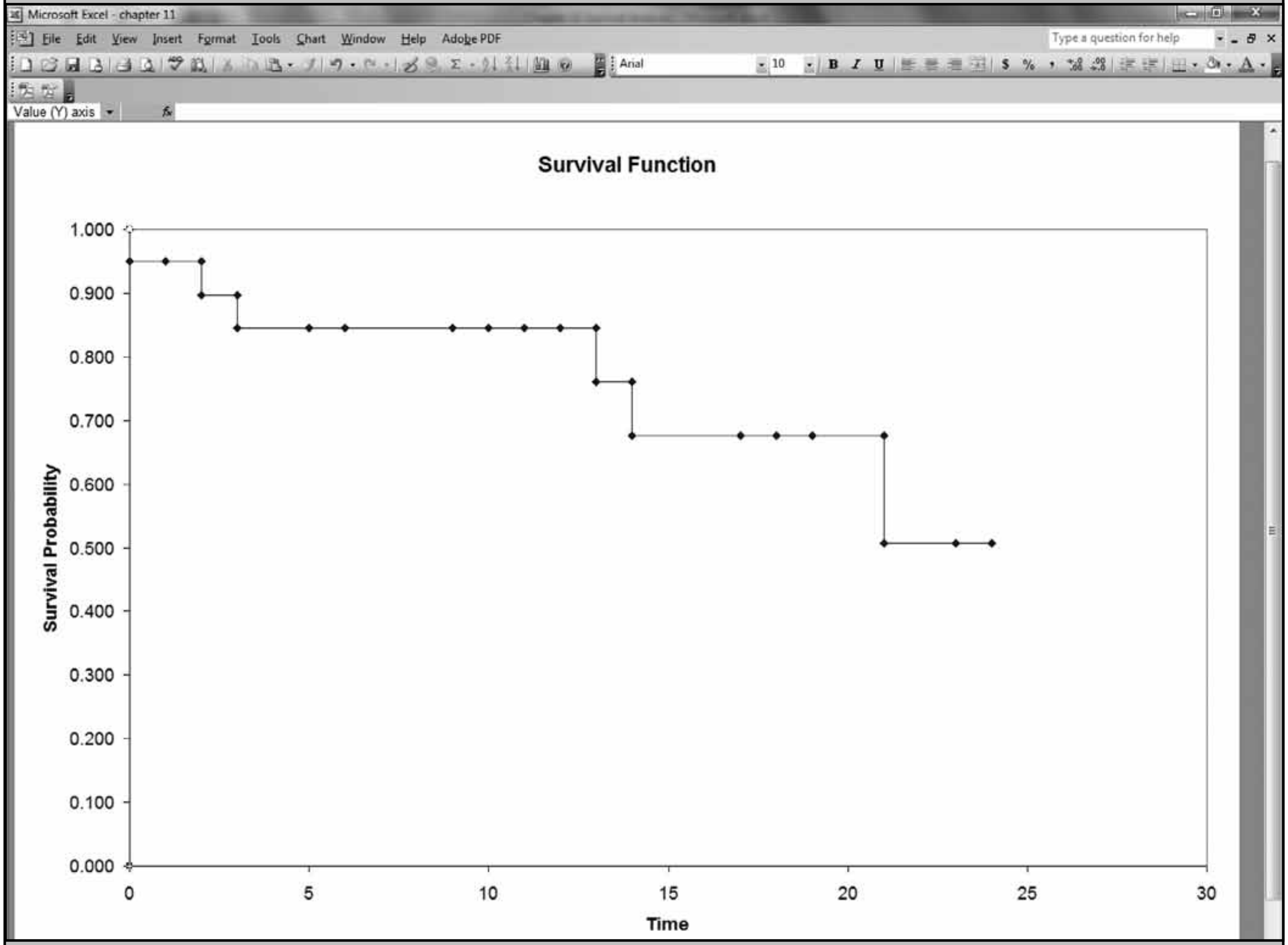
□ Values in reverse order

□ Value (X) axis crosses at maximum value

confidence limits. The results of the computations are shown in Figure 11-26. To facilitate interpretation, we first hide the columns in the worksheet that are not directly related to the graphical display. We then replace the formulas in the columns of the worksheet that are still showing by their calculated values. This is done by first highlighting the cells (columns D, J, O, and P in Figure 11-41), clicking on the copy icon on the top menu bar, and then selecting the **Values** option next to the paste icon on the top menu bar. Next we insert rows at each transition in survival probability (to produce a step function) as shown in Figure 11-41.

Because survival probabilities are between 0 and 1, inclusive, we recode any estimates to that range. Specifically, we recode the estimates of the upper limits from above 1.0 to

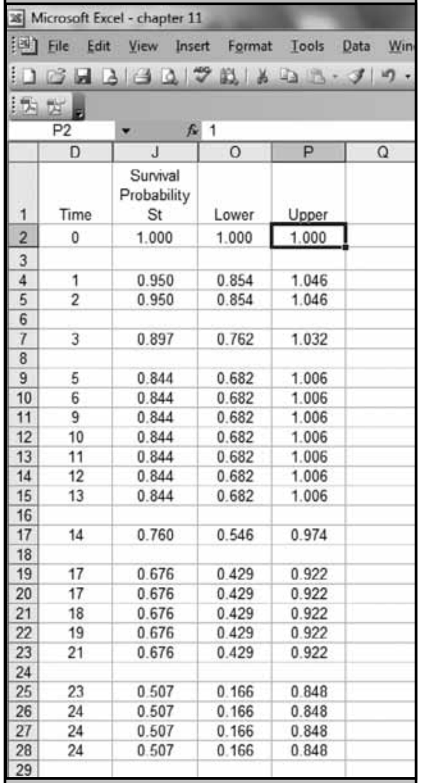
FIGURE 11-40 Plot of the Survival Function



1.0 as needed. This is shown in Figure 11-42. In addition, in each newly inserted (blank) row, we copy the time from the cell above and the survival probability and the lower and upper limits of the confidence interval from the cells below. For example, in cell D3 we enter "=D2" (the time from cell D2) and in cells J3, O3, and P3 we enter "=J4", "=O4", and "=P4", respectively. The same procedure is followed for rows 6, 8, 16, 18, and 24. For example, in cell D24 we enter "=D23", and in cells J24, O24, and P24 we enter "=J25", "=O25", and "=P25", respectively. The updated worksheet is shown in Figure 11-42.

We again use the Chart Wizard to create the display. The easiest way to produce the plot is to highlight the data shown in Figure 11-42 (specifically, the data for the *x*-axis in column D and the data for the *y*-axis in column J, column O, and column P). With the data highlighted, we click on the Chart

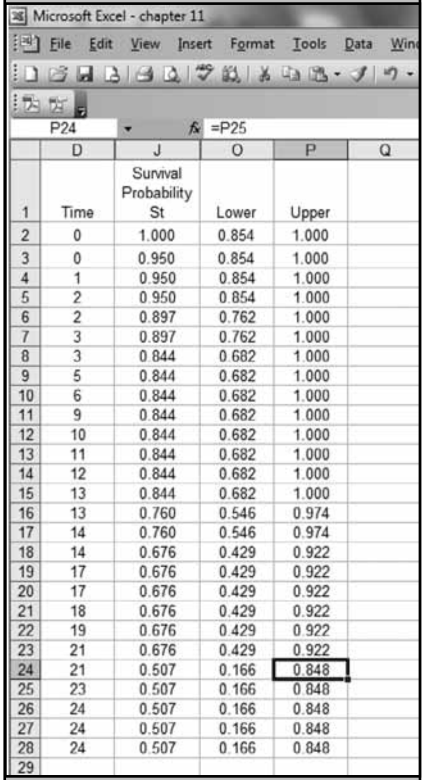
FIGURE 11-41 Inserting Rows at Each Transition Point


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Win

P2 fx 1

	D	J	O	P	Q
1	Time	Survival Probability St	Lower	Upper	
2	0	1.000	1.000	1.000	
3					
4	1	0.950	0.854	1.046	
5	2	0.950	0.854	1.046	
6					
7	3	0.897	0.762	1.032	
8					
9	5	0.844	0.682	1.006	
10	6	0.844	0.682	1.006	
11	9	0.844	0.682	1.006	
12	10	0.844	0.682	1.006	
13	11	0.844	0.682	1.006	
14	12	0.844	0.682	1.006	
15	13	0.844	0.682	1.006	
16					
17	14	0.760	0.546	0.974	
18					
19	17	0.676	0.429	0.922	
20	17	0.676	0.429	0.922	
21	18	0.676	0.429	0.922	
22	19	0.676	0.429	0.922	
23	21	0.676	0.429	0.922	
24					
25	23	0.507	0.166	0.848	
26	24	0.507	0.166	0.848	
27	24	0.507	0.166	0.848	
28	24	0.507	0.166	0.848	
29					

FIGURE 11-42 Data to Display Survival Probabilities and 95% Confidence Limits


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Win

P24 fx =P25

	D	J	O	P	Q
1	Time	Survival Probability St	Lower	Upper	
2	0	1.000	0.854	1.000	
3	0	0.950	0.854	1.000	
4	1	0.950	0.854	1.000	
5	2	0.950	0.854	1.000	
6	2	0.897	0.762	1.000	
7	3	0.897	0.762	1.000	
8	3	0.844	0.682	1.000	
9	5	0.844	0.682	1.000	
10	6	0.844	0.682	1.000	
11	9	0.844	0.682	1.000	
12	10	0.844	0.682	1.000	
13	11	0.844	0.682	1.000	
14	12	0.844	0.682	1.000	
15	13	0.844	0.682	1.000	
16	13	0.760	0.546	0.974	
17	14	0.760	0.546	0.974	
18	14	0.676	0.429	0.922	
19	17	0.676	0.429	0.922	
20	17	0.676	0.429	0.922	
21	18	0.676	0.429	0.922	
22	19	0.676	0.429	0.922	
23	21	0.676	0.429	0.922	
24	21	0.507	0.166	0.848	
25	23	0.507	0.166	0.848	
26	24	0.507	0.166	0.848	
27	24	0.507	0.166	0.848	
28	24	0.507	0.166	0.848	
29					

Wizard icon as shown in Figure 11-43, and we select the XY (Scatter) plot under Chart Types and the chart subtype showing in the bottom left corner.

When we click Next, Excel produces a preliminary version of the display, and Excel then asks for a chart title as well as labels for the *x*- and *y*-axes (see Figure 11-44).

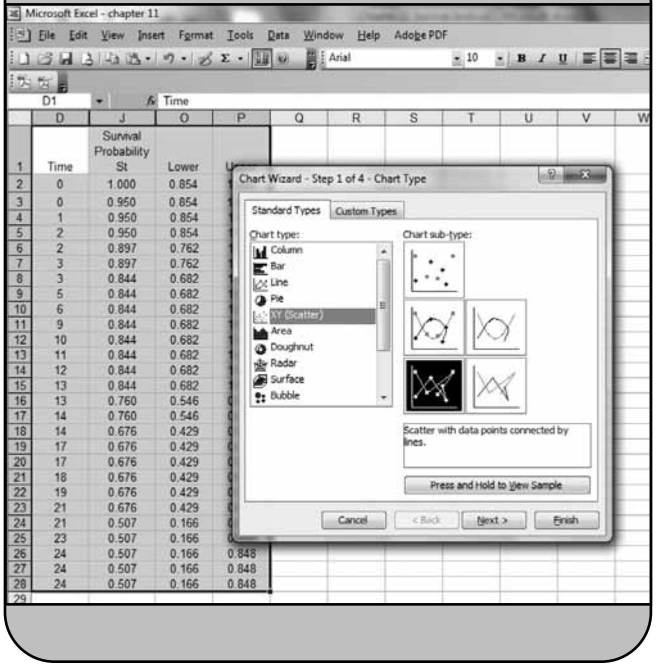
Once we enter the title and labels, Excel prompts for a location for the display (see Figure 11-31). If we select a new worksheet as the location, Excel produces the display shown in Figure 11-45.

Again, there are a number of formatting changes that we make prior to presenting the display. First, we change the background color to white or to no color, remove the legend and horizontal lines, and rescale the *y*-axis to a maximum of 1.0. (All of these changes were illustrated in Example 11.5.) The formatted display is shown in Figure 11-46.

