lects 2,3,4,5,6,7,8,9,10 and 11
Dr.Sabah Auda

Introduction to Statistics & Data Analysis

FIFTH EDITION

Roxy Peck
California Polytechnic State University,
San Luis Obispo, CA

Chris Olsen
Grinnell College
Grinnell, IA

Jay Devore
California Polytechnic State University,
San Luis Obispo, CA

Prepared by

Melissa M. Sovak
California University of Pennsylvania, California, PA

CENGAGE Learning®

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

ISBN-13: 978-1-305-26891-3
ISBN-10: 1-305-26891-1

Cengage Learning
20 Channel Center Street, 4th Floor
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at: **www.cengage.com/global**.

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage Learning Solutions, visit **www.cengage.com**.

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com**.

Printed in the United States of America
1 2 3 4 5 6 7 17 16 15 14 13

Table of Contents

Introduction To Excel

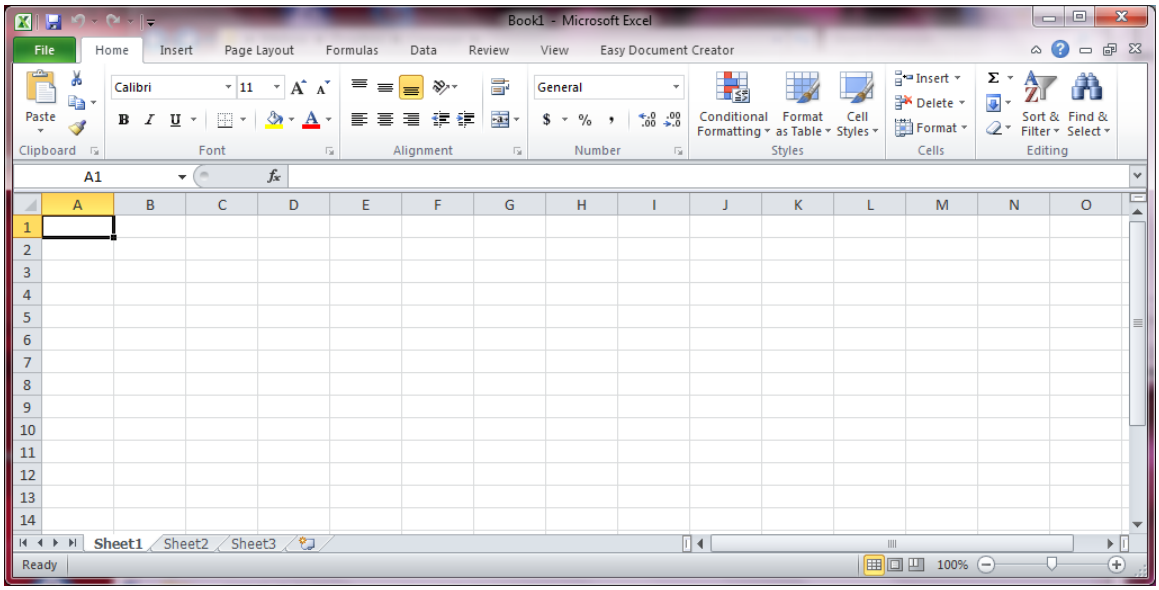
Getting Started with Excel

This chapter covers the basic structure and commands of Excel for Windows. After reading this chapter you should be able to:

1. Start Excel
2. Identify the Main Menu Ribbon
3. Install the Data Analysis Tool (if needed)
4. Enter Data in Excel
5. Save the Data File
6. Exit Excel

Starting Excel

Excel is a computer software program designed to create spreadsheets. It also includes many formulas and other tools to complete statistical analyses. You can start Excel by finding it in your program list or clicking on the Excel icon on your desktop. Once you start Excel, you should see a window like the one pictured below:



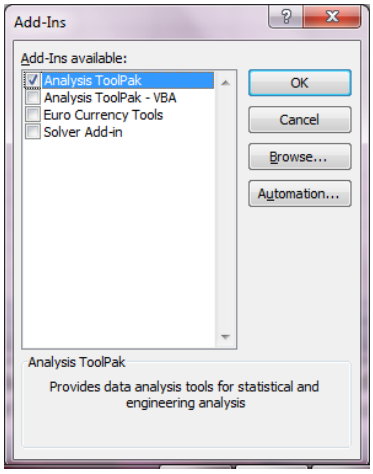
Across the top of the screen, there are several menu options that lead to submenus to run data analysis. These menus are

File Home Insert Page Layout Formulas Data Review View

Installing the Data Analysis Tool

One tool that we will use heavily to complete data analysis is the Data Analysis package. This can be found under the Data tab. If you do not see Data Analysis when you click on the Data tab, you will need to install it.

To install this package, click on the **File** tab and click **Options**. Click **Add-ins** and click **Go** near the bottom of the screen. Check the box next to Analysis Toolpak and click **OK.**



Entering Data

Excel is a spreadsheet. Typically, a column contains the data for one variable with each individual observation being in a row. Columns are designated A, B, C ... You can then type data in directly by inputting values into a particularly cell and pressing Enter.

You can also read in datasets from the text by using the **File>Open** command. This will open a dialog box from which you can navigate to the file, select it and click Open to display it in in Excel.

Saving Files

To save a Worksheet, select the worksheet to make it active. Then choose **File>Save As…**

This will open a dialog box that will allow you to navigate to the appropriate folder to save the file. Once you have found the appropriate folder, type a filename into the box title File name: and click Save.

Excel uses the file extension .xlsx to save files.

Exiting Excel

To exit Excel, choose **File>Exit**.

Chapter 1
The Role of Statistics and the Data Analysis Problem

One of the most useful ways to begin an initial exploration of data is to use techniques that result in a graphical representation of the data. These can quickly reveal characteristics of the variable being examined. There are a variety of graphical techniques used for variables.

In this chapter, we will learn how to create a frequency bar chart for categorical data. We will use the tab

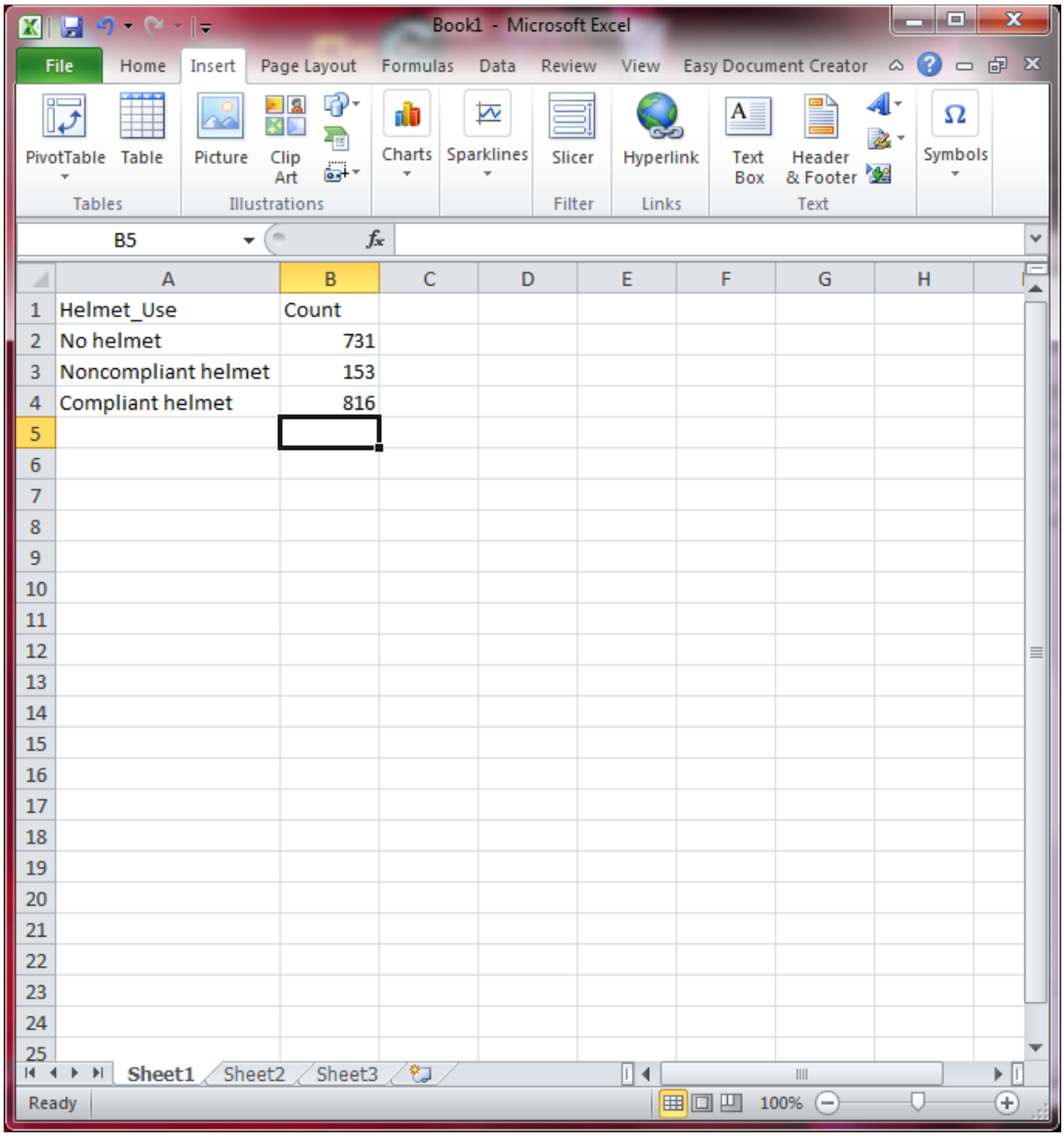
Insert

to create the bar chart.

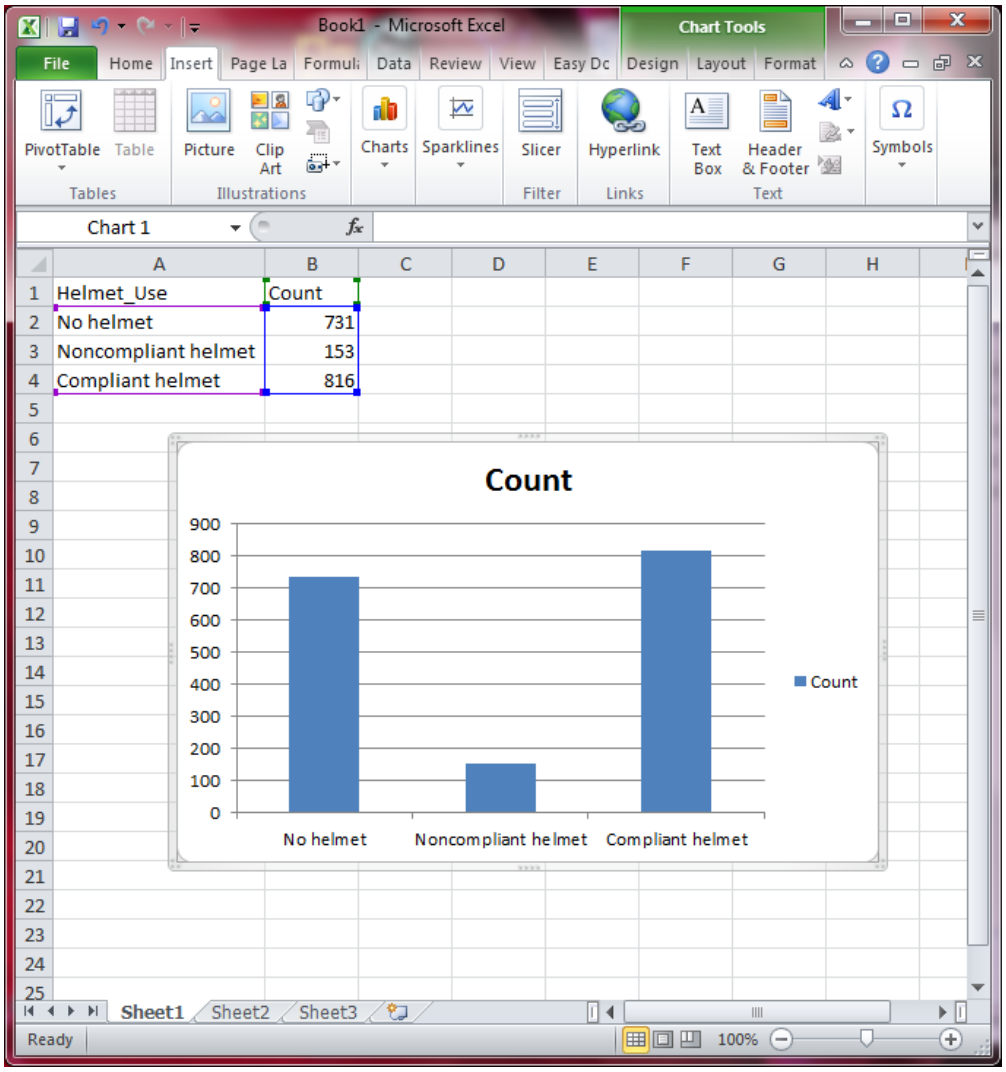
Example 1.8 – Bar Chart

The U.S. Department of Transportation establishes standards for motorcycle helmets. To ensure a certain degree of safety, helmets should reach the bottom of the motorcyclist's ears. The report "Motorcycle Helmet Use in 2005 – Overall Results" (National Highway Traffic Safety Administration, August 2005) summarized data collected in June of 2005 by observing 1700 motorcyclists nationwide at selected roadway locations. In total, there were 731 riders who wore no helmet, 153 who wore a noncompliant helmet, and 816 who wore a compliant helmet.

In this example, we will use the motorcycle helmet data to create a bar chart. Begin by entering the data into the Excel spreadsheet. Enter the categories for helmet use into column A and title it Helmet_Use. Enter the frequencies into B, titling it Count. Once the dataset is entered, you screen should look like the image below.



Next, we will create the bar chart. Click **Insert** and select **Charts>Column** and select the first option under 2-D Column. Excel will automatically select the data entered and create a bar chart. You should see the following bar chart.


Book1 - Microsoft Excel
Chart Tools
File
Home
Insert
Page La
Formul
Data
Review
View
Easy Dc
Design
Layout
Format
PivotTable
Table
Picture
Clip Art
Charts
Sparklines
Slicer
Hyperlink
Text Box
Header & Footer
Symbols
Tables
Illustrations
Filter
Links
Text
Chart 1
Helmet_Use
Count
No helmet
731
Noncompliant helmet
153
Compliant helmet
816
Count
900
800
700
600
500
400
300
200
100
0
No helmet
Noncompliant helmet
Compliant helmet
Count
Sheet1
Sheet2
Sheet3
Ready
100%

Chapter 2
Collecting Data Sensibly

Section 2.2 of the text discusses random sampling. We can use Excel to create a random sample of data. We can create many types of random samples using the formula

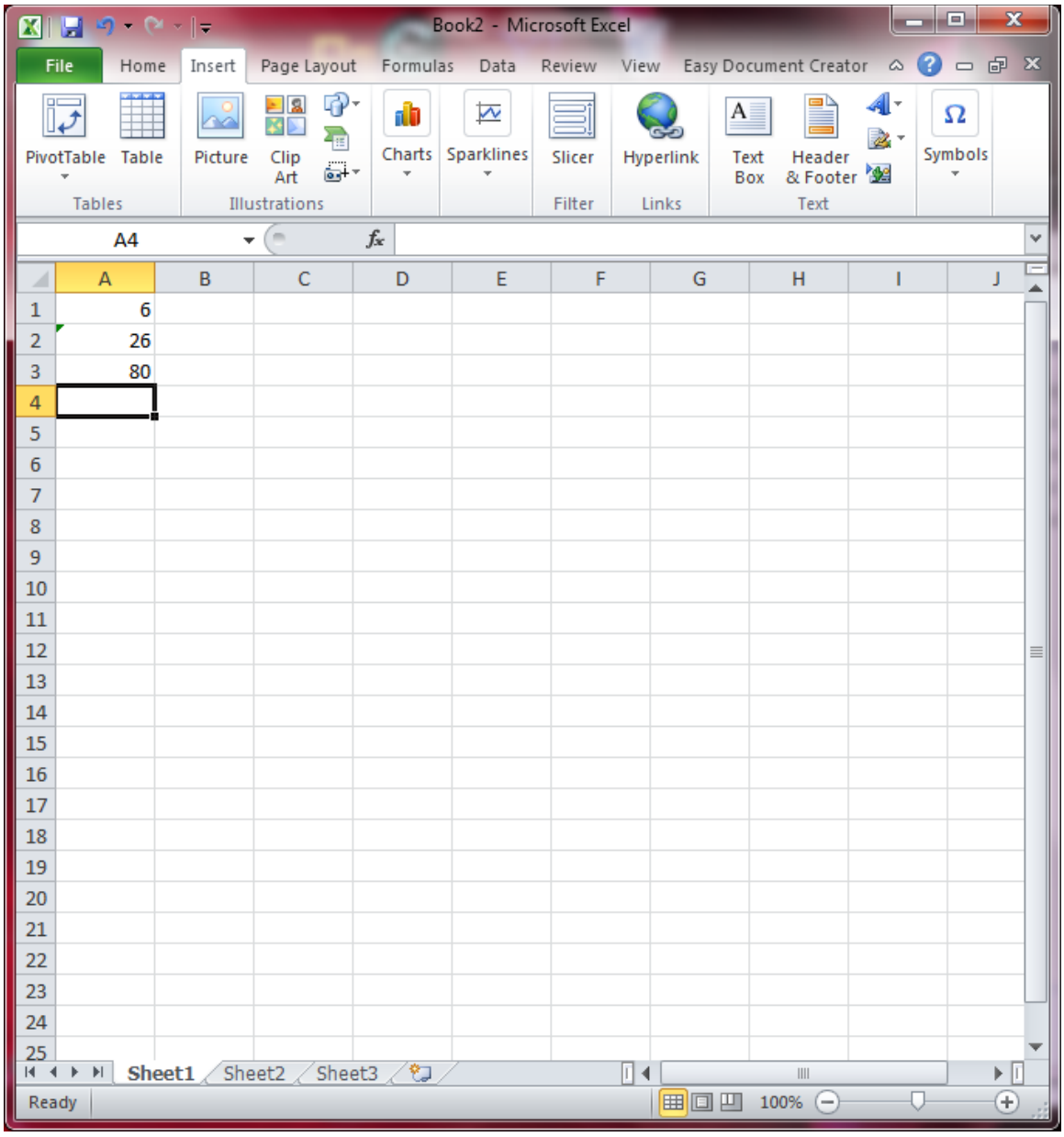
RANDBETWEEN

Example 2.3 – Selecting a Random Sample

Breaking strength is an important characteristic of glass soda bottles. Suppose that we want to measure the breaking strength of each bottle in a random sample of size $n = 3$ selected from four crates containing a total of 100 bottles (the population).

We will start with an empty spreadsheet in Excel. To randomly select 3 bottles from bottles numbered from 1 to 100, we click in cell A1 and type =RANDBETWEEN(1,100). Press **Enter**. Do this two more times in cells A2 and A3.

The results will be displayed as below. Note: Your results will be different from those below.


Book2 - Microsoft Excel
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Easy Document Creator
PivotTable
Table
Picture
Clip Art
Charts
Sparklines
Slicer
Hyperlink
Text Box
Header & Footer
Symbols
Tables
Illustrations
Filter
Links
Text
A4
6
26
80
Sheet1
Sheet2
Sheet3
Ready
100%

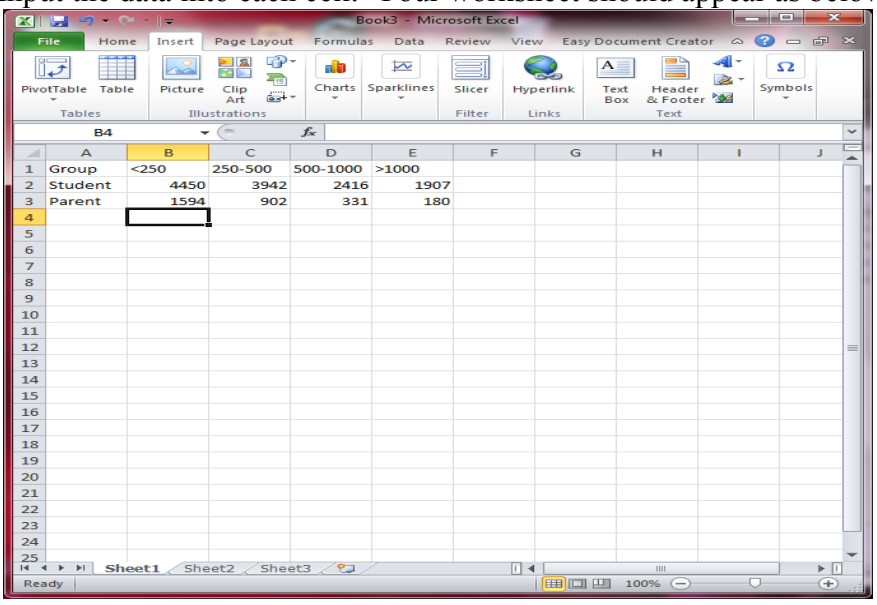
Chapter 3
Graphical Methods for Describing Data

Chapter 3 describes a number of graphical displays for both categorical and quantitative data. In this chapter, we will use the **Insert>Charts** menu to create several different types of graphs.

Example 3.1 – Comparative Bar Charts

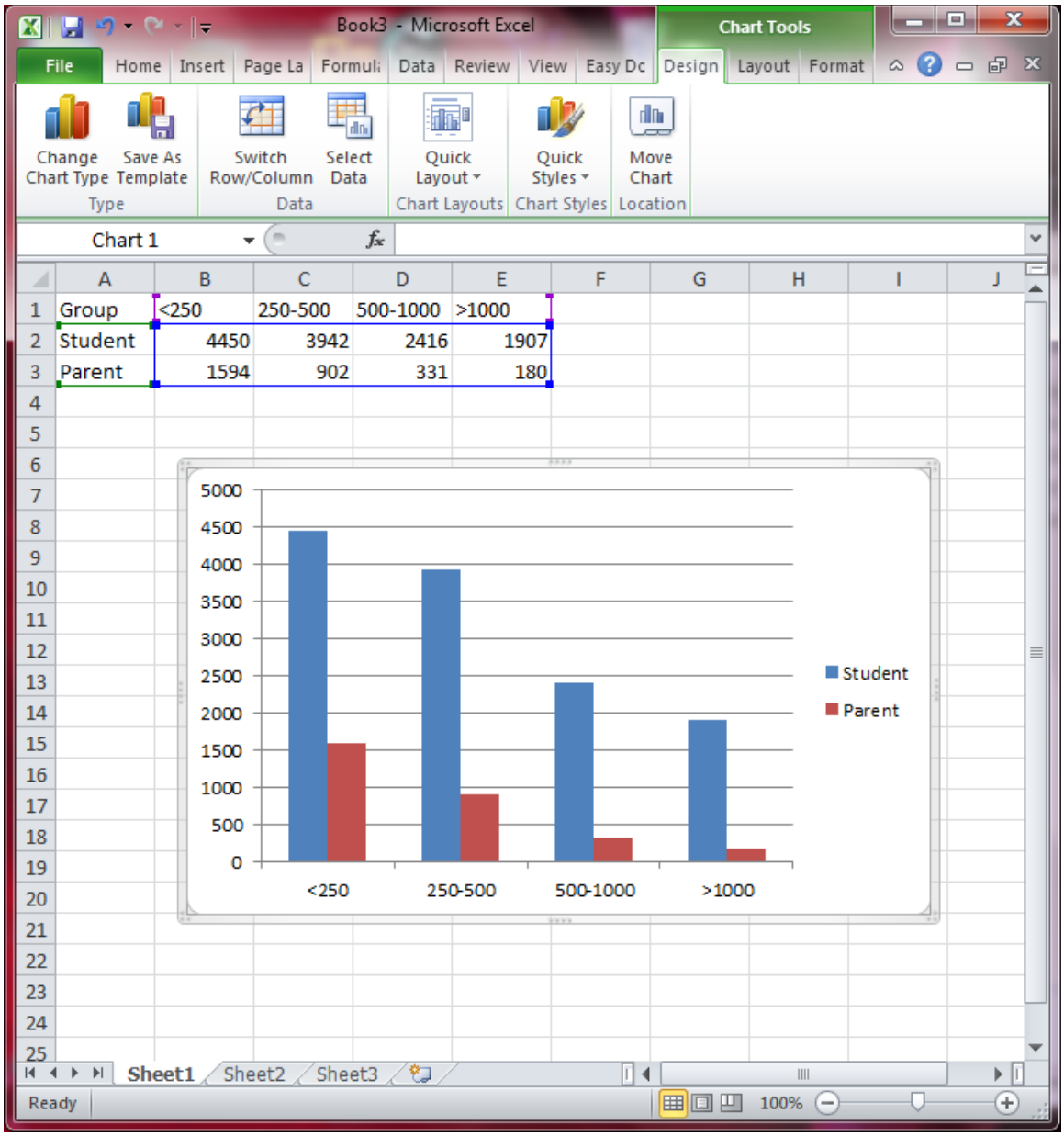
Each year The Princeton Review conducts a survey of high school students who are applying to college and parents of college applicants. The report "2009 College Hopes & Worries Survey Findings" (www.princetonreview.com/uploadedFiles/Test_Preparation/Hopes_and_Worries/colleg_hopes_worries_details.pdf) included a summary of how 12,715 high school students responded to the question "Ideally how far from home would you like the college you attend to be?" Also included was a summary of how 3007 parents of students applying to college responded to the question "How far from home would you like the college your child attends to be?"

We begin by entering the data table. Enter Group into C1 and enter values "Student and Parent". Enter into C2-C5 the variable titles "<250", "250-500", "500-1000" and ">1000". Input the data into each cell. Your worksheet should appear as below.



Click **Insert>Charts>Column** and select the first option under 2-D Column**.** Excel will automatically select the data table input and create the comparative bar chart.

The comparative bar chart is output as below.