An additional formatting step is needed to change the colors of the lines (the estimated survival probability and the lower and upper limits of the 95% confidence interval) to black. This is done by clicking on any part of the line (one at a time). For example, suppose we click on the estimated survival probability line (the middle line shown in Figure 11-46). This brings up the dialog box shown in Figure 11-47, where we specify a new line color of black under the **Line** option in the Patterns tab. We also change the color of the marker (both the foreground and the background) to black using the same approach. This is done under the **Markers** option in the **Patterns** tab.

When we click **OK**, Excel updates the color scheme. We do the same for the upper and lower limits of the 95% confidence interval. We also perform one additional step in formatting the confidence limits. We change the style of the lines to dashed

FIGURE 11-43 Selecting the Chart Type



Microsoft Excel - chapter 11

D1 | Time

	D	J	O	P
1	Time	Survival Probability St	Lower	U
2	0	1.000	0.854	
3	0	0.950	0.854	
4	1	0.950	0.854	
5	2	0.950	0.854	
6	2	0.897	0.762	
7	3	0.897	0.762	
8	3	0.844	0.682	
9	5	0.844	0.682	
10	6	0.844	0.682	
11	9	0.844	0.682	
12	10	0.844	0.682	
13	11	0.844	0.682	
14	12	0.844	0.682	
15	13	0.844	0.682	
16	13	0.760	0.546	
17	14	0.760	0.546	
18	14	0.676	0.429	
19	17	0.676	0.429	
20	17	0.676	0.429	
21	18	0.676	0.429	
22	19	0.676	0.429	
23	21	0.676	0.429	
24	21	0.507	0.166	
25	23	0.507	0.166	
26	24	0.507	0.166	0.848
27	24	0.507	0.166	0.848
28	24	0.507	0.166	0.848

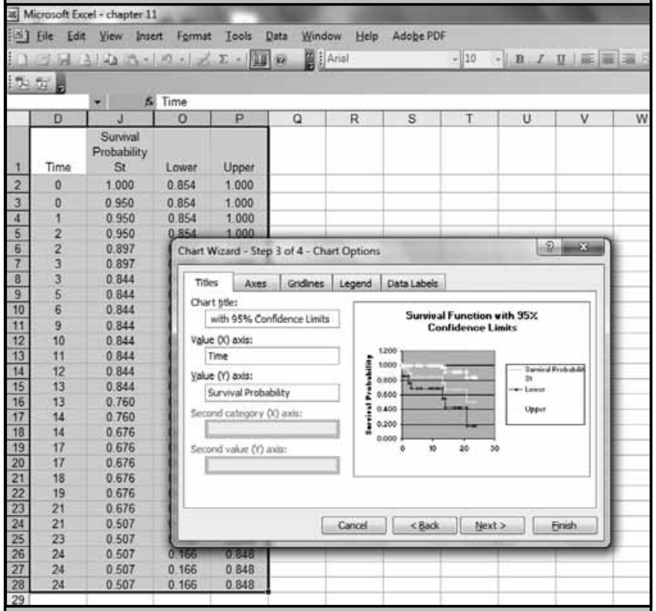
Chart Wizard - Step 1 of 4 - Chart Type

Standard Types | Custom Types

Chart type: XY (Scatter)

Scatter with data points connected by lines.

FIGURE 11-44 Specifying a Title and Labels for the Axes



Microsoft Excel - chapter 11

Time

	D	J	O	P
1	Time	Survival Probability St	Lower	Upper
2	0	1.000	0.854	1.000
3	0	0.950	0.854	1.000
4	1	0.950	0.854	1.000
5	2	0.950	0.854	1.000
6	2	0.897		
7	3	0.897		
8	3	0.844		
9	5	0.844		
10	6	0.844		
11	9	0.844		
12	10	0.844		
13	11	0.844		
14	12	0.844		
15	13	0.844		
16	13	0.760		
17	14	0.760		
18	14	0.676		
19	17	0.676		
20	17	0.676		
21	18	0.676		
22	19	0.676		
23	21	0.676		
24	21	0.507		
25	23	0.507		
26	24	0.507	0.166	0.848
27	24	0.507	0.166	0.848
28	24	0.507	0.166	0.848

Chart Wizard - Step 3 of 4 - Chart Options

Titles | Axes | Gridlines | Legend | Data Labels

Chart title: with 95% Confidence Limits

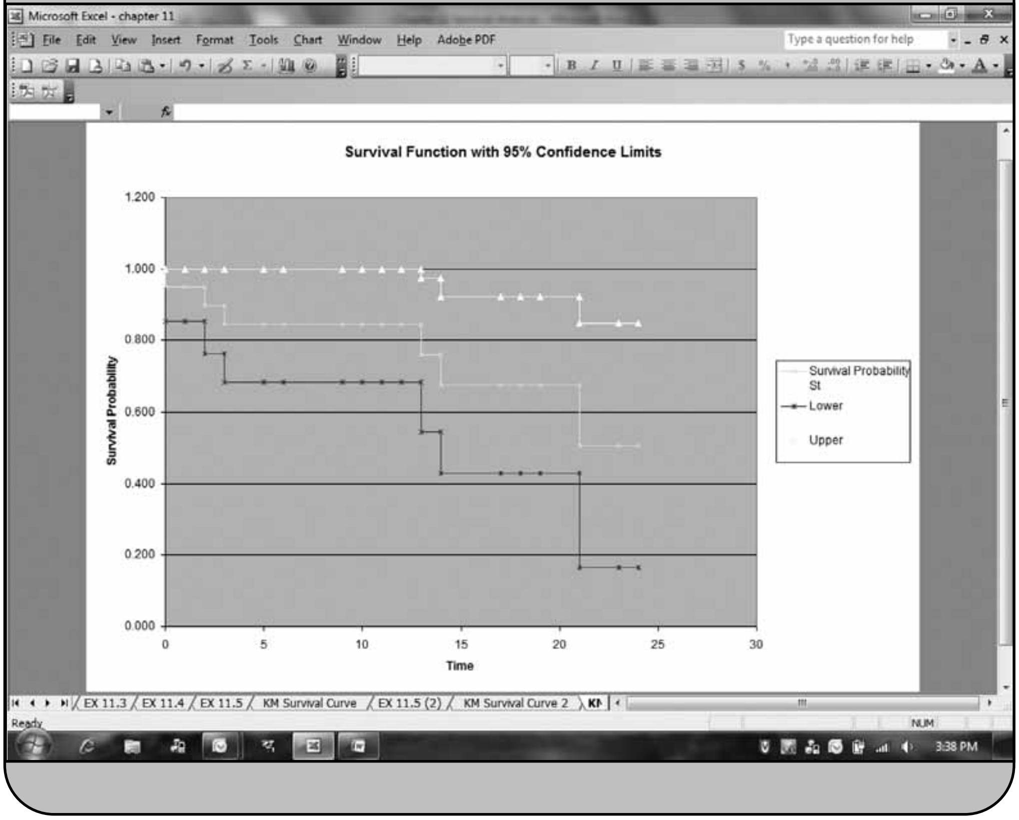
Value (X) axis: Time

Value (Y) axis: Survival Probability

Second category (X) axis:

Second value (Y) axis:

FIGURE 11-45 The Display



Microsoft Excel - chapter 11

FIGURE 11-46 The Formatted Display

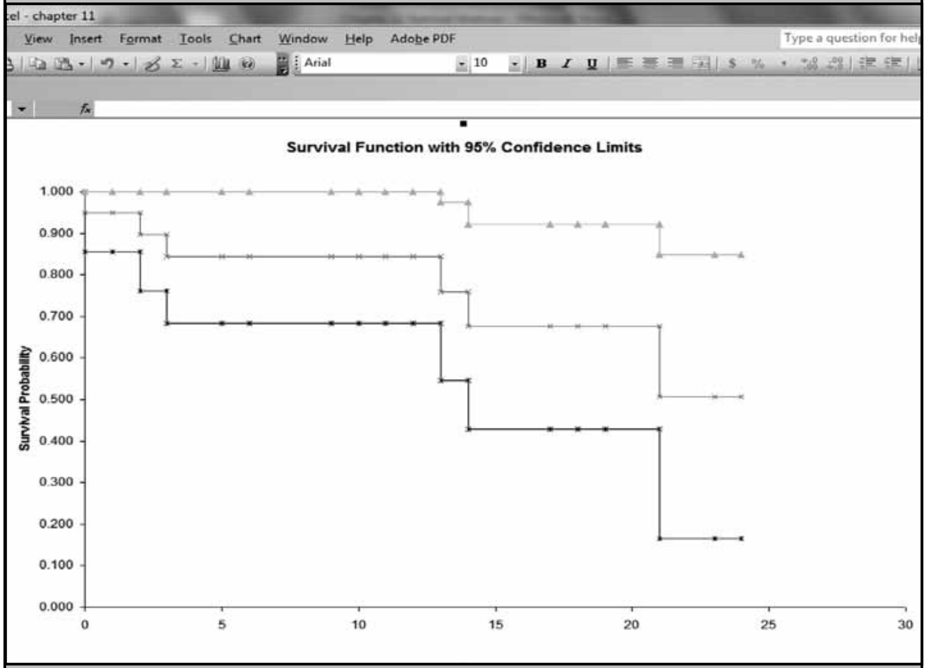
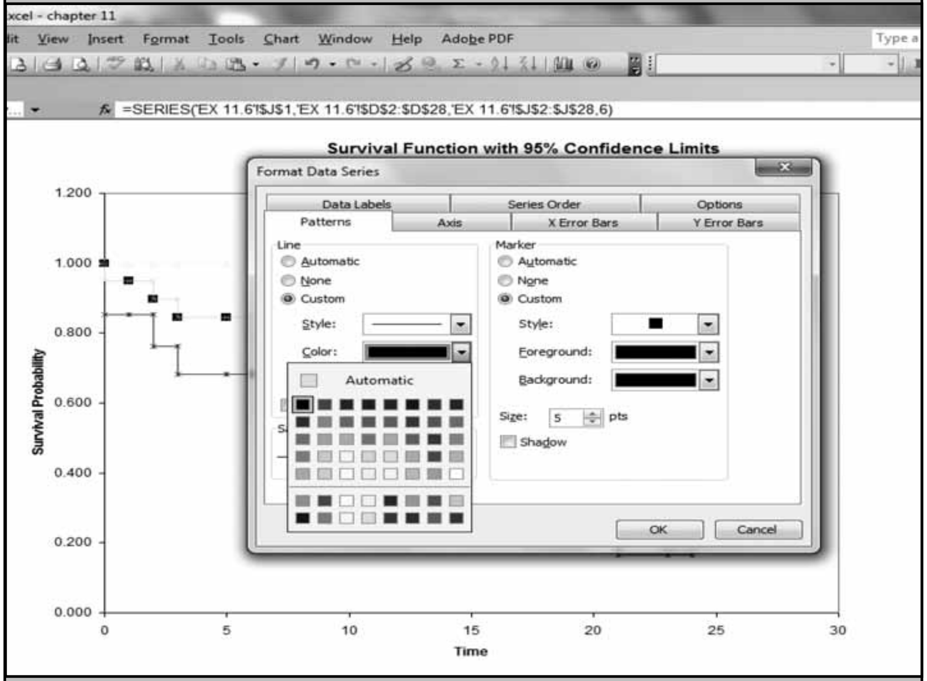


FIGURE 11-47 Changing the Line Color



=SERIES('EX 11.6'!$J$1,'EX 11.6'!$D$2:$D$28,'EX 11.6'!$J$2:$J$28,6)

Line
○ Automatic
○ None
● Custom
Style:
Color:

Marker
○ Automatic
○ None
● Custom
Style:
Foreground:
Background:
Size: 5 pts
□ Shadow

(as opposed to solid) to distinguish between the estimates of the survival probabilities (shown in solid black) from the confidence limits (shown in dashed black). The style of the lines can be changed by clicking on the arrow to the right of the Style box under the Patterns tab shown in Figure 11-47. The final display is shown in Figure 11-48.