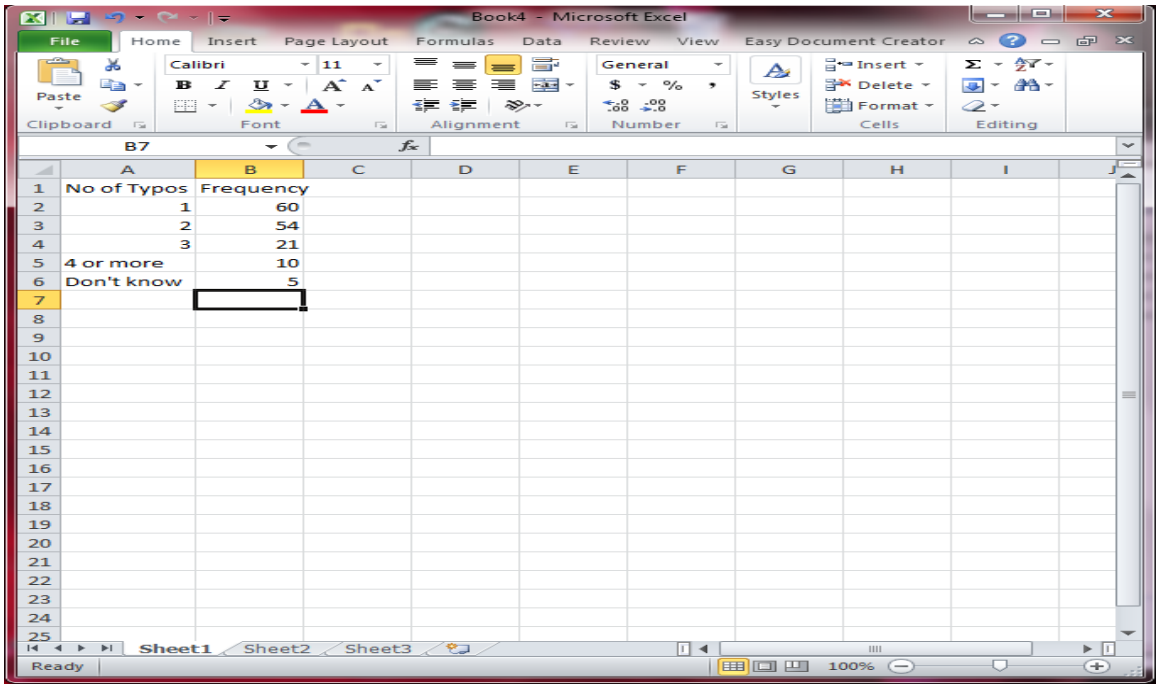


Example 3.3 – Pie Charts

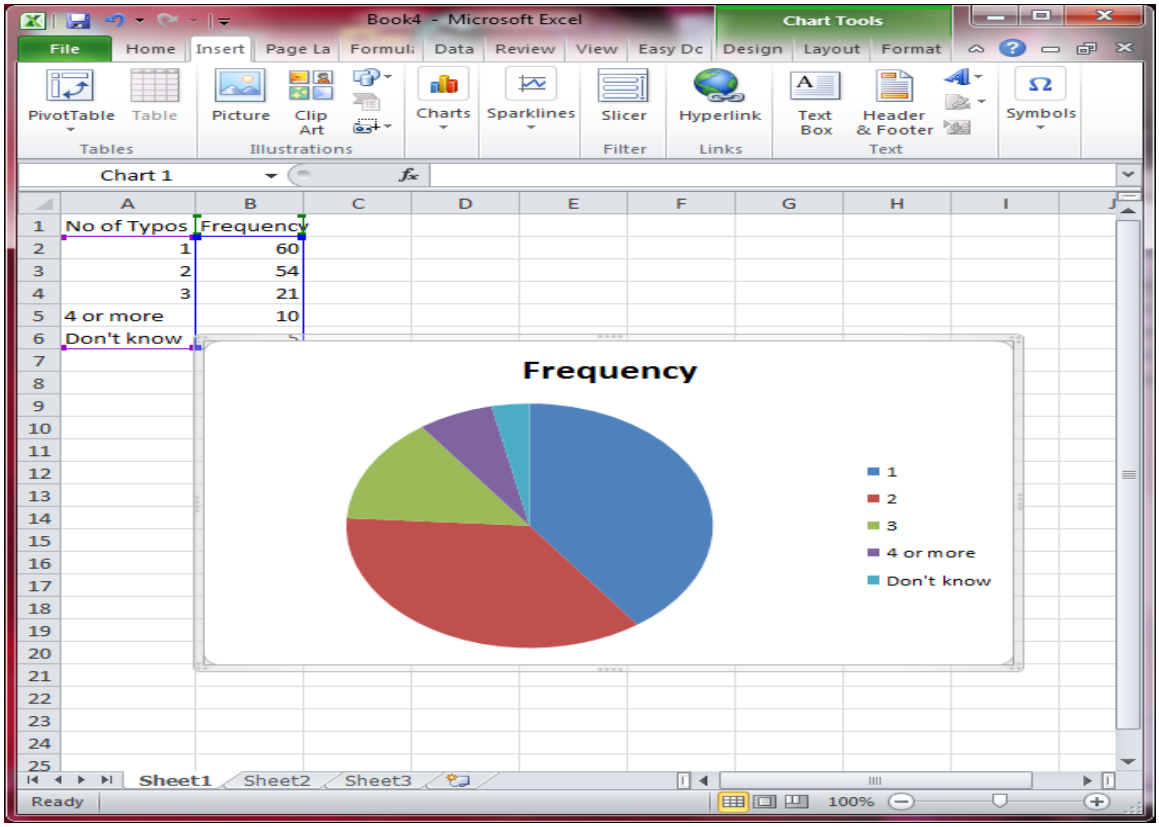
Typos on a resume do not make a very good impression when applying for a job. Senior executives were asked how many typos in a resume would make them not consider a job candidate ("Job Seekers Need a Keen Eye,", *USA Today*, August 3, 2009).

Begin by inputting the dataset by entering Number of Typos into column A and the Frequency for each into column B.

Your worksheet should look like the one below.



To create the pie chart, use the command **Insert>Charts>Pie** and select the first option under 2-D Pie. Excel will automatically select the data table and produce the chart below.



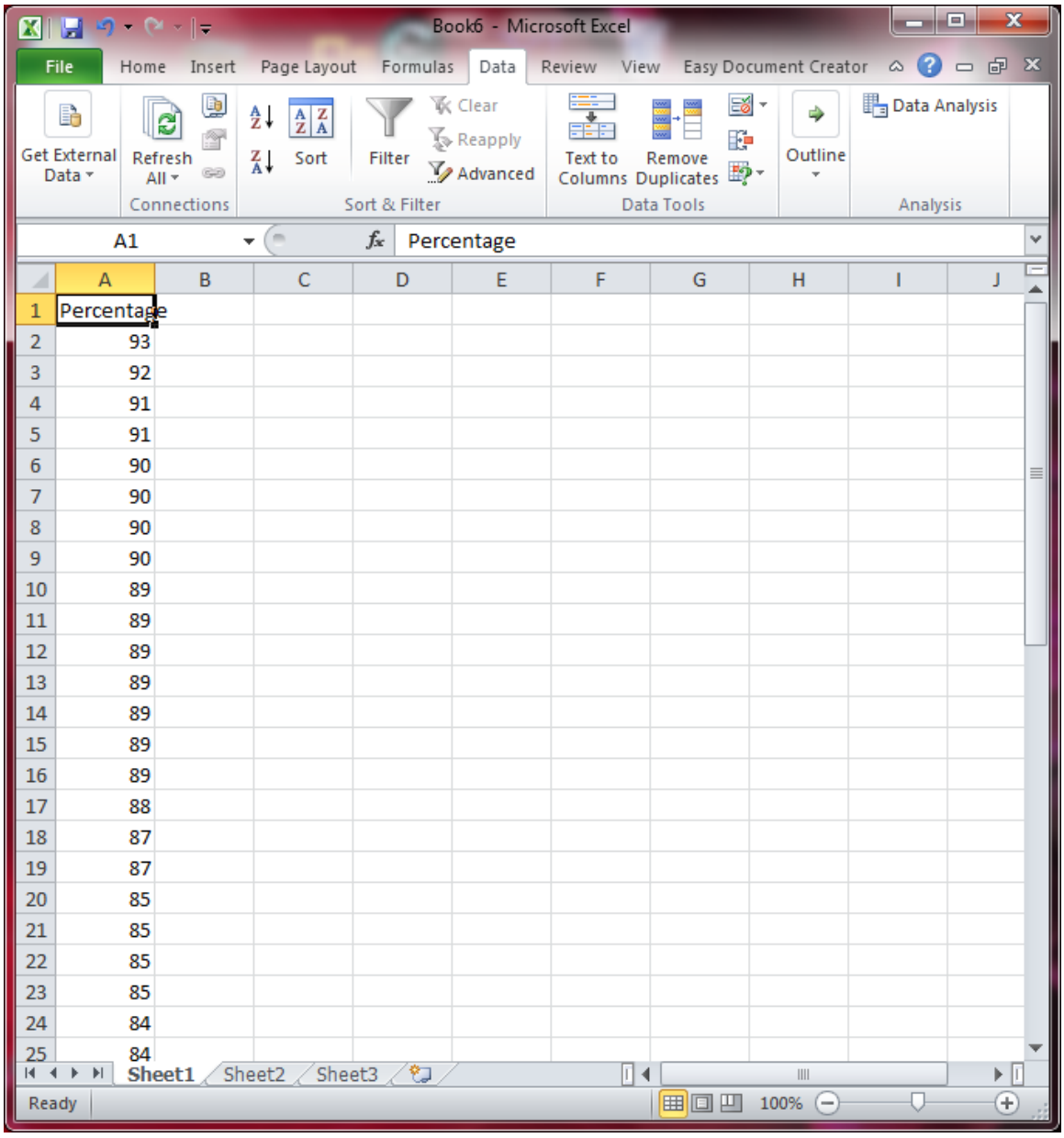
Example 3.8 – Stem-and-Leaf Displays

Excel does not support automatic creation of stem-and-leaf displays.

Example 3.15 – Histogram

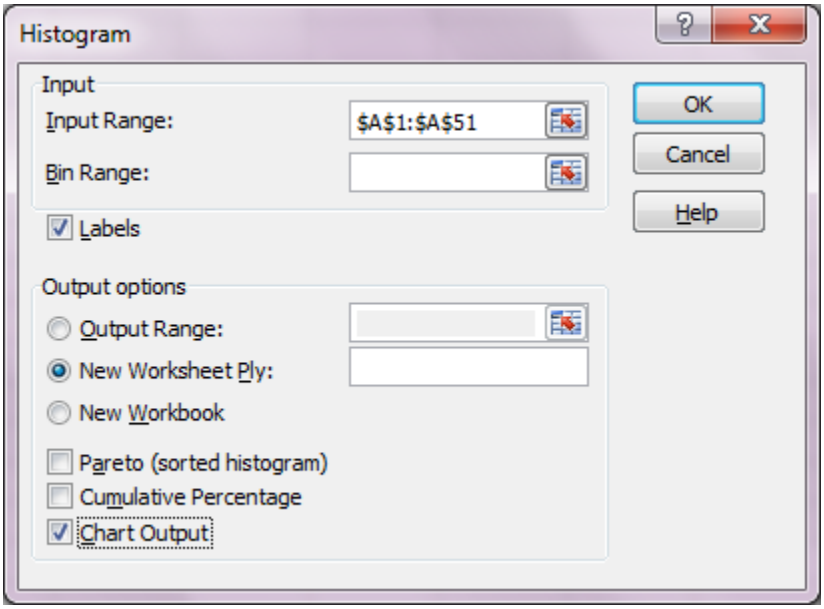
States differ widely in the percentage of college students who attend college in their home state. The percentages of freshmen who attend college in their home state for each of the 50 states are shown in the text (*The Chronicle of Higher Education*, August 23, 2013).

Begin by entering the data into a column, title Percentage

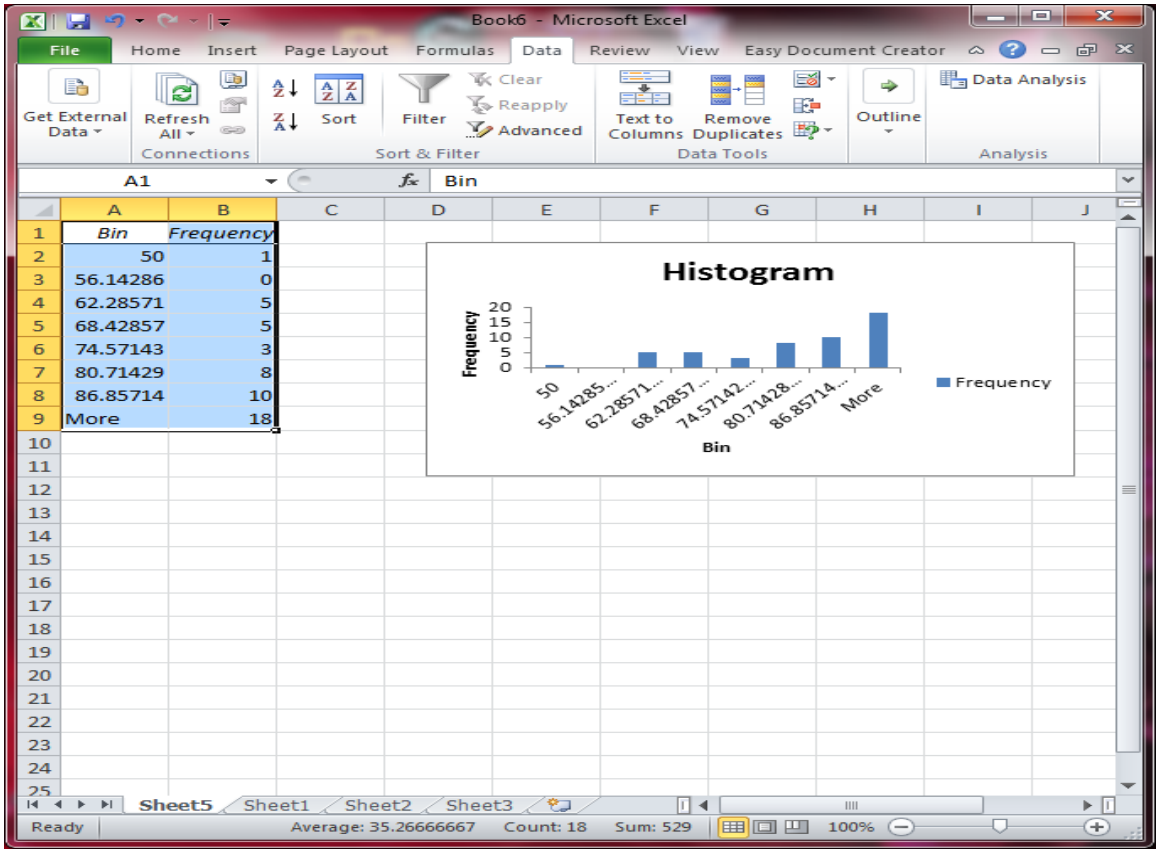


	A
1	Percentage
2	93
3	92
4	91
5	91
6	90
7	90
8	90
9	90
10	89
11	89
12	89
13	89
14	89
15	89
16	89
17	88
18	87
19	87
20	85
21	85
22	85
23	85
24	84
25	84

Click **Data>Data Analysis**. Select **Histogram.** Click **OK**. Select the data in column A including the Label for Input Range. Check the box next to Labels. Check the box next to Chart Output. Click **OK.**



The results follow.

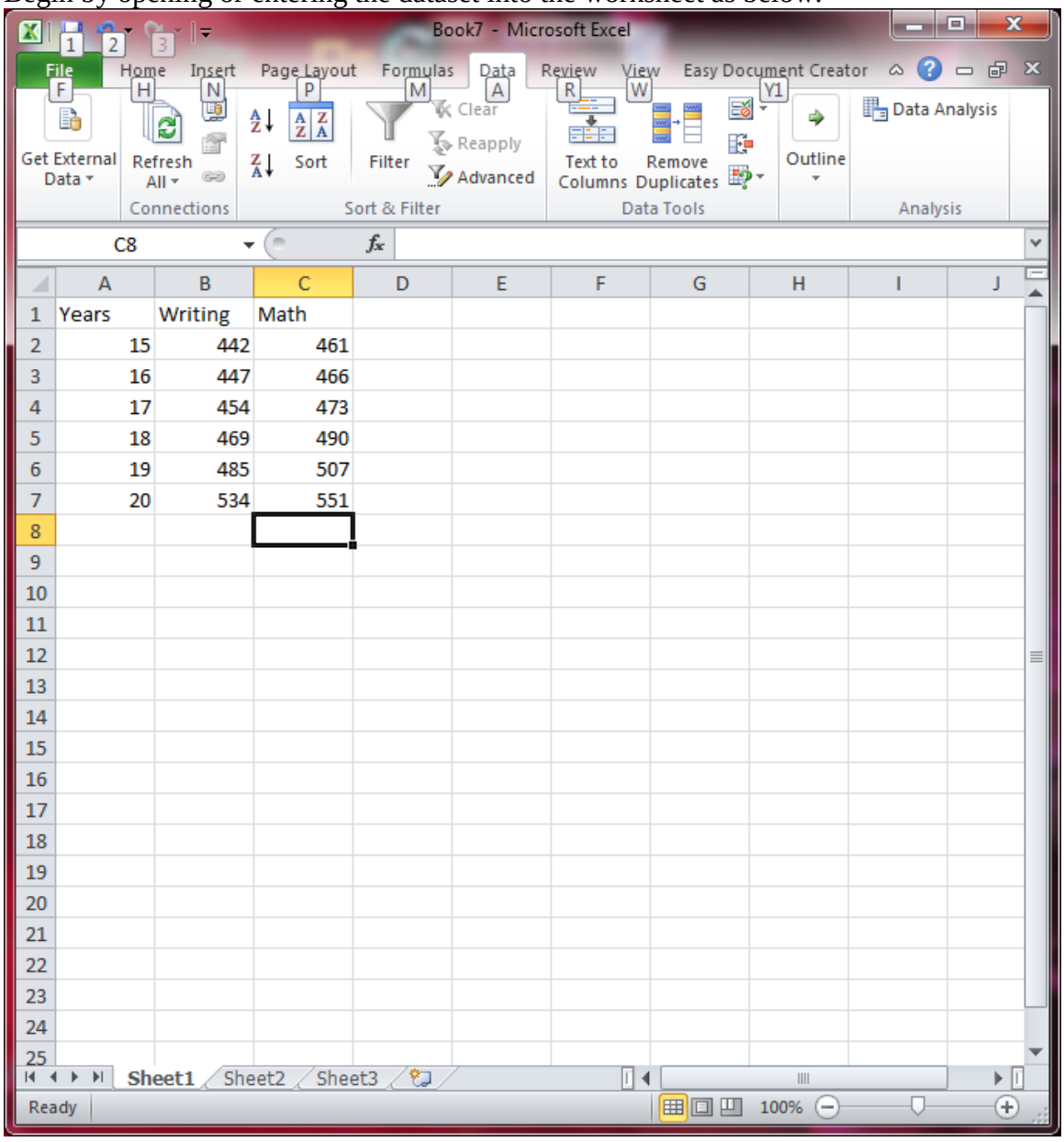


Bin	Frequency
50	1
56.14286	0
62.28571	5
68.42857	5
74.57143	3
80.71429	8
86.85714	10
More	18

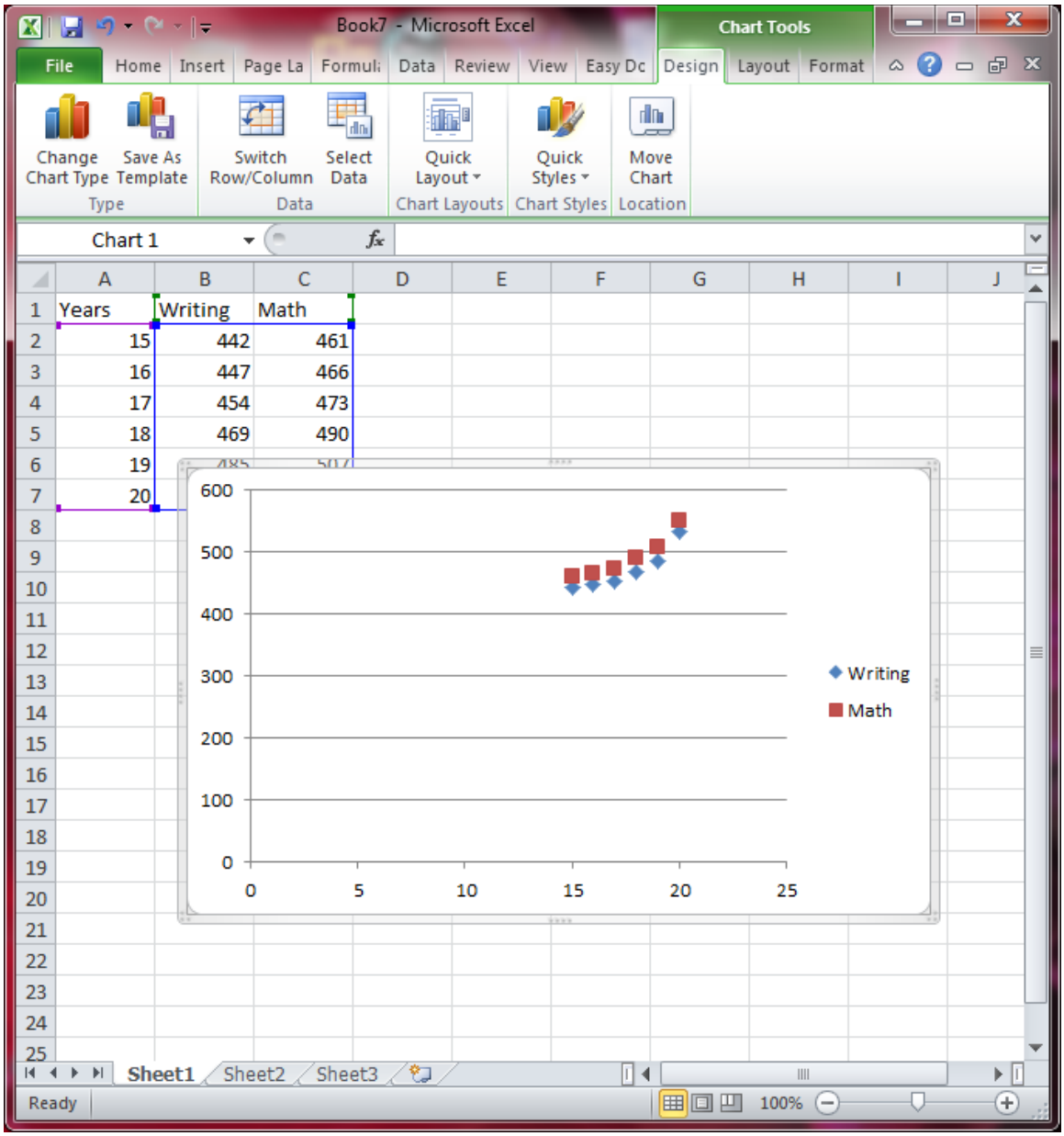
Example 3.21 – Scatterplots

The report title "2007 College Bound Seniors" (College Board, 2007) included data showing the average score on the writing and math sections of the SAT for groups of high school seniors completing different numbers of years of study in six core academic subjects (arts and music, English, foreign languages, mathematics, natural sciences, and social sciences and history.

Begin by opening or entering the dataset into the worksheet as below.



We will create a scatterplot by selecting **Insert>Charts>Scatter**. Select the first option. Excel will automatically select the data set input and display the scatterplot below.


Book7 - Microsoft Excel
Chart Tools
File
Home
Insert
Page La
Formuli
Data
Review
View
Easy Dc
Design
Layout
Format
Change Chart Type
Save As Template
Type
Switch Row/Column
Select Data
Data
Quick Layout
Chart Layouts
Quick Styles
Chart Styles
Move Chart
Location
Chart 1
Years
Writing
Math
15 442 461
16 447 466
17 454 473
18 469 490
19
20
600
500
400
300
200
100
0
0 5 10 15 20 25
Writing
Math
Sheet1
Sheet2
Sheet3
Ready
100%

Chapter 4
Numerical Methods for Describing Data

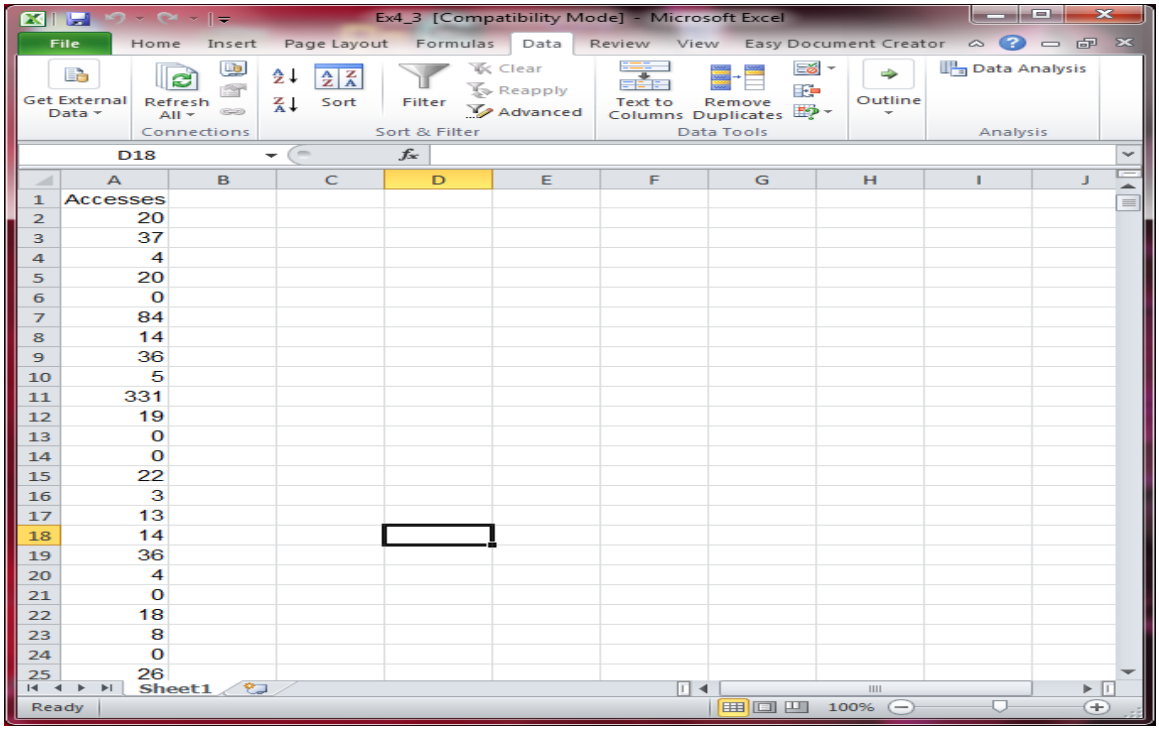
In previous chapters, we examined graphical methods for displaying data. These provide a visual picture of the data, but do not provide any numerical summaries. In this chapter, we compute numerical summaries for quantitative data. We use the menu command