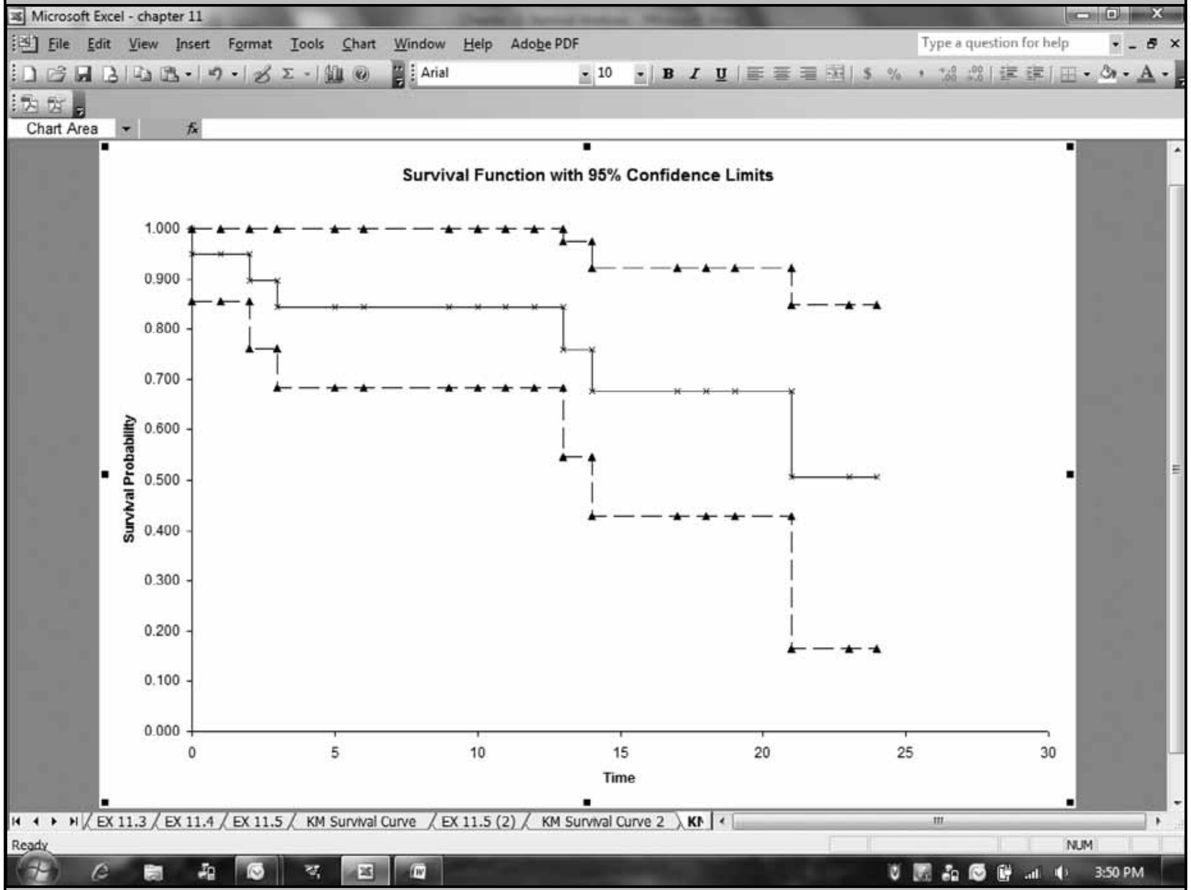
Notice that we also select triangles as markers for both the upper and lower limits of the 95% confidence interval.

11.3 COMPARING SURVIVAL CURVES

In many survival analysis applications, we are interested in assessing whether there are differences in survival among different groups of participants. For example, in a clinical trial with a survival outcome, we are often interested in comparing survival between participants receiving a new drug as compared to a placebo (or other appropriate comparator). In an observational study, we might be interested in comparing survival between men and women, or between participants with and without a particular risk factor (e.g., hypertension or diabetes).

There are several tests available to compare survival among independent groups. In Chapter 11 of the textbook we presented the log-rank test, which is a popular test to compare survival between two or more independent groups. The test compares the entire survival experience between groups and can be thought of as a test of whether the survival curves are identical (overlapping) or not. Survival curves are estimated for each group, considered separately, using the Kaplan–Meier approach and are compared statistically using the log-rank test. The log-rank test is computed using the five-step approach for hypothesis testing. The test statistic for the log-rank test is as follows: $\chi^2 = \sum \frac{(\Sigma O_{jt} - \Sigma E_{jt})^2}{\Sigma E_{jt}}$, where $\Sigma\ O_{jt}$ represents the

FIGURE 11-48 Estimates of Survival Probabilities with 95% Confidence Limits



sum of the observed number of events in the *j*th group (e.g., $j = 1, 2$) and $\Sigma\, E_{jt}$ represents the sum of the expected number of events in the *j*th group over time and is approximately distributed as a χ^2 test statistic. Using Excel we can compute the value of the test statistic and compute a *p*-value using the CHIDIST function. The log-rank test statistic has degrees of freedom equal to $k - 1$, where *k* represents the number of comparison groups.

Example 11.7. In Example 11.3 in the textbook we analyzed a small clinical trial comparing two combination treatments in patients with advanced gastric cancer. Twenty participants with stage IV gastric cancer who consented to participate in the trial were randomly assigned to receive chemotherapy before surgery or chemotherapy after surgery. The primary outcome was death and participants were followed for up to 48 months (4 years) following enrollment into the trial. The experiences of participants in each arm of the trial are shown in Table 11-3. The question of interest is whether there is a difference in survival between the two treatments. The test is run at a 5% level of significance and the hypotheses are as follows:

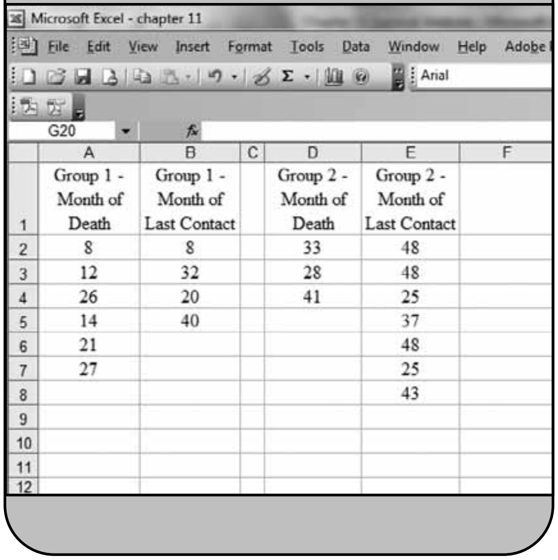
H_0: The two survival curves are identical (or $S_{1t} = S_{2t}$).

H_1: The two survival curves are not identical (or $S_{1t} \neq S_{2t}$ at any time *t*).

TABLE 11-3 Month of Death or Month of Last Contact in Each Treatment Group

Chemotherapy Before Surgery		Chemotherapy After Surgery	
Month of Death	Month of Last Contact	Month of Death	Month of Last Contact
8	8	33	48
12	32	28	48
26	20	41	25
14	40		37
21			48
27			25
			43

FIGURE 11-49 Data for the Log-Rank Test

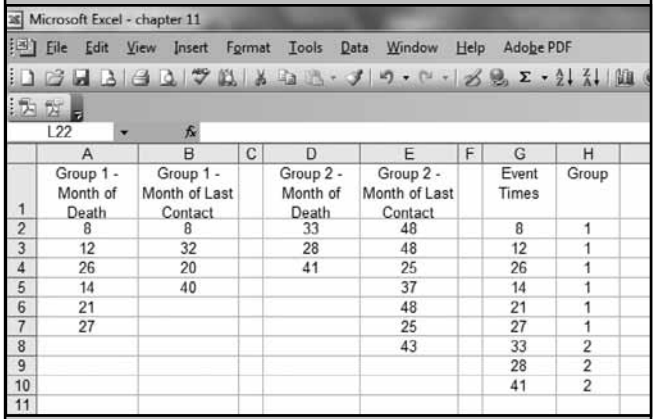

Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe
Arial
G20
fx

	A	B	C	D	E	F
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact	
2	8	8		33	48	
3	12	32		28	48	
4	26	20		41	25	
5	14	40			37	
6	21				48	
7	27				25	
8					43	
9						
10						
11						
12						

The data are entered into Excel as shown in Figure 11-49.

To prepare the data for the computations, we now create two new variables, event times and group. The event times are the times of the observed events (deaths in this example). Group indicates whether the events occurred in group 1 or in group 2. The two new variables are shown in Figure 11-50.

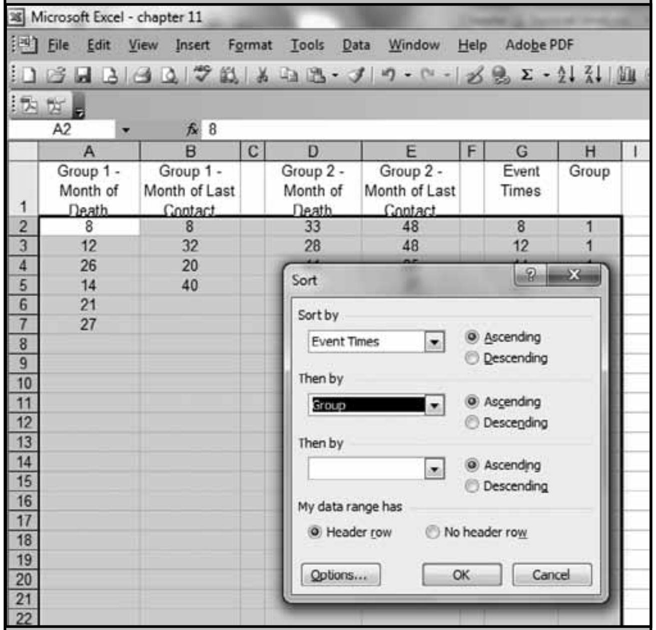
FIGURE 11-50 Creating New Variables


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
L22
fx

	A	B	C	D	E	F	G	H
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group
2	8	8		33	48		8	1
3	12	32		28	48		12	1
4	26	20		41	25		26	1
5	14	40			37		14	1
6	21				48		21	1
7	27				25		27	1
8					43		33	2
9							28	2
10							41	2
11								

Notice that the event times are not sorted; they are simply copied from column A and column D. The next step is to sort the data. Specifically, we sort the data by the two new variables, first by event times in ascending order and then by group in ascending order. This is done by highlighting the data (column A through column H) and selecting sort under the Data option on the top menu bar. Once we select the Sort option, Excel presents the dialog box shown in Figure 11-51. We spec-

FIGURE 11-51 Sorting the Data



Microsoft Excel - chapter 11

A2 =8

	A	B	C	D	E	F	G	H	I
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group	
2	8	8		33	48		8	1	
3	12	32		28	48		12	1	
4	26	20							
5	14	40							
6	21								
7	27								

Sort

Sort by: Event Times — ● Ascending ○ Descending

Then by: Group — ● Ascending ○ Descending

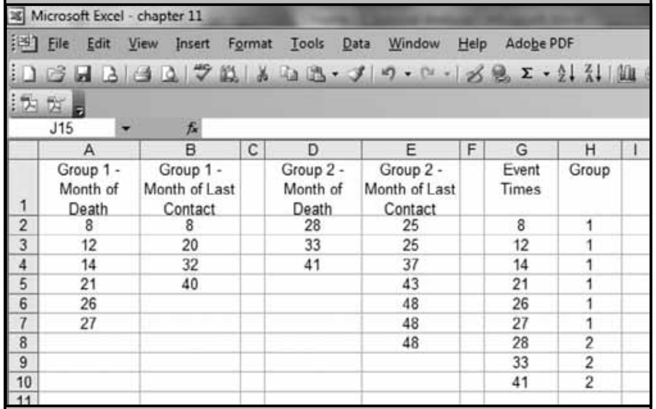
Then by: — ● Ascending ○ Descending

My data range has: ● Header row ○ No header row

ify that the primary sort variable is event times and that we want the data sorted by event times in ascending order. We specify that we then want the data sorted by group in ascending order. This is shown in Figure 11-51. Notice that we also indicate that there is a header row in the dataset using the option at the bottom of dialog box. When we indicate that there is a Header row, Excel unselects the first row.

The sorted data are shown in Figure 11-52.