Data>Data Analysis>Descriptive Statistics

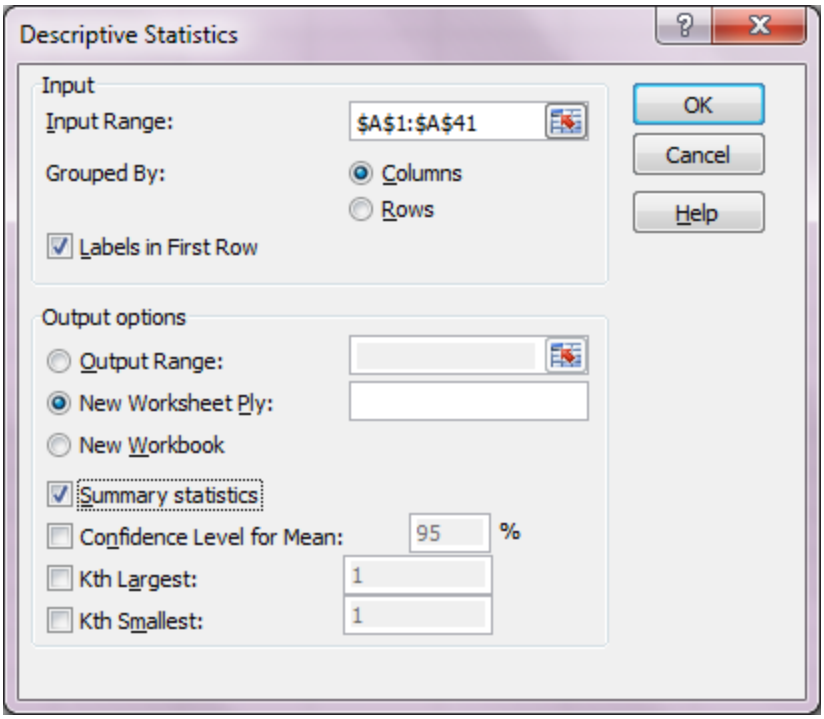
Example 4.3 – Calculating the Mean

Forty students were enrolled in a section of a general education course in statistical reasoning during one fall quarter at Cal Poly, San Luis Obispo. The instructor made course materials, grades and lecture notes available to students on a class web site, and course management software kept track of how often each student accessed any of the web pages on the class site. One month after the course began, the instructor requested a report that indicated how many times each student had accessed a web page on the class site. The 40 observations are listed in the text.

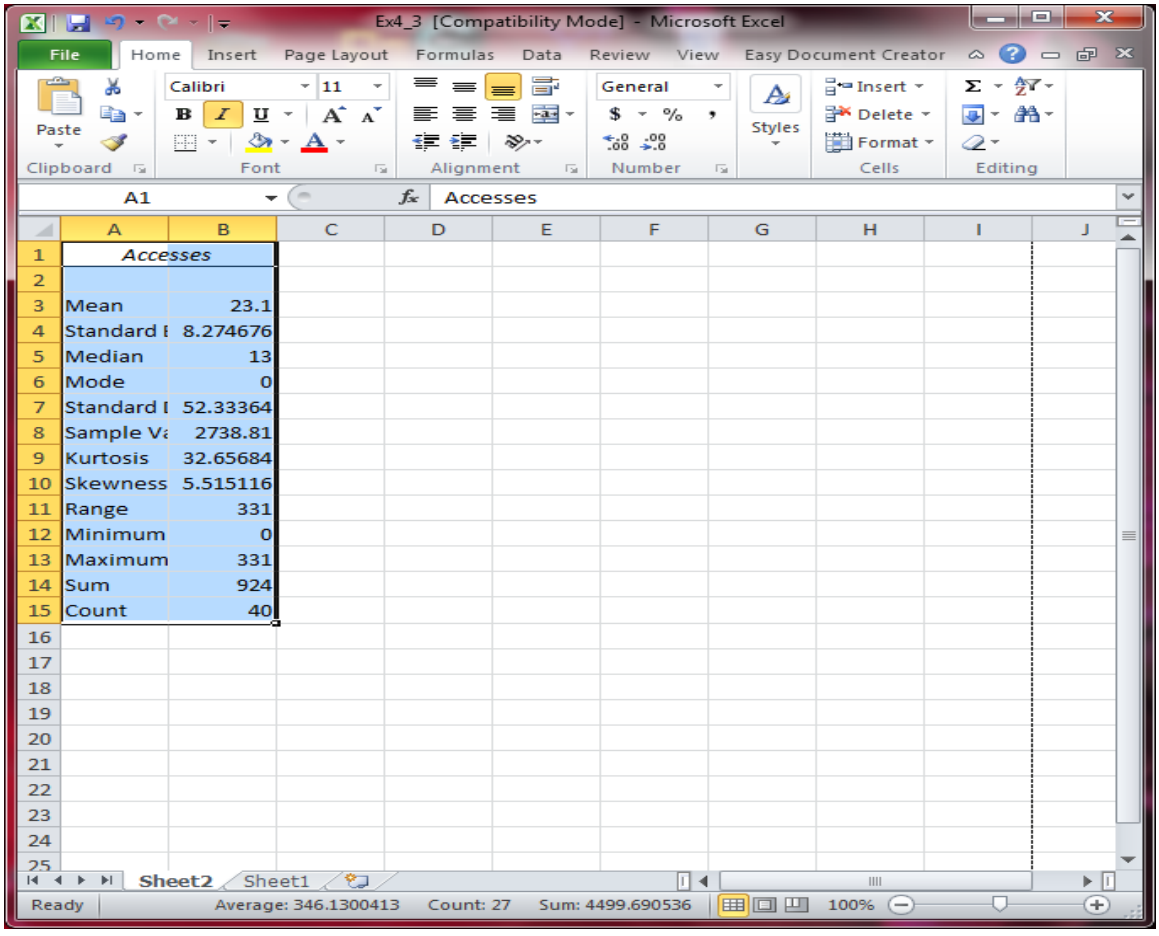
We begin by entering or opening the dataset in Excel.



Click **Data>Data Analysis>Descriptive Statistics.** Click **OK**. Select the data for Input Range, including the label. Check the button next to Labels in First Row. Check the box next to Summary statistics. Click **OK**.



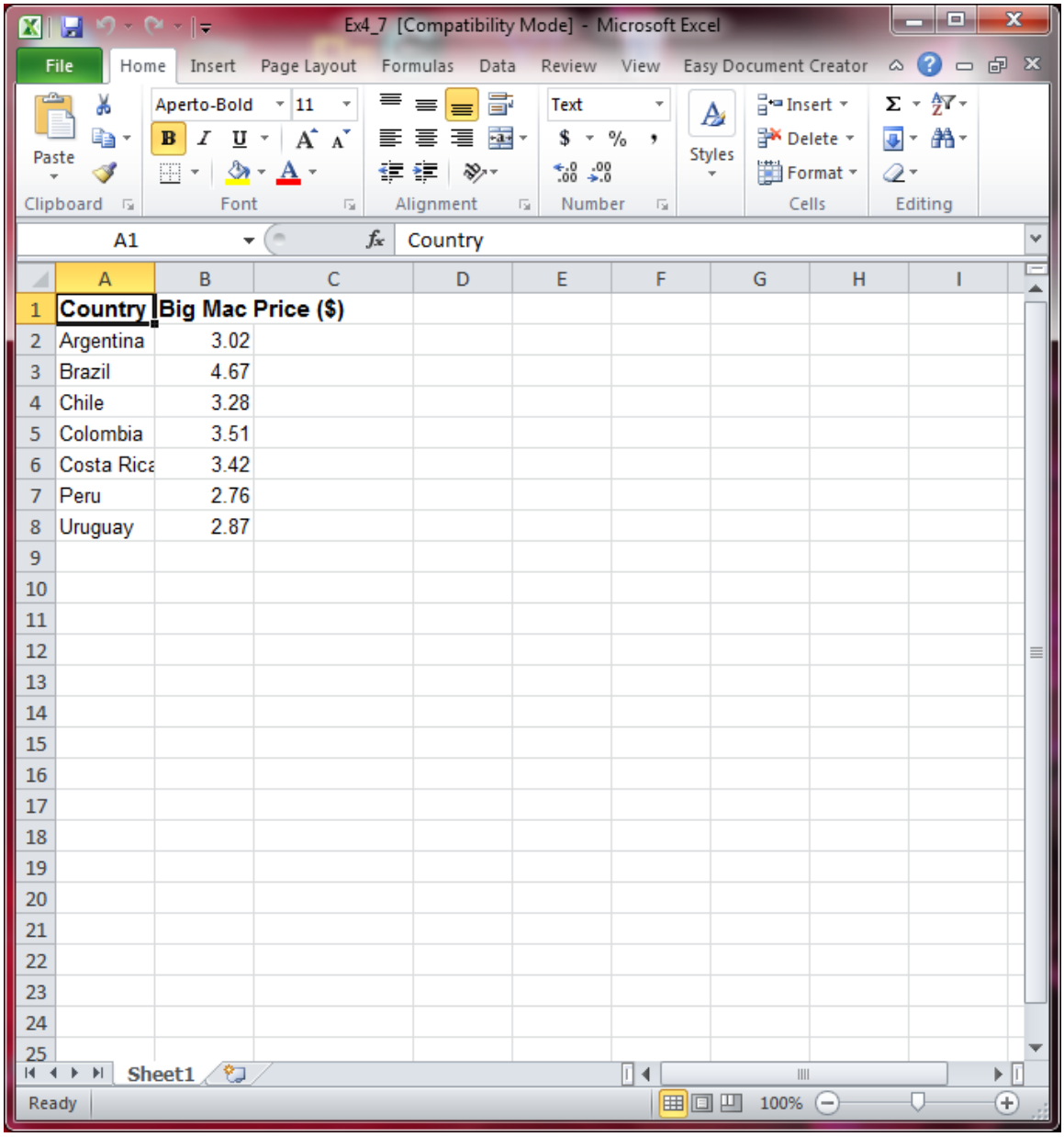
The results are shown below.



Example 4.8 – Calculating the Standard Deviation

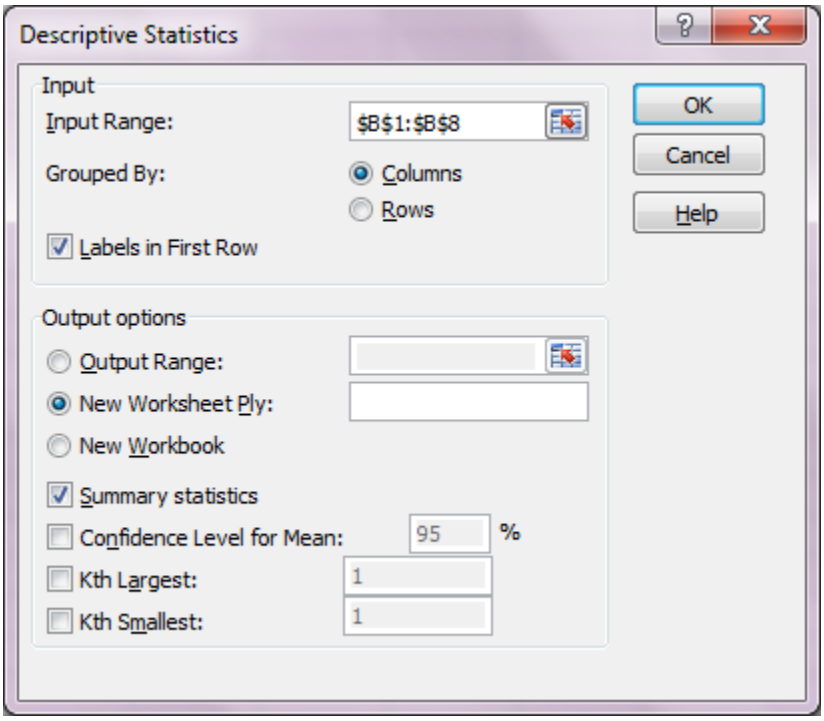
McDonald's fast-food restaurants are now found in many countries around the world. But the cost of a Big Mac varies from country to country. Table 4.3 in the text shows data on the cost of a Big Mac taken from the article "The Big Mac Index" (*The Economist*, July 11, 2013).