FIGURE 11-52 Sorted Data



Microsoft Excel - chapter 11

J15

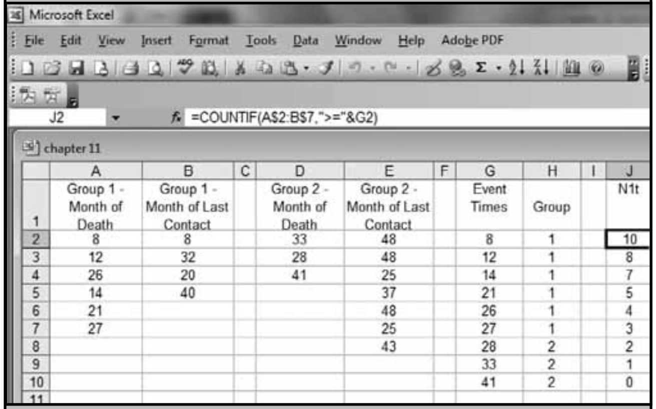
	A	B	C	D	E	F	G	H	I
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group	
2	8	8		28	25		8	1	
3	12	20		33	25		12	1	
4	14	32		41	37		14	1	
5	21	40			43		21	1	
6	26				48		26	1	
7	27				48		27	1	
8					48		28	2	
9							33	2	
10							41	2	

Next, we want to compute the numbers of participants at risk in each group at each observed event time. To do this, we use the COUNTIF function. To compute the numbers of participants in group 1 at risk at each event time, N_{1t}, we count the numbers of participants in group 1 with observed times (either month of death or month of last contact) that are greater than or equal to each observed event time in column G. Specifically, in cell J2 we enter "=COUNTIF(A$2:B$7,">="&G2)". Notice that we specify the range of the data for group 1 as cell A$2 through cell B$7 using absolute cell references so that we continue to refer to the same data range when we copy the formula from cell J2 to cell J3 through cell J10. The numbers of participants in group 1 at risk at each event time are shown in Figure 11-53.

We use the same approach to compute the numbers of participants in group 2 at risk at each event time, N_{2t}. We count the numbers of participants in group 2 with observed times (either month of death or month of last contact) that are greater than or equal to each observed event time in column G. Specifically, in cell K2 we enter "=COUNTIF(D$2:E$8,">=" &G2)". Again, we specify the range of the data for group 2 as cell D$2 through cell E$8 using absolute cell references so that we continue to refer to the same data range when we copy the formula from cell K2 to cell K3 through cell K10. The numbers of participants in group 2 at risk at each event time are shown in Figure 11-54.

In the next step, we compute the numbers of observed events in each group, O_{1t} and O_{2t}. These are computed in columns L and M using the IF function. Specifically, in column L we compute the numbers of observed events in group 1 by specifying "=IF(H2=1,1,0)" in cell L2 and copying this

FIGURE 11-53 Computing the Number of Participants in Group 1 at Risk at Each Event Time



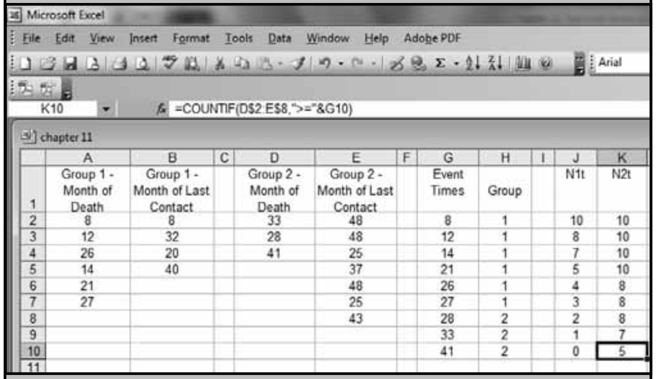
Microsoft Excel

J2 =COUNTIF(A$2:B$7,">="&G2)

chapter 11

	A	B	C	D	E	F	G	H	I	J
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group		N1t
2	8	8		33	48		8	1		10
3	12	32		28	48		12	1		8
4	26	20		41	25		14	1		7
5	14	40			37		21	1		5
6	21				48		26	1		4
7	27				25		27	1		3
8					43		28	2		2
9							33	2		1
10							41	2		0

FIGURE 11-54 Computing the Number of Participants in Group 2 at Risk at Each Event Time


Microsoft Excel
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
chapter 11

K10 =COUNTIF(D$2:E$8,">="&G10)

	A	B	C	D	E	F	G	H	I	J	K
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group		N1t	N2t
2	8	8		33	48		8	1		10	10
3	12	32		28	48		12	1		8	10
4	26	20		41	25		14	1		7	10
5	14	40			37		21	1		5	10
6	21				48		26	1		4	8
7	27				25		27	1		3	8
8					43		28	2		2	8
9							33	2		1	7
10							41	2		0	5
11											

formula into cell L3 through cell L10. Similarly, in column M we compute the numbers of observed events in group 2 by specifying "=IF(H2=2,1,0)" in cell M2 and copying this formula into cell M3 through cell M10. The numbers of observed events in each group are shown in Figure 11-55.

Next, we compute the total number of participants at risk at each event time, N_t, by summing the numbers of participants at risk across groups (i.e., summing N_{1t} and N_{2t}) at each event time. This is shown in Figure 11-56.

Next, we compute the total number of observed events at each event time, O_t, by summing the numbers of observed events across groups (i.e., summing O_{1t} and O_{2t}) at each event time. This is shown in Figure 11-57.

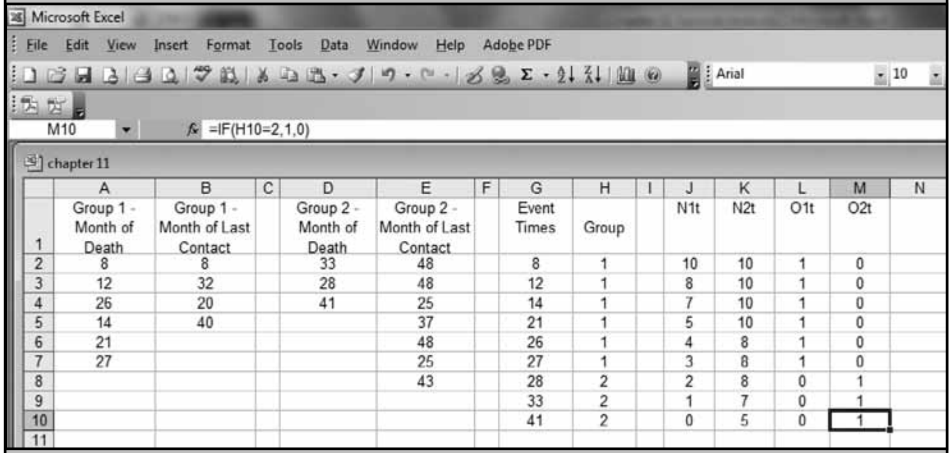
Next we compute the expected number of events in each group, E_{1t} and E_{2t}, at each event time using: $E_{1t} = N_{1t} \times (O_t/N_t)$ for group 1 and $E_{2t} = N_{2t} \times (O_t/N_t)$ for group 2. Specifically, in column P we compute the expected number of events in group 1 by entering: "=J2*(O2/N2)" in cell P2 and copying this formula into cell P3 through cell P10. This is shown in Figure 11-58. We do the same for group 2 by entering "=K2*(O2/N2)" in cell Q2 and copying this formula into cell Q3 through cell Q10.

The final step is to compute the test statistic. However, to do so we need the total number of observed events in each group (i.e., the sums of O_{1t} and O_{2t}, respectively) and the total number of expected events in each group (i.e., the sums of E_{1t} and E_{2t}, respectively). These are computed using the SUM function as shown in Figure 11-59.

The next step is to compute the log-rank test statistic: $\chi^2 = \sum \frac{(\Sigma O_{jt} - \Sigma E_{jt})^2}{\Sigma E_{jt}}$, where ΣO_{jt} represents the sum of the observed number of events in the jth group (e.g., $j = 1, 2$) and ΣE_{jt} represents the sum of the expected number of events in the jth group over time. We compute the test statistic in two steps. First, we take the ratio of the square the difference between the sum of the observed and the sum of the expected numbers of events to the sum of the expected number of events in each group (see Figure 11-60), and then we sum to produce the test statistic (see Figure 11-61).

Thus, $\chi^2 = 6.148$. We now compute a p-value for the test using the χ^2 distribution and the CHIDIST function in Excel.

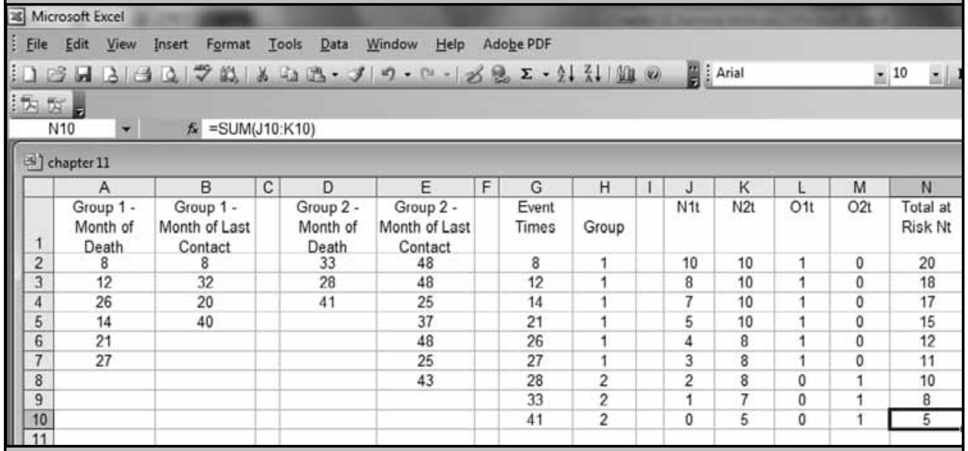
FIGURE 11-55 Computing the Number of Observed Events in Each Group at Each Event Time


Microsoft Excel
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
chapter 11

M10 =IF(H10=2,1,0)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group		N1t	N2t	O1t	O2t	
2	8	8		33	48		8	1		10	10	1	0	
3	12	32		28	48		12	1		8	10	1	0	
4	26	20		41	25		14	1		7	10	1	0	
5	14	40			37		21	1		5	10	1	0	
6	21				48		26	1		4	8	1	0	
7	27				25		27	1		3	8	1	0	
8					43		28	2		2	8	0	1	
9							33	2		1	7	0	1	
10							41	2		0	5	0	1	
11														

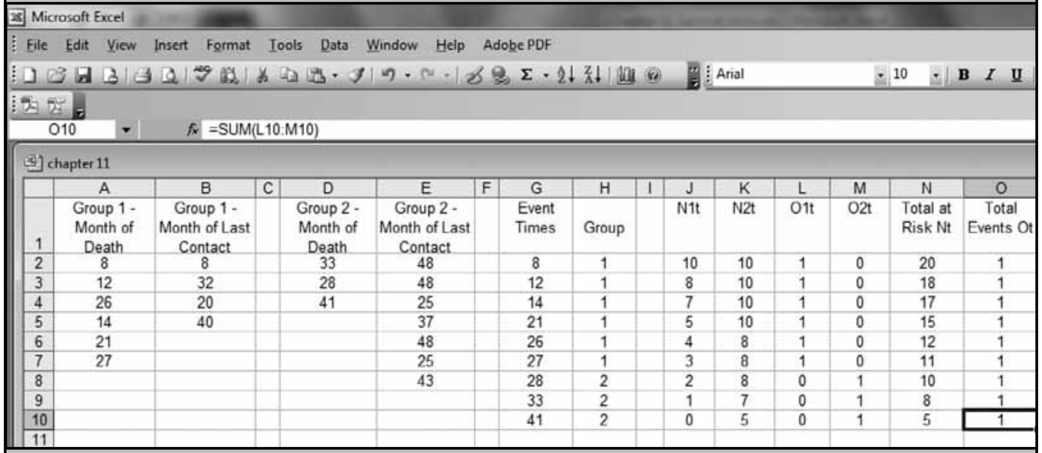
FIGURE 11-56 Computing the Total Number of Participants at Risk at Each Event Time


Microsoft Excel
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
chapter 11

N10 =SUM(J10:K10)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group		N1t	N2t	O1t	O2t	Total at Risk Nt
2	8	8		33	48		8	1		10	10	1	0	20
3	12	32		28	48		12	1		8	10	1	0	18
4	26	20		41	25		14	1		7	10	1	0	17
5	14	40			37		21	1		5	10	1	0	15
6	21				48		26	1		4	8	1	0	12
7	27				25		27	1		3	8	1	0	11
8					43		28	2		2	8	0	1	10
9							33	2		1	7	0	1	8
10							41	2		0	5	0	1	5
11														

FIGURE 11-57 Computing the Total Number of Observed Events at Each Event Time


Microsoft Excel
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B I U
chapter 11

O10 =SUM(L10:M10)

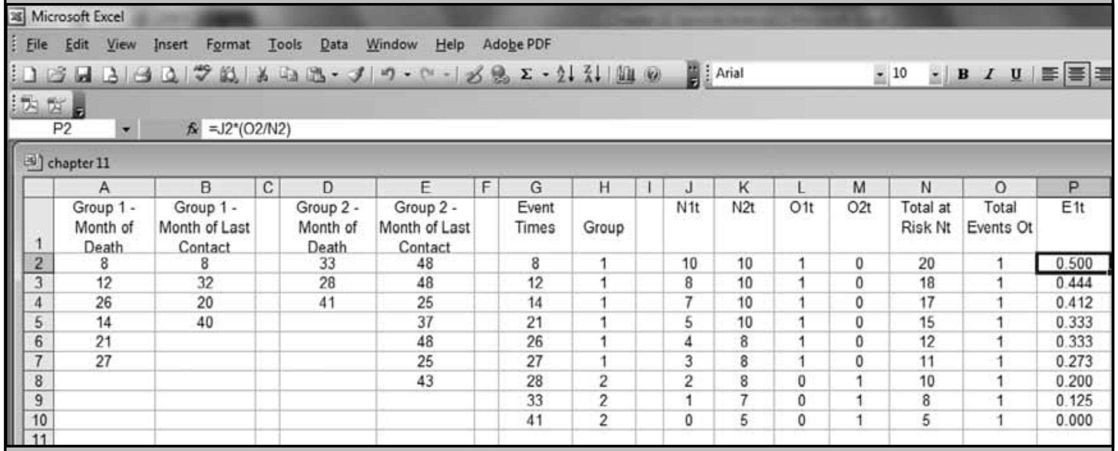
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group		N1t	N2t	O1t	O2t	Total at Risk Nt	Total Events Ot
2	8	8		33	48		8	1		10	10	1	0	20	1
3	12	32		28	48		12	1		8	10	1	0	18	1
4	26	20		41	25		14	1		7	10	1	0	17	1
5	14	40			37		21	1		5	10	1	0	15	1
6	21				48		26	1		4	8	1	0	12	1
7	27				25		27	1		3	8	1	0	11	1
8					43		28	2		2	8	0	1	10	1
9							33	2		1	7	0	1	8	1
10							41	2		0	5	0	1	5	1
11															

The CHIDIST function requires specification of the test statistic (χ^2) and the degrees of freedom. The log-rank statistic has degrees of freedom equal to $k - 1$, where k represents the number of comparison groups. In this example, $k = 2$, so the test statistic has 1 degree of freedom. The p-value for the test is computed as shown in Figure 11-62.

Because the p-value = 0.013 is less than the level of significance $\alpha = 0.05$, we reject H_0. We have significant evidence at $\alpha = 0.05$ to show that the two survival curves are different.

Again, while it is tedious to construct the template for the computation of the log-rank test the first time, once the structure is entered into Excel, implementation of the test for new

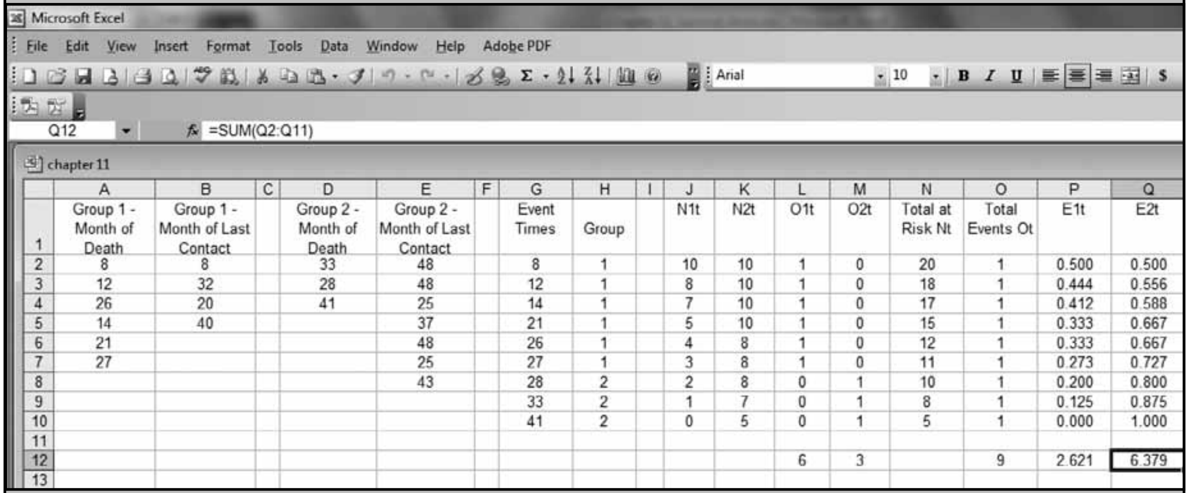
FIGURE 11-58 Computing the Number of Expected Events in Group 1 at Each Event Time


Microsoft Excel
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial 10
chapter 11

P2 =J2*(O2/N2)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group		N1t	N2t	O1t	O2t	Total at Risk Nt	Total Events Ot	E1t
2	8	8		33	48		8	1		10	10	1	0	20	1	0.500
3	12	32		28	48		12	1		8	10	1	0	18	1	0.444
4	26	20		41	25		14	1		7	10	1	0	17	1	0.412
5	14	40			37		21	1		5	10	1	0	15	1	0.333
6	21				48		26	1		4	8	1	0	12	1	0.333
7	27				25		27	1		3	8	1	0	11	1	0.273
8					43		28	2		2	8	0	1	10	1	0.200
9							33	2		1	7	0	1	8	1	0.125
10							41	2		0	5	0	1	5	1	0.000
11																

FIGURE 11-59 Computing the Total Numbers of Observed and Expected Events in Each Group


Microsoft Excel
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial 10
chapter 11

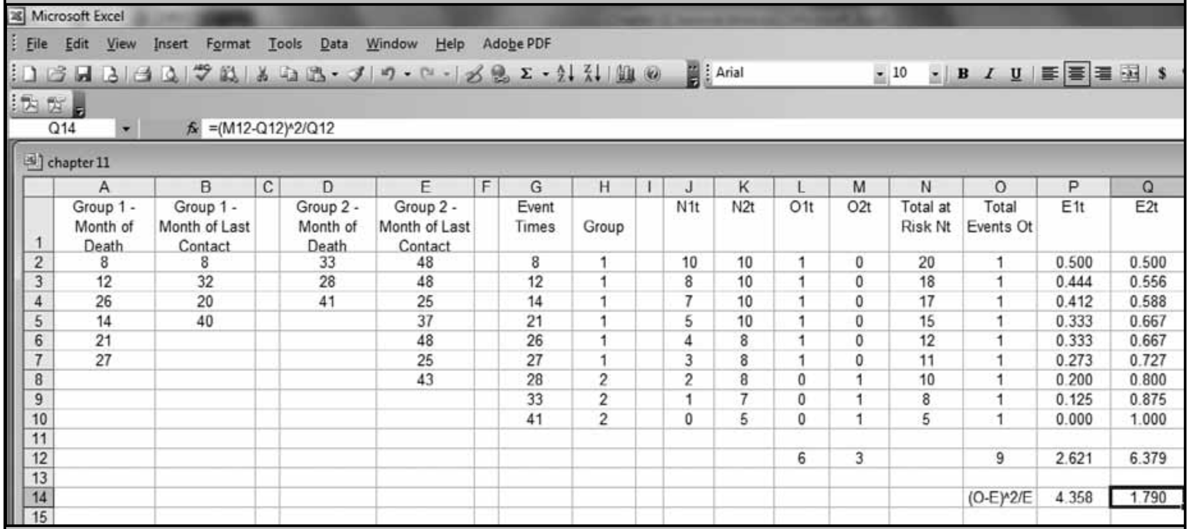
Q12 =SUM(Q2:Q11)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group		N1t	N2t	O1t	O2t	Total at Risk Nt	Total Events Ot	E1t	E2t
2	8	8		33	48		8	1		10	10	1	0	20	1	0.500	0.500
3	12	32		28	48		12	1		8	10	1	0	18	1	0.444	0.556
4	26	20		41	25		14	1		7	10	1	0	17	1	0.412	0.588
5	14	40			37		21	1		5	10	1	0	15	1	0.333	0.667
6	21				48		26	1		4	8	1	0	12	1	0.333	0.667
7	27				25		27	1		3	8	1	0	11	1	0.273	0.727
8					43		28	2		2	8	0	1	10	1	0.200	0.800
9							33	2		1	7	0	1	8	1	0.125	0.875
10							41	2		0	5	0	1	5	1	0.000	1.000
11																	
12												6	3		9	2.621	6.379
13																	

datasets is relatively easy. Only the data range in the formulas must be modified.

Example 11.8. In Example 11.4 in the textbook, we evaluated the efficacy of a brief intervention to prevent alcohol consumption in pregnancy. Pregnant women with a history of heavy alcohol consumption were recruited into the study and randomized to receive either the brief intervention focused on abstinence from alcohol or standard prenatal care. The outcome of interest was relapse to drinking. Women were recruited into the study at approximately 18 weeks gestation and followed through the course of pregnancy to delivery (approximately 39 weeks gestation). The data are shown in

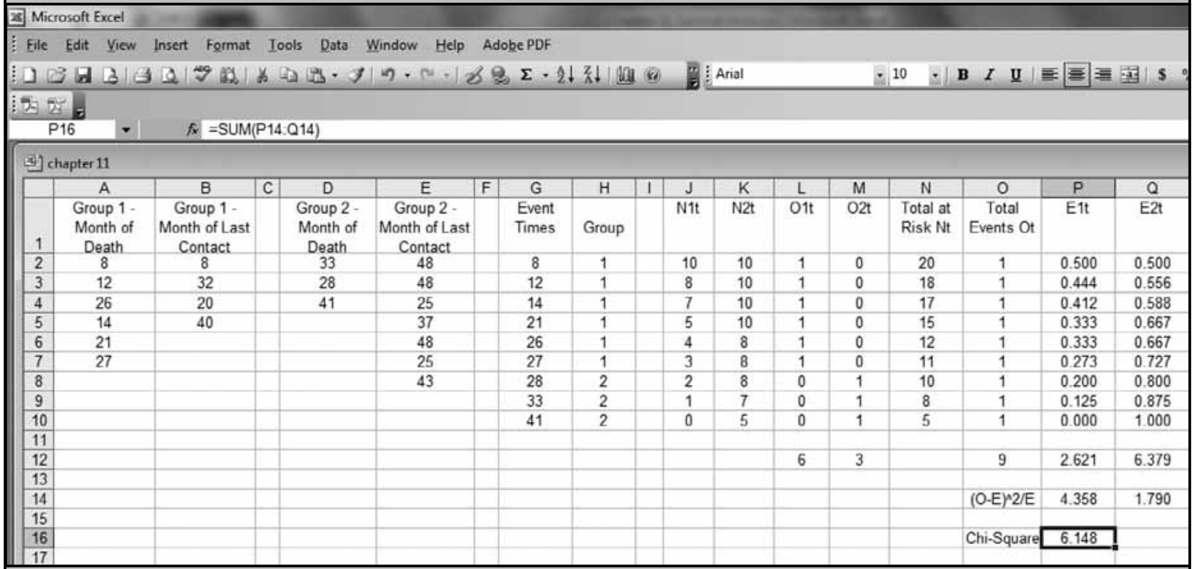
FIGURE 11-60 Computing the Log-Rank Test Statistic


Microsoft Excel
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial 10
chapter 11

Q14 =(M12-Q12)^2/Q12

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group		N1t	N2t	O1t	O2t	Total at Risk Nt	Total Events Ot	E1t	E2t
2	8	8		33	48		8	1		10	10	1	0	20	1	0.500	0.500
3	12	32		28	48		12	1		8	10	1	0	18	1	0.444	0.556
4	26	20		41	25		14	1		7	10	1	0	17	1	0.412	0.588
5	14	40			37		21	1		5	10	1	0	15	1	0.333	0.667
6	21				48		26	1		4	8	1	0	12	1	0.333	0.667
7	27				25		27	1		3	8	1	0	11	1	0.273	0.727
8					43		28	2		2	8	0	1	10	1	0.200	0.800
9							33	2		1	7	0	1	8	1	0.125	0.875
10							41	2		0	5	0	1	5	1	0.000	1.000
11																	
12												6	3		9	2.621	6.379
13																	
14															(O-E)^2/E	4.358	1.790
15																	