We begin by entering this data or opening the dataset in Excel.

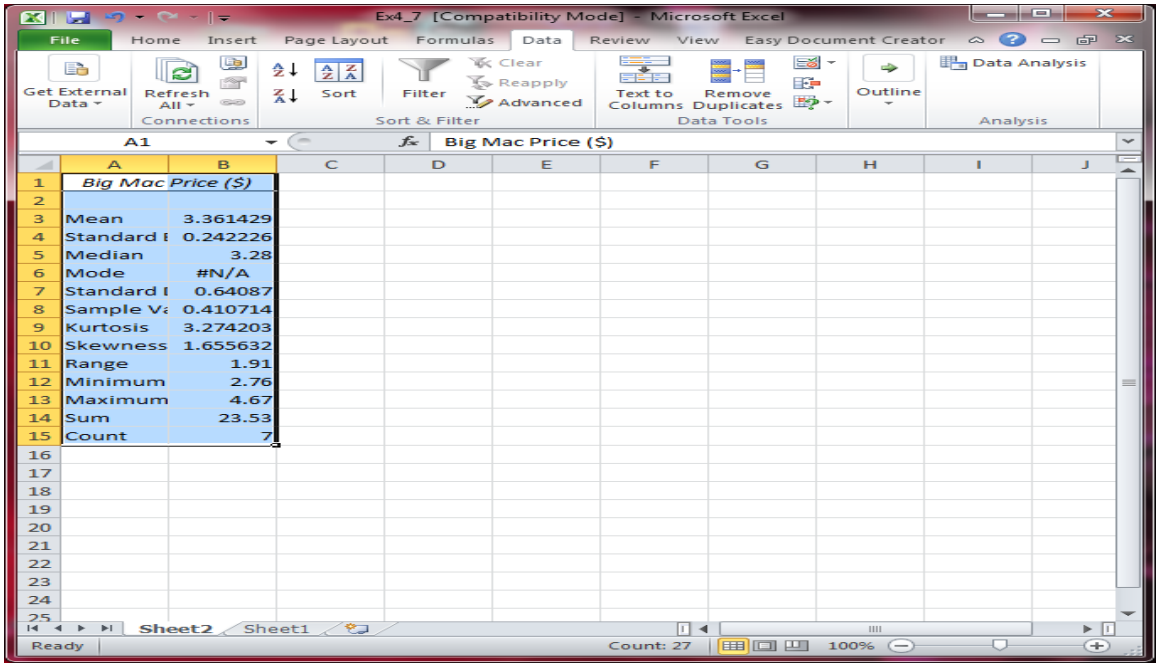


	A	B
1	Country	Big Mac Price ($)
2	Argentina	3.02
3	Brazil	4.67
4	Chile	3.28
5	Colombia	3.51
6	Costa Rica	3.42
7	Peru	2.76
8	Uruguay	2.87

To calculate the standard deviation, click **Data>Data Analysis>Descriptive Statistics.** Select the Big Mac Price data for Input Range. Check the box next to Labels in First Row and check the box next to Summary Statistics. Click **OK**.



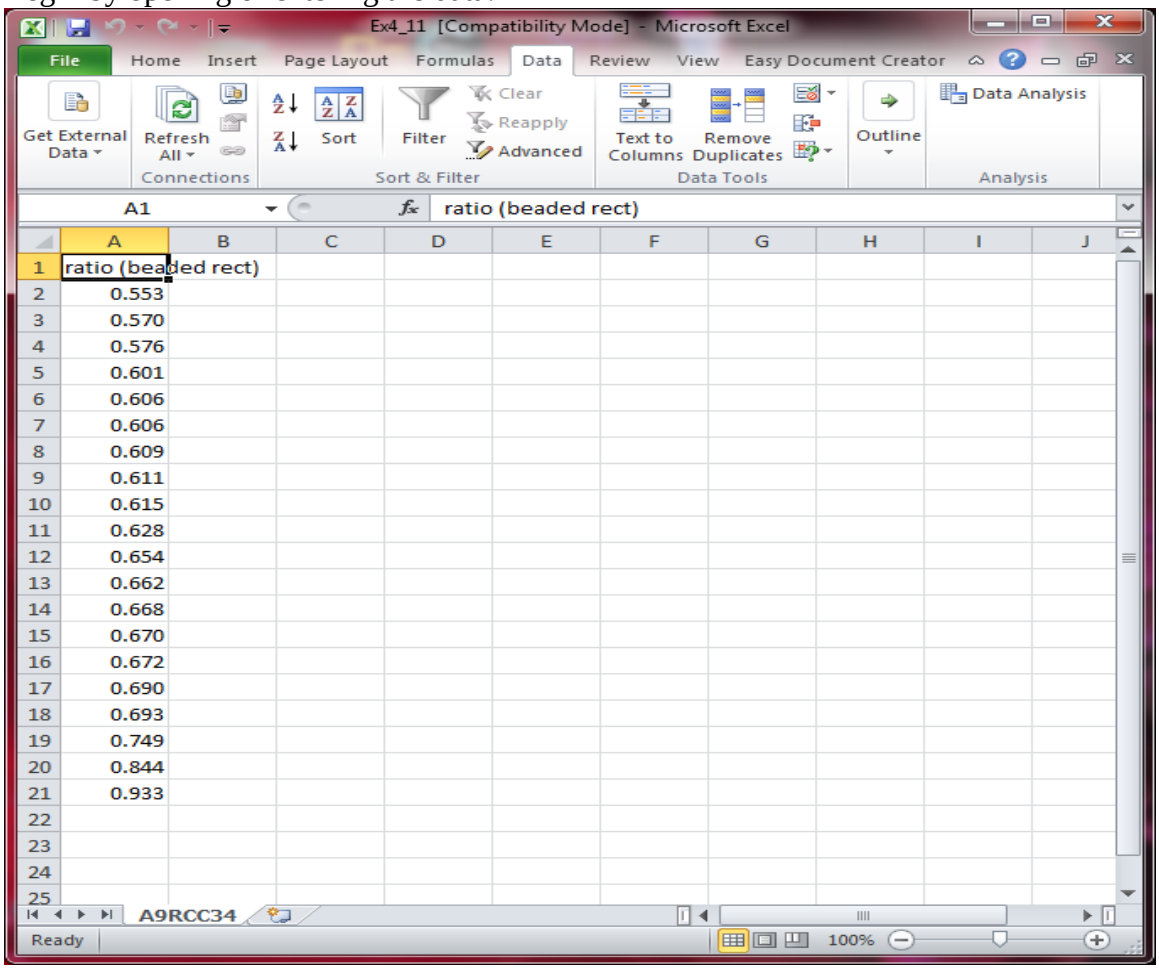
The results are displayed below.



Example 4.11 – Quartiles and IQR

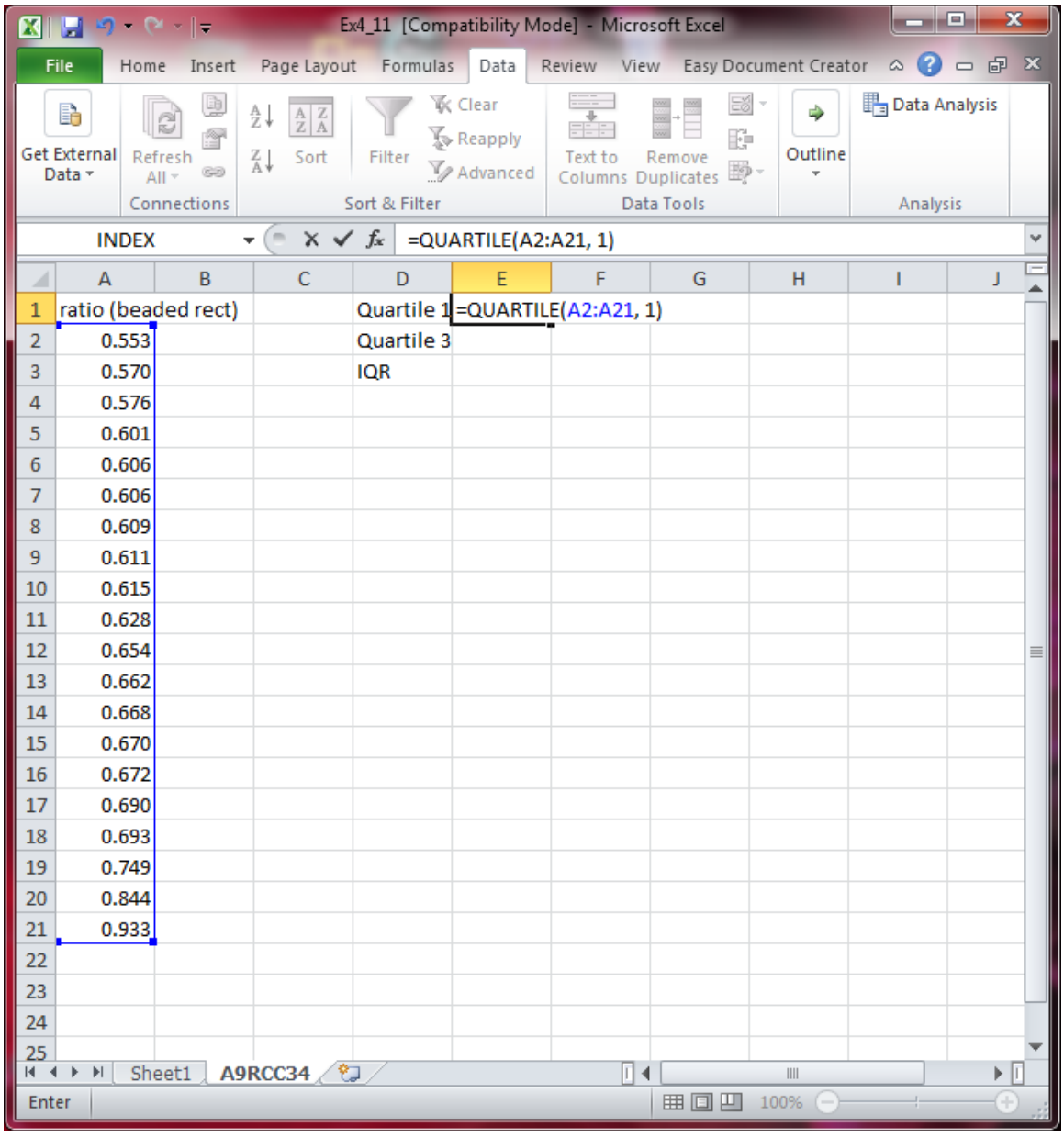
The accompanying data from Example 4.11 of the text came from an anthropological study of rectangular shapes (*Lowie's Selected Papers in Anthropology*, Cora Dubios, ed. [Berkeley, CA: University of California Press, 1960]: 137-142). Observations were made on the variable x = width/length for a sample of n = 20 beaded rectangles used in Shoshani Indian leather handicrafts.

Begin by opening or entering the data.