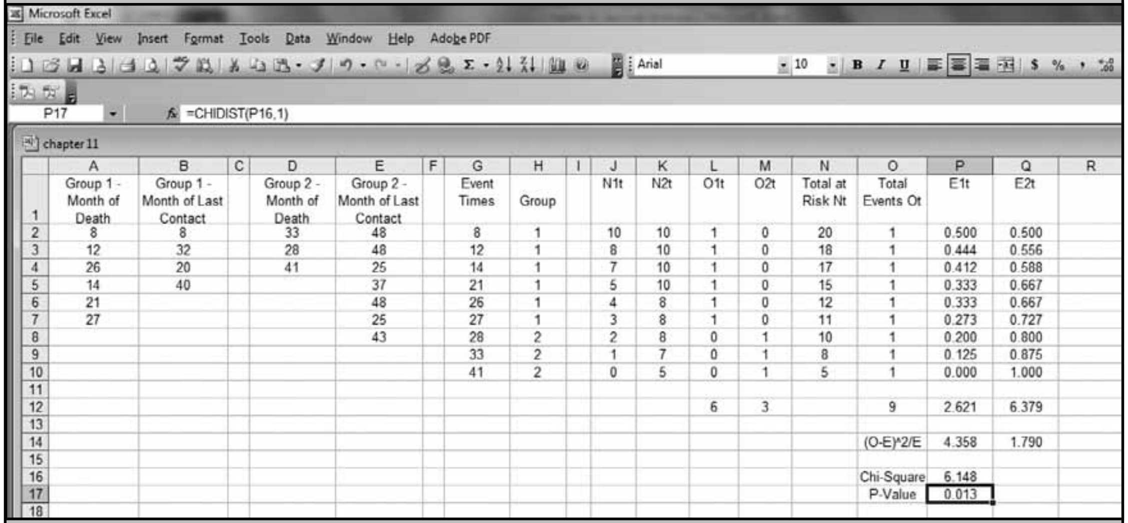
FIGURE 11-61 The Log-Rank Test Statistic


Microsoft Excel
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial 10
chapter 11

P16 =SUM(P14:Q14)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group		N1t	N2t	O1t	O2t	Total at Risk Nt	Total Events Ot	E1t	E2t
2	8	8		33	48		8	1		10	10	1	0	20	1	0.500	0.500
3	12	32		28	48		12	1		8	10	1	0	18	1	0.444	0.556
4	26	20		41	25		14	1		7	10	1	0	17	1	0.412	0.588
5	14	40			37		21	1		5	10	1	0	15	1	0.333	0.667
6	21				48		26	1		4	8	1	0	12	1	0.333	0.667
7	27				25		27	1		3	8	1	0	11	1	0.273	0.727
8					43		28	2		2	8	0	1	10	1	0.200	0.800
9							33	2		1	7	0	1	8	1	0.125	0.875
10							41	2		0	5	0	1	5	1	0.000	1.000
11																	
12												6	3		9	2.621	6.379
13																	
14															(O-E)^2/E	4.358	1.790
15																	
16															Chi-Square	6.148	
17																	

FIGURE 11-62 Computing the *p*-Value for the Log-Rank Test


Microsoft Excel
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
chapter 11

P17 =CHIDIST(P16,1)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Group 2 - Month of Death	Group 2 - Month of Last Contact		Event Times	Group		N1t	N2t	O1t	O2t	Total at Risk Nt	Total Events Ot	E1t	E2t	
2	8	8		33	48		8	1		10	10	1	0	20	1	0.500	0.500	
3	12	32		28	48		12	1		8	10	1	0	18	1	0.444	0.556	
4	26	20		41	25		14	1		7	10	1	0	17	1	0.412	0.588	
5	14	40			37		21	1		5	10	1	0	15	1	0.333	0.667	
6	21				48		26	1		4	8	1	0	12	1	0.333	0.667	
7	27				25		27	1		3	8	1	0	11	1	0.273	0.727	
8					43		28	2		2	8	0	1	10	1	0.200	0.800	
9							33	2		1	7	0	1	8	1	0.125	0.875	
10							41	2		0	5	0	1	5	1	0.000	1.000	
11																		
12												6	3		9	2.621	6.379	
13																		
14															(O-E)^2/E	4.358	1.790	
15																		
16															Chi-Square	6.148		
17															P-Value	0.013		
18																		

Table 11-4 and indicate whether women relapsed to drinking and if so, the time of their first drink measured in the number of weeks from randomization. For women who did not relapse, we recorded the number of weeks from randomization that they were alcohol-free.

The question of interest is whether there is a difference in time to relapse between women assigned to standard prenatal care as compared to those assigned to the brief intervention.

The test is run at a 5% level of significance and the hypotheses are as follows:

H_0: Relapse-free time is identical between groups.

H_1: Relapse-free time is not identical between groups.

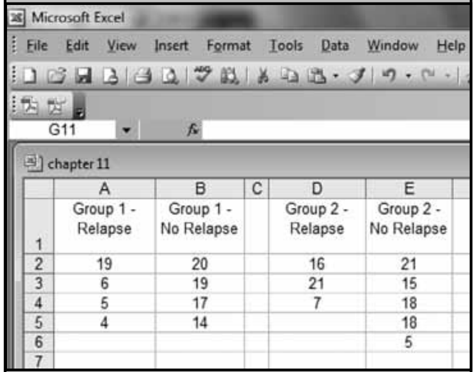
TABLE 11-4 Number of Weeks to First Drink or Number of Weeks Alcohol Free by Treatment Group

Standard Prenatal Care		Brief Intervention	
Relapse	No Relapse	Relapse	No Relapse
19	20	16	21
6	19	21	15
5	17	7	18
4	14		18
			5

To conduct the test, we first copy the worksheet with all of the formulas we developed to conduct the log-rank test in Example 11.7, and we enter the data in Table 11-4 into the worksheet as shown in Figure 11-63.

Next we copy the observed event times in column A and column D into column G and indicate which group (1 or 2) that each event occurred. We then sort the data by the event times and group variables as shown in Figure 11-51. The sorted data are shown in Figure 11-64.

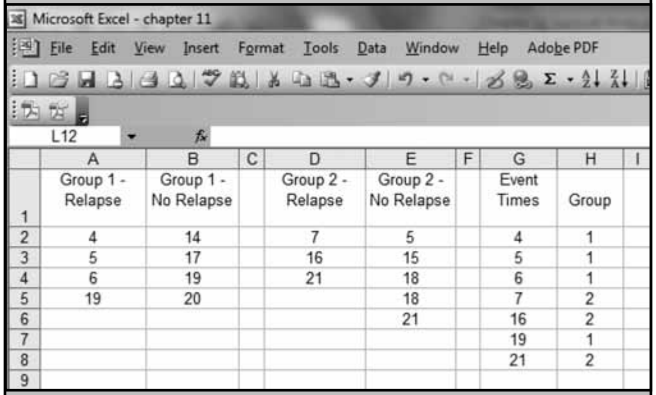
FIGURE 11-63 Data for the Log Rank Test



G11

	A	B	C	D	E
1	Group 1 - Relapse	Group 1 - No Relapse		Group 2 - Relapse	Group 2 - No Relapse
2	19	20		16	21
3	6	19		21	15
4	5	17		7	18
5	4	14			18
6					5
7					

FIGURE 11-64 Sorted Data for Analysis


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF

L12

	A	B	C	D	E	F	G	H	I
1	Group 1 - Relapse	Group 1 - No Relapse		Group 2 - Relapse	Group 2 - No Relapse		Event Times	Group	
2	4	14		7	5		4	1	
3	5	17		16	15		5	1	
4	6	19		21	18		6	1	
5	19	20			18		7	2	
6					21		16	2	
7							19	1	
8							21	2	
9									

Next, we update the formulas used to compute the numbers of participants at risk in each group, N_{1t} and N_{2t}, to reflect the data ranges for these data (i.e., A$2:B$5 for group 1 and D$2:E$6 for group 2). Once the data ranges are updated in cells J2 and K2, the formulas are copied into cell J3 through cell K8 (see formula in top menu bar for cell K8), as shown in Figure 11-65. Excel then computes the observed numbers of events, the total numbers of participants and events, and the expected numbers of events in each group at each event time. The test statistic is then computed along with a p-value.

For this example, $\chi^2 = 0.727$ (degrees of freedom $= k - 1 = 2 - 1 = 1$) and the p-value is $p = 0.394$. Because the p-value exceeds the level of significance ($\alpha = 0.05$), we do not reject H_0. We do not have statistically significant evidence at $\alpha = 0.05$ to show that the time to relapse is different between groups.

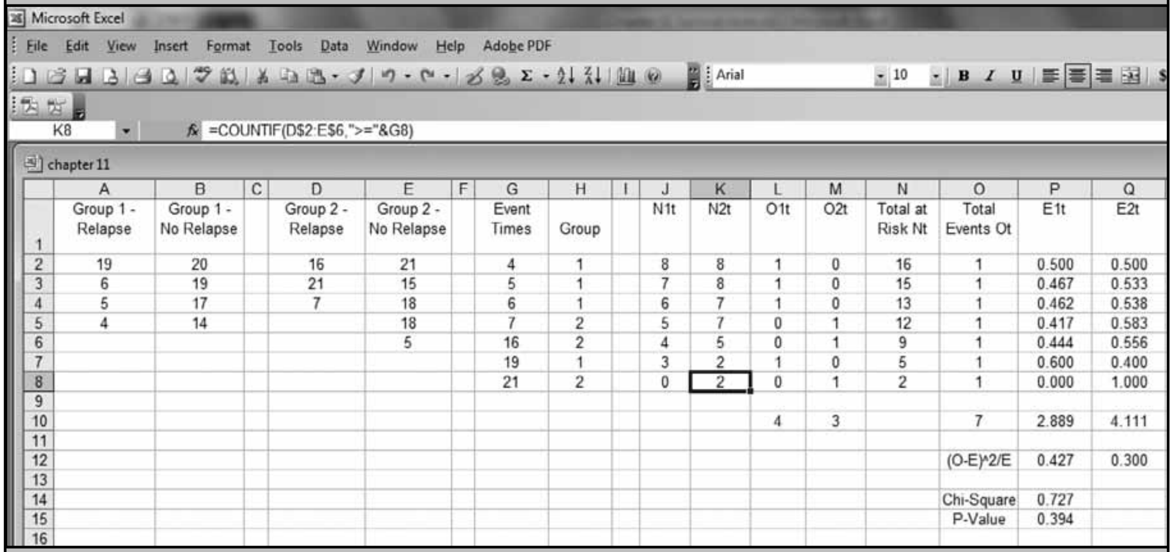
11.4 COMPARING TWO SURVIVAL CURVES GRAPHICALLY

In Section 11.2, we used Excel to generate graphical displays of a survival curve in one sample. We outlined the steps to produce graphical displays of the survival curve as well as displays that incorporated 95% confidence limits around the estimated survival probabilities. Here we present techniques to produce graphical displays of Kaplan–Meier survival functions in two independent groups using Excel.

Example 11.9. Consider again Example 11.7 where we compared survival between two competing combination treatments in patients with advanced gastric cancer. In Example 11.7 we conducted a log-rank test and found a statistically significant difference in survival between competing treatments ($p = 0.013$). Here we use Excel to produce a graphical display of the survival functions.

The first step is to produce the Kaplan–Meier estimates of the survival probabilities for each group using the template

FIGURE 11-65 The Log-Rank Test Statistic and *p*-Value


Microsoft Excel
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
B I U
chapter 11

K8 =COUNTIF(D$2:E$6,">="&G8)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Group 1 - Relapse	Group 1 - No Relapse		Group 2 - Relapse	Group 2 - No Relapse		Event Times	Group		N1t	N2t	O1t	O2t	Total at Risk Nt	Total Events Ot	E1t	E2t
2	19	20		16	21		4	1		8	8	1	0	16	1	0.500	0.500
3	6	19		21	15		5	1		7	8	1	0	15	1	0.467	0.533
4	5	17		7	18		6	1		6	7	1	0	13	1	0.462	0.538
5	4	14			18		7	2		5	7	0	1	12	1	0.417	0.583
6					5		16	2		4	5	0	1	9	1	0.444	0.556
7							19	1		3	2	1	0	5	1	0.600	0.400
8							21	2		0	2	0	1	2	1	0.000	1.000
9																	
10												4	3		7	2.889	4.111
11																	
12															(O-E)^2/E	0.427	0.300
13																	
14															Chi-Square	0.727	
15															P-Value	0.394	
16																	

we developed in Excel to compute survival probabilities using the Kaplan–Meier approach shown in Figure 11-20. We copy the worksheet (Figure 11-20) and enter the data for group 1. Here we need only to adjust the data ranges for the time and event variables (column D and column E) to produce the survival estimates shown in Figure 11-66.

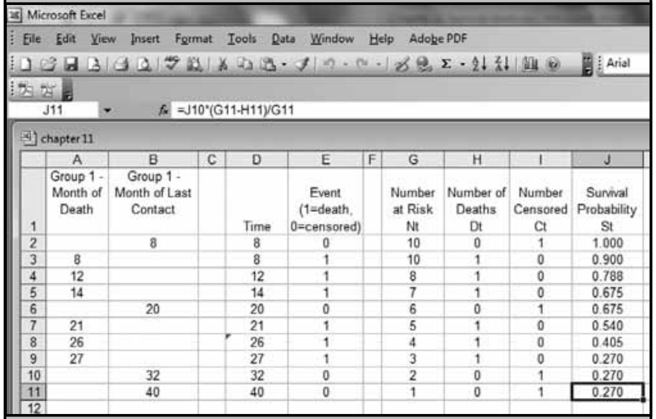
Next we copy the data for group 2 into the worksheet and again copy the formulas from row 2 through row 11 to produce the survival estimates for the data in group 2 (see specifically row 14 through row 23), as shown in Figure 11-67. We again need to update the data ranges for the computations. For example, to compute the number of participants at risk in group 2, we specify the data range as D$14:D$23 in cell G23 (see top menu bar).

Next we must prepare the data for the display. First, we hide the columns that are not directly involved in the display using the **Format Column** option on the top menu bar. We retain the time (column D) and the survival probability (column J). Next we convert the formulas we used to compute the time and survival probability variables to their calculated values by highlighting those columns, and clicking on the copy icon and then on the **Values** option to the right of the paste icon. These steps produce the data shown in Figure 11-68.

In order to produce a display with two survival curves, we need some additional formatting. First, we move the survival probability for group 2 into column K as shown in Figure 11-69, and we remove the header rows for the group 2 data (i.e., remove rows 12 and 13 in Figure 11-68) using the **Delete** option under **Edit** on the top menu bar. We also rename the columns to S_{1t} and S_{2t} to represent survival in groups 1 and 2, respectively.

Once again, we must format the data to produce the step functions for the survival probabilities. We first insert a row and enter "0" for the time and "1" for the survival probability in each group (i.e., enter "0" for time and "1" for S_{1t} in cell D2 and in cell J2 for group 1, and "0" for time and "1" for S_{2t} in cell D13 and in cell K13 for group 2). This is shown in Figure 11-70.

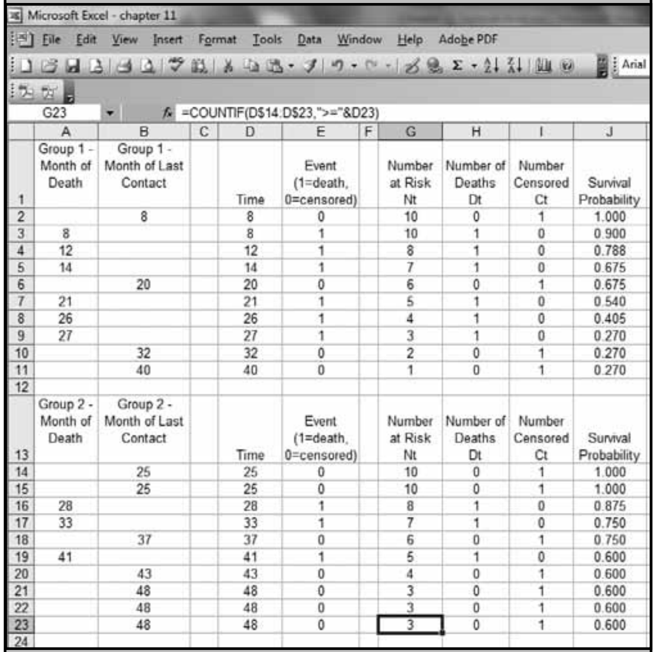
FIGURE 11-66 Computing Survival Probabilities for Group 1


Microsoft Excel
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
chapter 11

J11 =J10*(G11-H11)/G11

	A	B	C	D	E	F	G	H	I	J
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Time	Event (1=death, 0=censored)		Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability St
2		8		8	0		10	0	1	1.000
3	8			8	1		10	1	0	0.900
4	12			12	1		8	1	0	0.788
5	14			14	1		7	1	0	0.675
6		20		20	0		6	0	1	0.675
7	21			21	1		5	1	0	0.540
8	26			26	1		4	1	0	0.405
9	27			27	1		3	1	0	0.270
10		32		32	0		2	0	1	0.270
11		40		40	0		1	0	1	0.270
12										

FIGURE 11-67 Computing Survival Probabilities for Group 2


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial

G23 =COUNTIF(D$14:D$23,">="&D23)

	A	B	C	D	E	F	G	H	I	J
1	Group 1 - Month of Death	Group 1 - Month of Last Contact		Time	Event (1=death, 0=censored)		Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability
2		8		8	0		10	0	1	1.000
3	8			8	1		10	1	0	0.900
4	12			12	1		8	1	0	0.788
5	14			14	1		7	1	0	0.675
6		20		20	0		6	0	1	0.675
7	21			21	1		5	1	0	0.540
8	26			26	1		4	1	0	0.405
9	27			27	1		3	1	0	0.270
10		32		32	0		2	0	1	0.270
11		40		40	0		1	0	1	0.270
12										
13	Group 2 - Month of Death	Group 2 - Month of Last Contact		Time	Event (1=death, 0=censored)		Number at Risk Nt	Number of Deaths Dt	Number Censored Ct	Survival Probability
14		25		25	0		10	0	1	1.000
15		25		25	0		10	0	1	1.000
16	28			28	1		8	1	0	0.875
17	33			33	1		7	1	0	0.750
18		37		37	0		6	0	1	0.750
19	41			41	1		5	1	0	0.600
20		43		43	0		4	0	1	0.600
21		48		48	0		3	0	1	0.600
22		48		48	0		3	0	1	0.600
23		48		48	0		3	0	1	0.600
24										

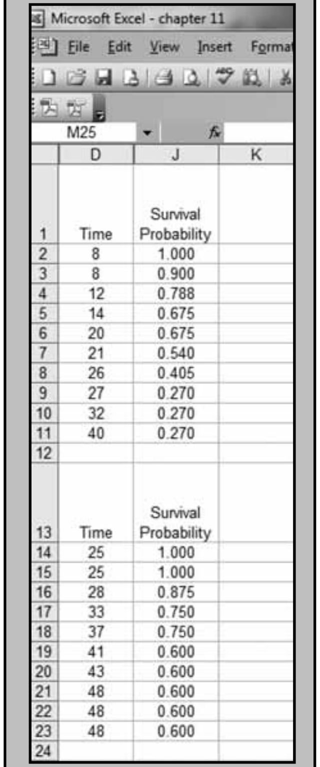
Next we insert rows at each transition point (i.e., changes in survival probability), and we then copy the time from the row above and the survival probability from the row below. The updated worksheet is shown in Figure 11-71.

We now use the Chart Wizard to create the display. We highlight the data (column D, column J, and column K) and select the chart icon in the top menu bar. Once we select the Chart Wizard, Excel opens a dialog box where we select the chart type XY (Scatter). This is shown in Figure 11-72.

Once we click Next, Excel then asks for a chart title as well as labels for the *x*- and *y*-axes. We enter this information and indicate that we would like to place the display in a new worksheet. We click Finish, and Excel generates the new worksheet containing the display shown in Figure 11-73.

The final step is to format the display for presentation. This involves changing the background from grey to white or no background color, removing the horizontal lines, rescaling the *y*-axis to a maximum of 1.0, rescaling the *x*-axis to a maximum of 50 months, and changing the line colors to black (solid and dashed to distinguish the groups). The final display is shown in Figure 11-74.

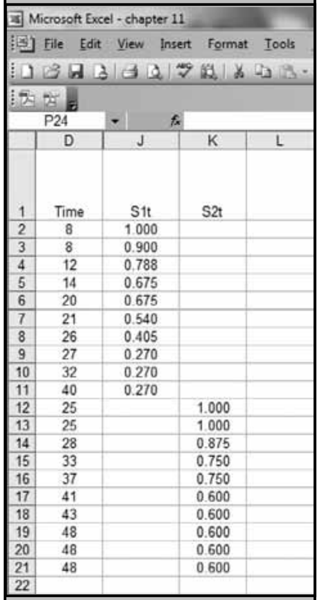
FIGURE 11-68 Data for the Display



M25

	D	J	K
1	Time	Survival Probability	
2	8	1.000	
3	8	0.900	
4	12	0.788	
5	14	0.675	
6	20	0.675	
7	21	0.540	
8	26	0.405	
9	27	0.270	
10	32	0.270	
11	40	0.270	
12			
13	Time	Survival Probability	
14	25	1.000	
15	25	1.000	
16	28	0.875	
17	33	0.750	
18	37	0.750	
19	41	0.600	
20	43	0.600	
21	48	0.600	
22	48	0.600	
23	48	0.600	
24			

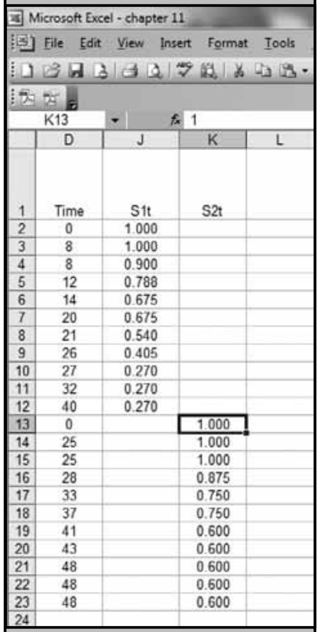
FIGURE 11-69 Formatting the Data for the Display



P24

	D	J	K	L
1	Time	S1t	S2t	
2	8	1.000		
3	8	0.900		
4	12	0.788		
5	14	0.675		
6	20	0.675		
7	21	0.540		
8	26	0.405		
9	27	0.270		
10	32	0.270		
11	40	0.270		
12	25		1.000	
13	25		1.000	
14	28		0.875	
15	33		0.750	
16	37		0.750	
17	41		0.600	
18	43		0.600	
19	48		0.600	
20	48		0.600	
21	48		0.600	
22				

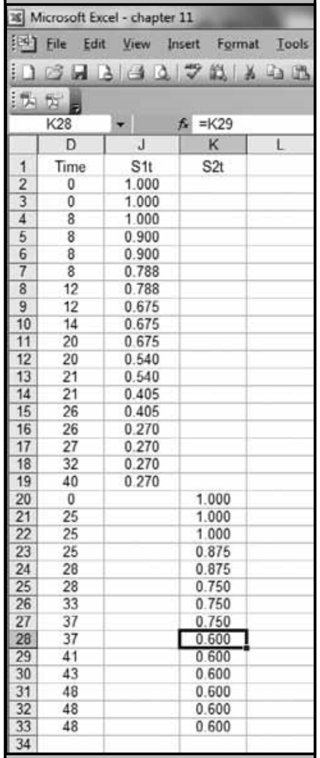
FIGURE 11-70 Data for the Display



K13 fx 1

	D	J	K	L
1	Time	S1t	S2t	
2	0	1.000		
3	8	1.000		
4	8	0.900		
5	12	0.788		
6	14	0.675		
7	20	0.675		
8	21	0.540		
9	26	0.405		
10	27	0.270		
11	32	0.270		
12	40	0.270		
13	0		1.000	
14	25		1.000	
15	25		1.000	
16	28		0.875	
17	33		0.750	
18	37		0.750	
19	41		0.600	
20	43		0.600	
21	48		0.600	
22	48		0.600	
23	48		0.600	
24				