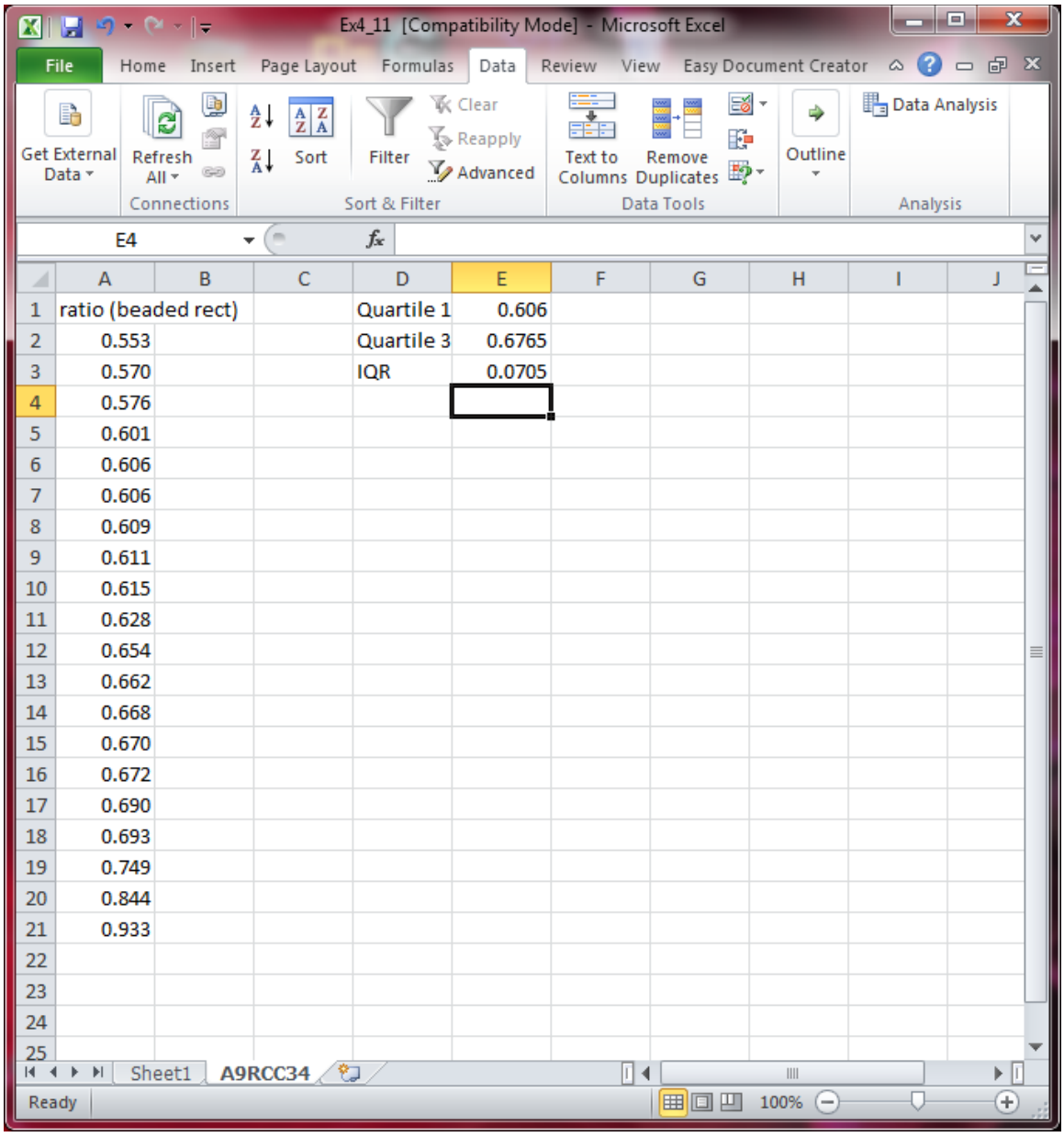


	A	B	C	D	E	F	G	H	I	J
1	ratio (beaded rect)									
2	0.553									
3	0.570									
4	0.576									
5	0.601									
6	0.606									
7	0.606									
8	0.609									
9	0.611									
10	0.615									
11	0.628									
12	0.654									
13	0.662									
14	0.668									
15	0.670									
16	0.672									
17	0.690									
18	0.693									
19	0.749									
20	0.844									
21	0.933									
22										
23										
24										
25										

To find the quartiles and IQR, we use the QUARTILE function in Excel. Click in cell D1 and input Quartile 1. Input Quartile 3 into D2 and IQR into D3. In cell E1, type =QUARTILE(A2:A21, 1). Press **Enter**.



In cell E2, input =Quartile(A2:A21, 3). In cell E3, input =E2-E1. The results follow.


Ex4_11 [Compatibility Mode] - Microsoft Excel
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Easy Document Creator
Get External Data
Refresh All
Connections
Sort
Filter
Clear
Reapply
Advanced
Sort & Filter
Text to Columns
Remove Duplicates
Data Tools
Outline
Data Analysis
Analysis
E4
ratio (beaded rect)
0.553
0.570
0.576
0.601
0.606
0.606
0.609
0.611
0.615
0.628
0.654
0.662
0.668
0.670
0.672
0.690
0.693
0.749
0.844
0.933
Quartile 1
0.606
Quartile 3
0.6765
IQR
0.0705
Sheet1
A9RCC34
Ready
100%

Chapter 5
Summarizing Bivariate Data

This chapter introduces methods for describing relationships among variables. The goal of this type of analysis is often to predict a characteristic of the response or dependent variable. The characteristics used to predict the response variable are called the independent or predictor variables.

In this chapter, we will learn how to compute the correlation and least squares regression fit to bivariate data in Excel. We will be using the commands

Data>Data Analysis>Correlation

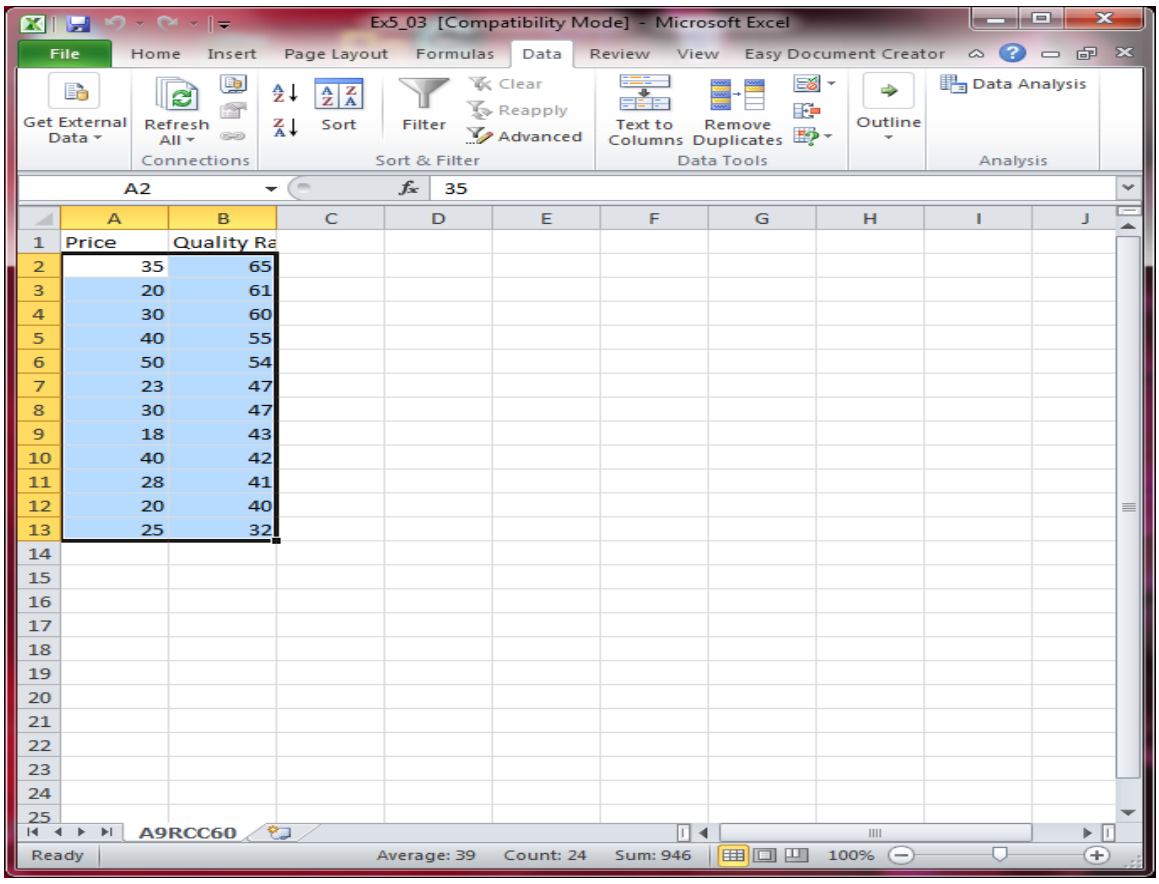
and

Data>Data Analysis>Regression.

Example 5.2 – Correlation

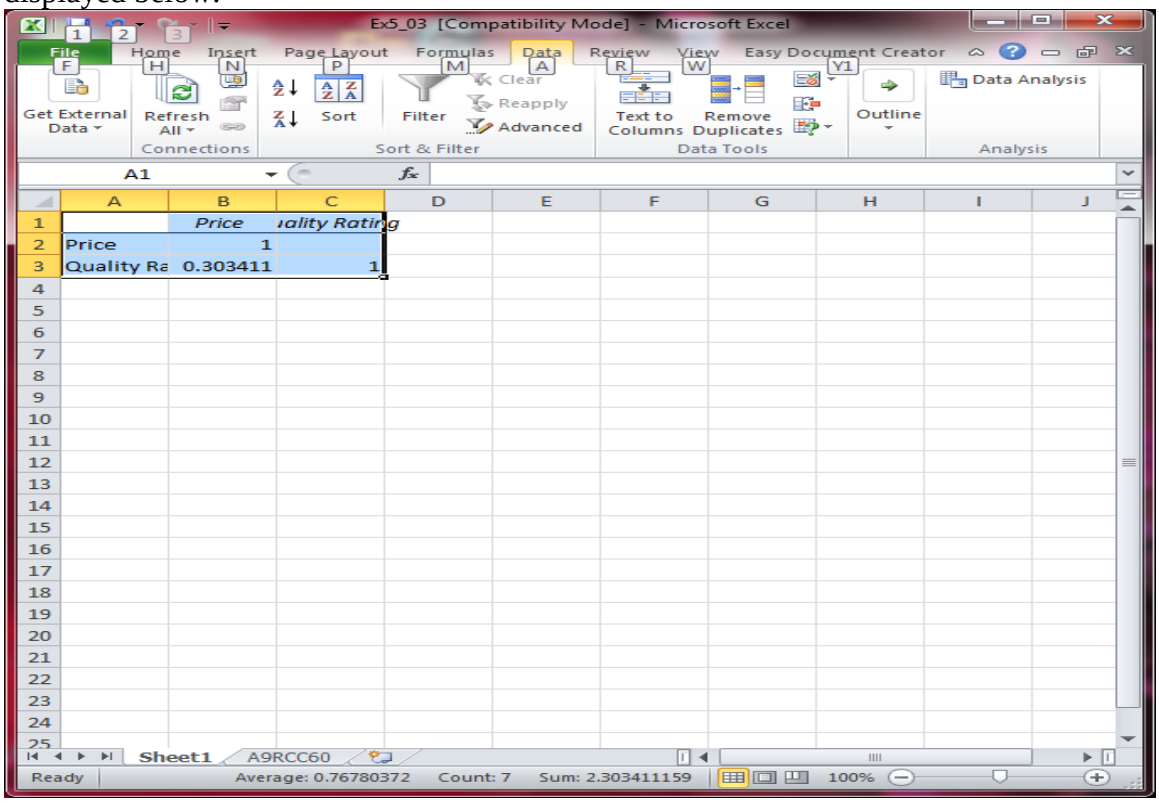
Are more expensive bike helmets safer than less expensive ones? The data from Example 5.2 of the text shows data on x = price and y = quality rating for 12 different brands of bike helmets appeared on the *Consumer Reports* web site.

Begin by opening or entering the data in Excel.



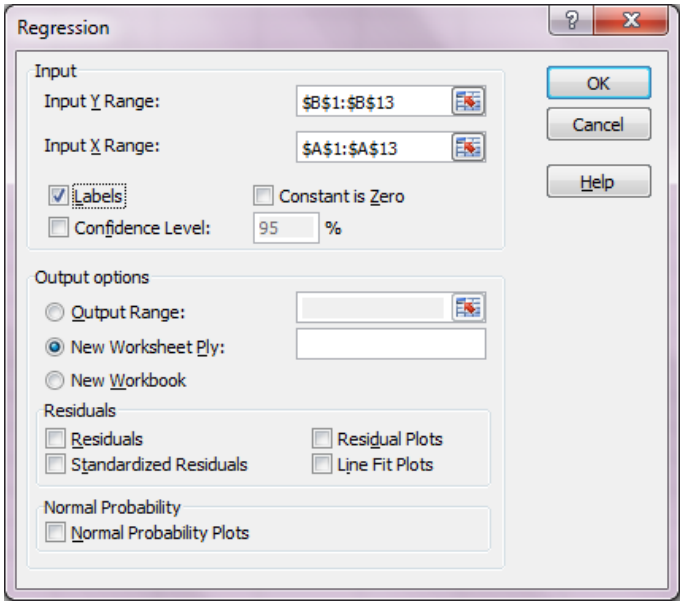
To calculate the correlation, select **Data>Data Analysis>Correlation**. Select the data as Input Range and check the box next to Labels in First Row. Click **OK**. The results are

displayed below.