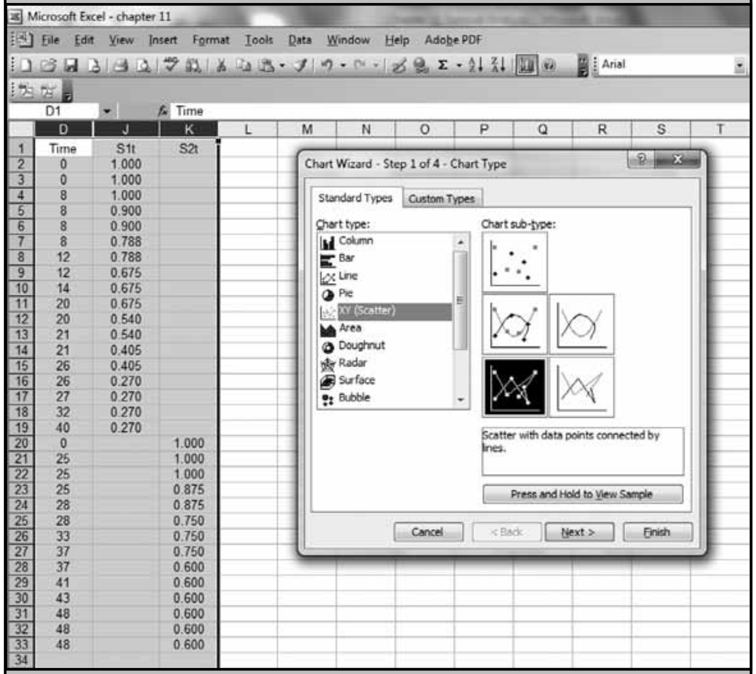
FIGURE 11-71 Data to Create Step Functions



K28 fx =K29

	D	J	K	L
1	Time	S1t	S2t	
2	0	1.000		
3	0	1.000		
4	8	1.000		
5	8	0.900		
6	8	0.900		
7	8	0.788		
8	12	0.788		
9	12	0.675		
10	14	0.675		
11	20	0.675		
12	20	0.540		
13	21	0.540		
14	21	0.405		
15	26	0.405		
16	26	0.270		
17	27	0.270		
18	32	0.270		
19	40	0.270		
20	0		1.000	
21	25		1.000	
22	25		1.000	
23	25		0.875	
24	28		0.875	
25	28		0.750	
26	33		0.750	
27	37		0.750	
28	37		0.600	
29	41		0.600	
30	43		0.600	
31	48		0.600	
32	48		0.600	
33	48		0.600	
34				

FIGURE 11-72 Using the Chart Wizard


Microsoft Excel - chapter 11
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
D1
Time
D J K L M N O P Q R S T

	D	J	K
1	Time	S1t	S2t
2	0	1.000	
3	0	1.000	
4	8	1.000	
5	8	0.900	
6	8	0.900	
7	8	0.788	
8	12	0.788	
9	12	0.675	
10	14	0.675	
11	20	0.675	
12	20	0.540	
13	21	0.540	
14	21	0.405	
15	26	0.405	
16	26	0.270	
17	27	0.270	
18	32	0.270	
19	40	0.270	
20	0		1.000
21	25		1.000
22	25		1.000
23	25		0.875
24	28		0.875
25	28		0.750
26	33		0.750
27	37		0.750
28	37		0.600
29	41		0.600
30	43		0.600
31	48		0.600
32	48		0.600
33	48		0.600
34			

Chart Wizard - Step 1 of 4 - Chart Type

Standard Types | Custom Types

Chart type: Column, Bar, Line, Pie, XY (Scatter), Area, Doughnut, Radar, Surface, Bubble

Chart sub-type:

Scatter with data points connected by lines.

FIGURE 11-73 The Display

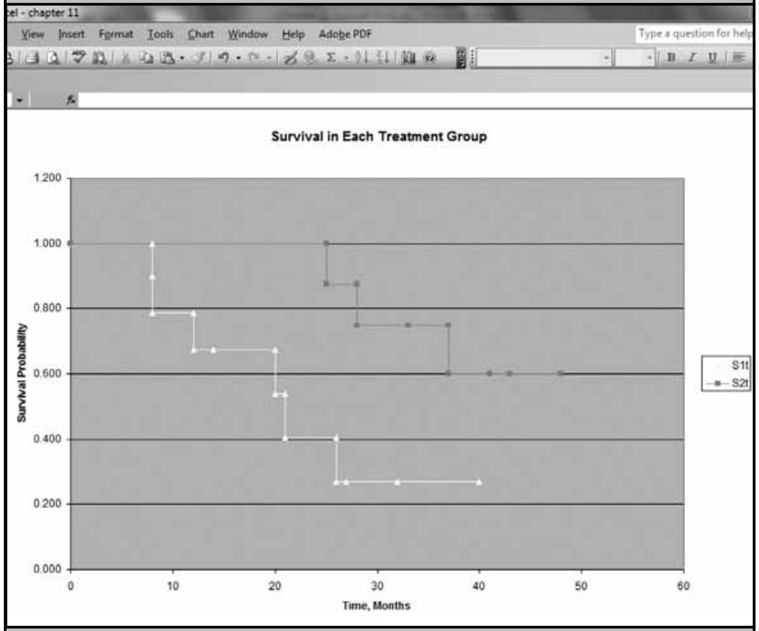
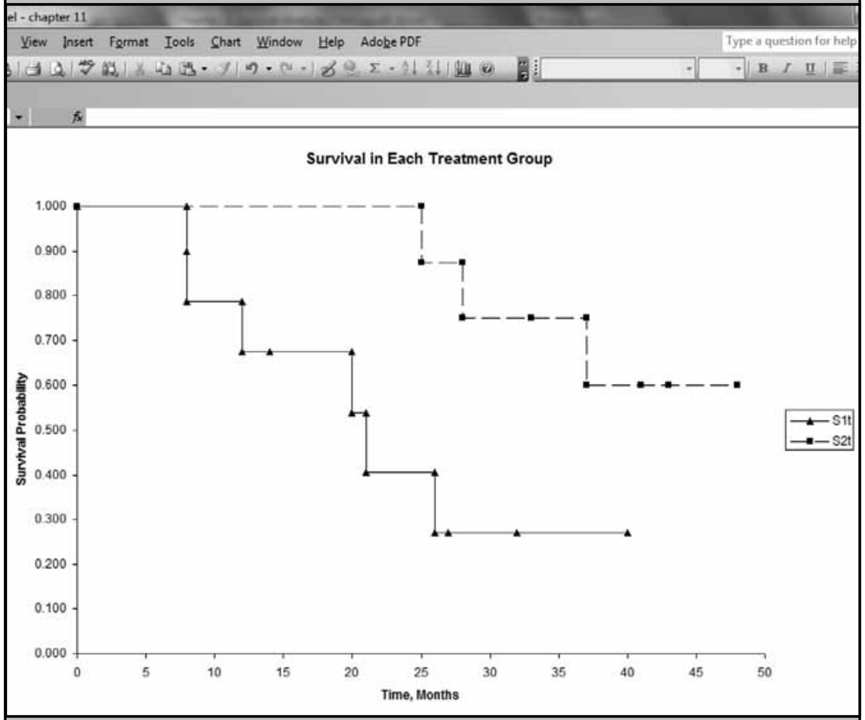


FIGURE 11-74 Survival in Each Treatment Group



11.5 PRACTICE PROBLEMS

1. A study is conducted to estimate survival in patients following kidney transplant. Key factors that adversely affect success of the transplant include advanced age and diabetes. This study involves 25 participants ($n = 25$) who are 65 years of age and older, and all have diabetes. Following transplant, each participant is followed for up to 10 years. The following are times to death, in years, or the time to last contact (at which time participant was known to be alive).

Deaths: 1.2, 2.5, 4.3, 5.6, 6.7, 7.3, and 8.1 years
Alive: 3.4, 4.1, 4.2, 5.7, 5.9, 6.3, 6.4, 6.5, 7.3, 8.2, 8.6, 8.9, 9.4, 9.5, 10, 10, 10, and 10 years

   a. Use the life table approach to estimate the survival function.
   b. Use the Kaplan–Meier approach to estimate the survival function.
   c. Graph the survival function based on the estimates in (b) using Excel.

2. A clinical trial is run to assess the effectiveness of a new anti-arrhythmic drug designed to prevent atrial fibrillation (AF). Thirty participants ($n = 30$) enroll in the trial and are randomized to receive the new drug or placebo. The primary outcome is AF and participants are followed for up to 12 months following randomization. The experiences of participants in each arm of the trial are shown in Table 11-5.
   a. Estimate the survival functions for each treatment group using the Kaplan–Meier approach.
   b. Test whether there is a significant difference in survival between treatment groups using the log-rank test and a 5% level of significance.
3. Using the results from Problem 2, sketch the survival functions for the new anti-arrhythmic drug and the placebo groups.
4. An observational cohort study is conducted to compare time to early failure in patients undergoing joint replacement surgery. Of specific interest is whether there is a difference in time to early failure between

TABLE 11-5 Data for Practice Problems 2 and 3

Placebo		New Drug	
Month of AF	Month of Last Contact	Month of AF	Month of Last Contact
4	5	7	6
6	5	9	6
7	6	12	7
8	7		8
9	8		9
11	9		9
	11		10
	12		10
	12		11
			11
			12
			12

patients who are considered obese versus those who are not. The study is run for 40 weeks, and times to early joint failure, measured in weeks, are shown in Table 11-6 for participants classified as obese or not at the time of surgery.

a. Estimate the survival functions (time to early joint failure) for each group using the Kaplan–Meier approach.

b. Test whether there is a significant difference in time to early joint failure between obese and non-obese patients undergoing joint replacement surgery using the log-rank test and a 5% level of significance.

TABLE 11-6 Data for Practice Problem 4

Obese		Not Obese	
Failure	No Failure	Failure	No Failure
28	39	27	37
25	41	31	36
31	37	34	39
32	35		40
	38		36
	36		32
	29		39
			41

5. Using the results from Problem 4, sketch the survival functions for each group (obese and non-obese).

6. A study of patients with stage I breast cancer is run to assess time to progression to stage II over an observation period of 15 years. The data are shown in Table 11-7. Times to progression are measured in years from the time at which the chemotherapy regimen was initiated. Of interest is whether there is a difference in time to progression between women on two different chemotherapy regimens.

   a. Estimate the survival functions (time to progression) for each chemotherapy regimen using the Kaplan–Meier approach.

   b. Test whether there is a significant difference in time to progression between treatment regimens using the log-rank test and a 5% level of significance.

7. A clinical trial is conducted to evaluate the efficacy of a new drug for prevention of hypertension in patients with pre-hypertension (defined as systolic blood pressure between 120 mmHg and 139 mmHg or diastolic blood pressure between 80 mmHg and 89 mmHg). A total of 20 patients are randomized to receive the new drug or a currently available drug for treatment of high blood pressure. Participants are followed for up to 12 months, and time to progression to hypertension is measured. The experiences of participants in each arm of the trial are shown in Table 11-8.

TABLE 11-7 Data for Practice Problem 6

Regimen 1		Regimen 2	
Progression	No Progression	Progression	No Progression
2	12	9	11
6	14	4	14
7	13	7	13
3	11		9
4	15		14
	10		13
	8		6
	6		9
	9		7
	12		

TABLE 11-8 Data for Practice Problem 7

New Drug		Currently Available Drug	
Hypertension	Free of Hypertension	Hypertension	Free of Hypertension
7	8	6	8
8	8	7	9
10	8	9	11
	9	10	11
	11	11	12
	12		
	12		

a. Estimate the survival functions (time to progression to hypertension) for each treatment group using the Kaplan–Meier approach.

b. Test whether there is a significant difference in time to progression between treatment groups using the log-rank test and a 5% level of significance.

8. The data in Table 11-9 reflect the time to first surgery in children born with congenital heart disease. Time is measured in years from birth up until the age of 10 years. Construct a life table using the Kaplan–Meier approach. Also include standard errors and 95% confidence limits for the estimates of survival probability.

TABLE 11-9 Data for Practice Problem 8

Participant Identification Number	Year of First Surgery	Year of Last Contact
1	8	
2		10
3		4
4	4	
5		7
6		6
7		9
8		5
9	3	
10		8
11		9
12		10
13		3
14	2	
15	6	
16		7
17		8
18		9

REFERENCE

1. Allison, P. *Survival Analysis Using the SAS System*, Cary, NC. SAS Institute, 1995.