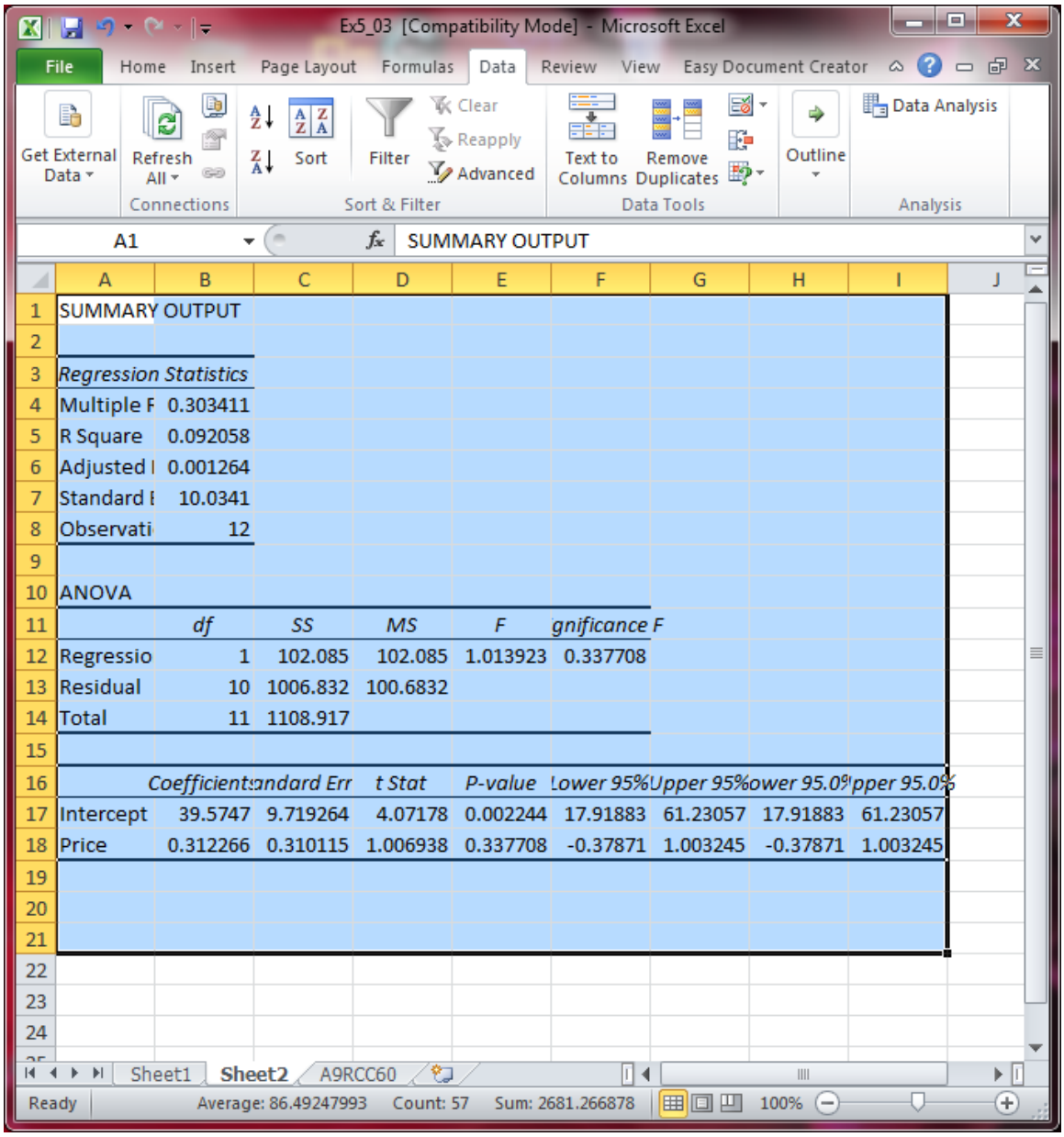


Example 5.2 – Regression Equation

For the example above, we will now fit the regression equation. Choose **Data>Data Analysis>Regression**. Input Quality Rating for Input Y Range and Price for Input X Range. Check the box next to Labels. Click **OK**.



The results follow.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple F	0.303411							
5	R Square	0.092058							
6	Adjusted I	0.001264							
7	Standard E	10.0341							
8	Observati	12							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regressio	1	102.085	102.085	1.013923	0.337708			
13	Residual	10	1006.832	100.6832					
14	Total	11	1108.917						
15									
16		*Coefficients*	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	39.5747	9.719264	4.07178	0.002244	17.91883	61.23057	17.91883	61.23057
18	Price	0.312266	0.310115	1.006938	0.337708	-0.37871	1.003245	-0.37871	1.003245
19									
20									
21									
22									
23									
24									

Notice that this also outputs the ANOVA table for assessing the fit of the line.

Chapter 6
Probability

There are no examples or material from Chapter 6.

Chapter 7
Random Variables and Probability Distributions

Chapter 7 explores further the concepts of probability, expanding to commonly used probability distributions. These distributions will be closely liked to the processes that we use for inference.

In this chapter, we learn how to use Excel to compute probabilities from the Normal distribution. We will use the formula

NORM.DIST

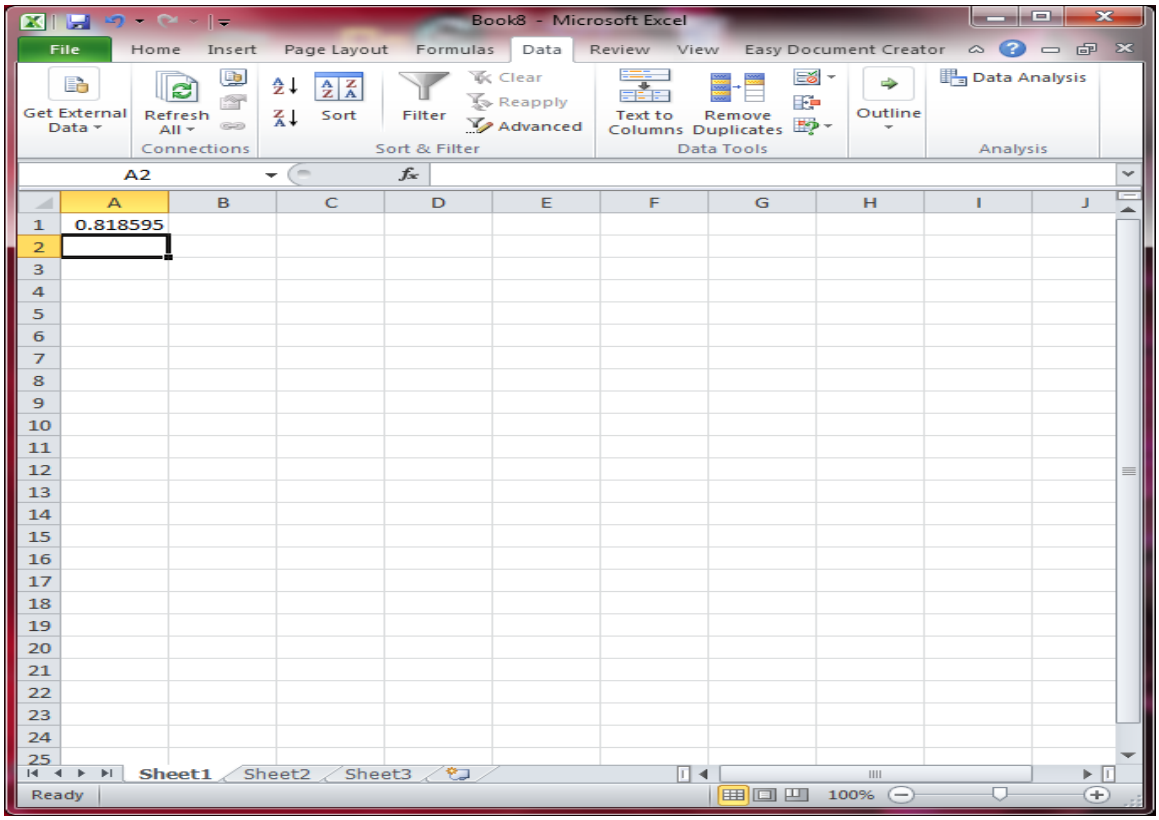
to compute probabilities.

Example 7.27 – Calculating Normal Probabilities

Data from the paper "Fetal Growth Parameters and Birth Weight: Their Relationship to Neonatal Body Composition" (*Ultrasound in Obstetrics and Gynecology* [2009]: 441-446) suggest that a normal distribution with mean $\mu = 3500$ grams and standard deviation $\sigma = 600$ grams is a reasonable model for the probability distribution of the continuous numerical variable x = birth weight of a randomly selected full-term baby. What proportion of birth weights are between 2900 and 4700 grams?

We do not need to enter any data in Excel to complete this problem. We begin with a blank worksheet. To find this probability, we must find the probability of being less than 4700 and 2900 separately and subtract the results.

To start, click in cell A1 and type =NORM.DIST(4700, 3500, 600, TRUE)-NORM.DIST(2900, 3500, 600, TRUE). Press **Enter**. The results are displayed in A1.


Book8 - Microsoft Excel
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Easy Document Creator
Get External Data
Refresh All
Connections
Sort
Filter
Clear
Reapply
Advanced
Sort & Filter
Text to Columns
Remove Duplicates
Data Tools
Outline
Data Analysis
Analysis
A2
0.818595
Sheet1
Sheet2
Sheet3
Ready
100%

Chapter 8
Sampling Variability and Sampling Distributions

There are no examples or material from Chapter 8.

Chapter 9
Estimation using a Single Sample

The objective of inferential statistics is to use sample data to decrease our uncertainty about the corresponding population. We can use confidence intervals to provide estimation for population parameters.

In this chapter, we will use Excel to create confidence intervals for the population mean. We will use the menu commands

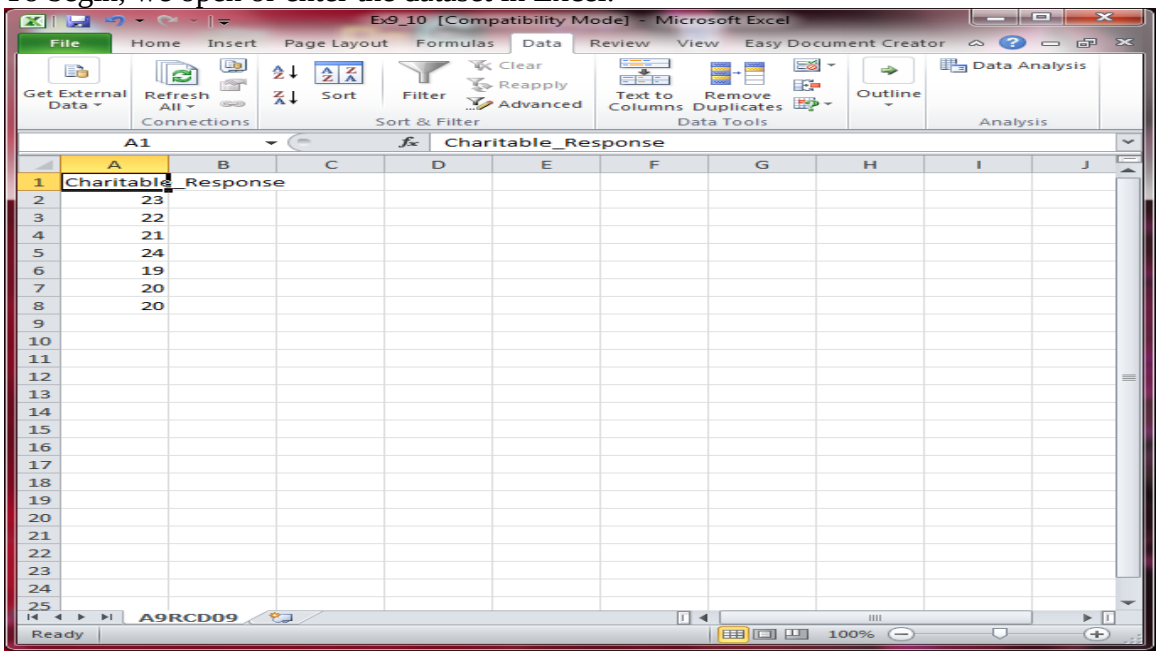
Data>Data Analysis

to find a confidence interval for the population mean.

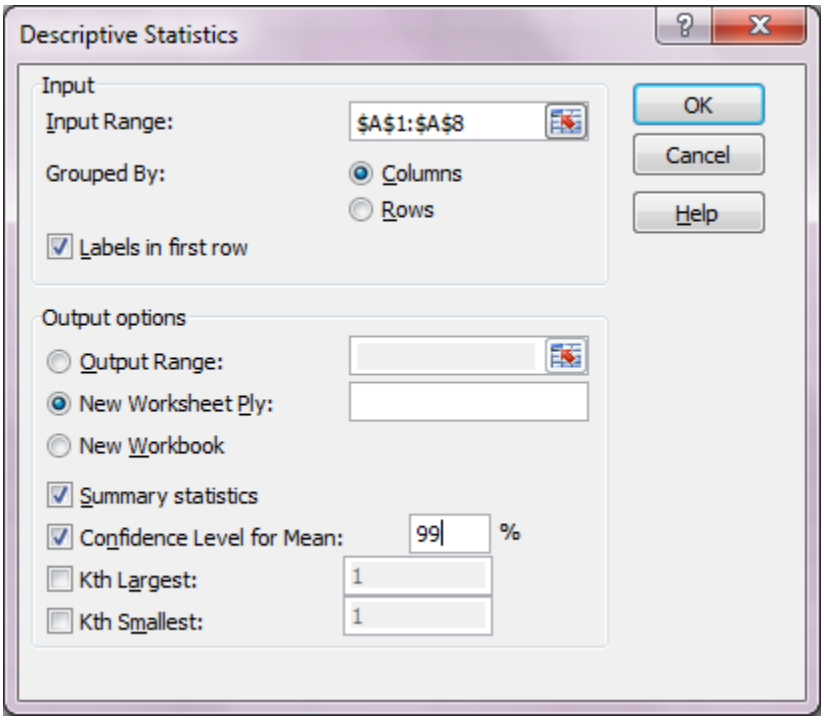
Example 9.10 – Confidence Interval for the Mean

The article "Chimps Aren't Charitable" (*Newsday*, November 2, 2005) summarizes the results of a research study published in the journal *Nature*. In this study, chimpanzees learned to use an apparatus that dispensed food when either of the two ropes was pulled. When one of the ropes was pulled, only the chimp controlling the apparatus received food. When the other rope was pulled, food was dispensed both to the chimp controlling the apparatus and also to a chimp in the adjoining cage. The data (listed in the text) represents the number of times out of 36 trials that each of the seven chimps chose the option that would provide food to both chimps (the "charitable" response).

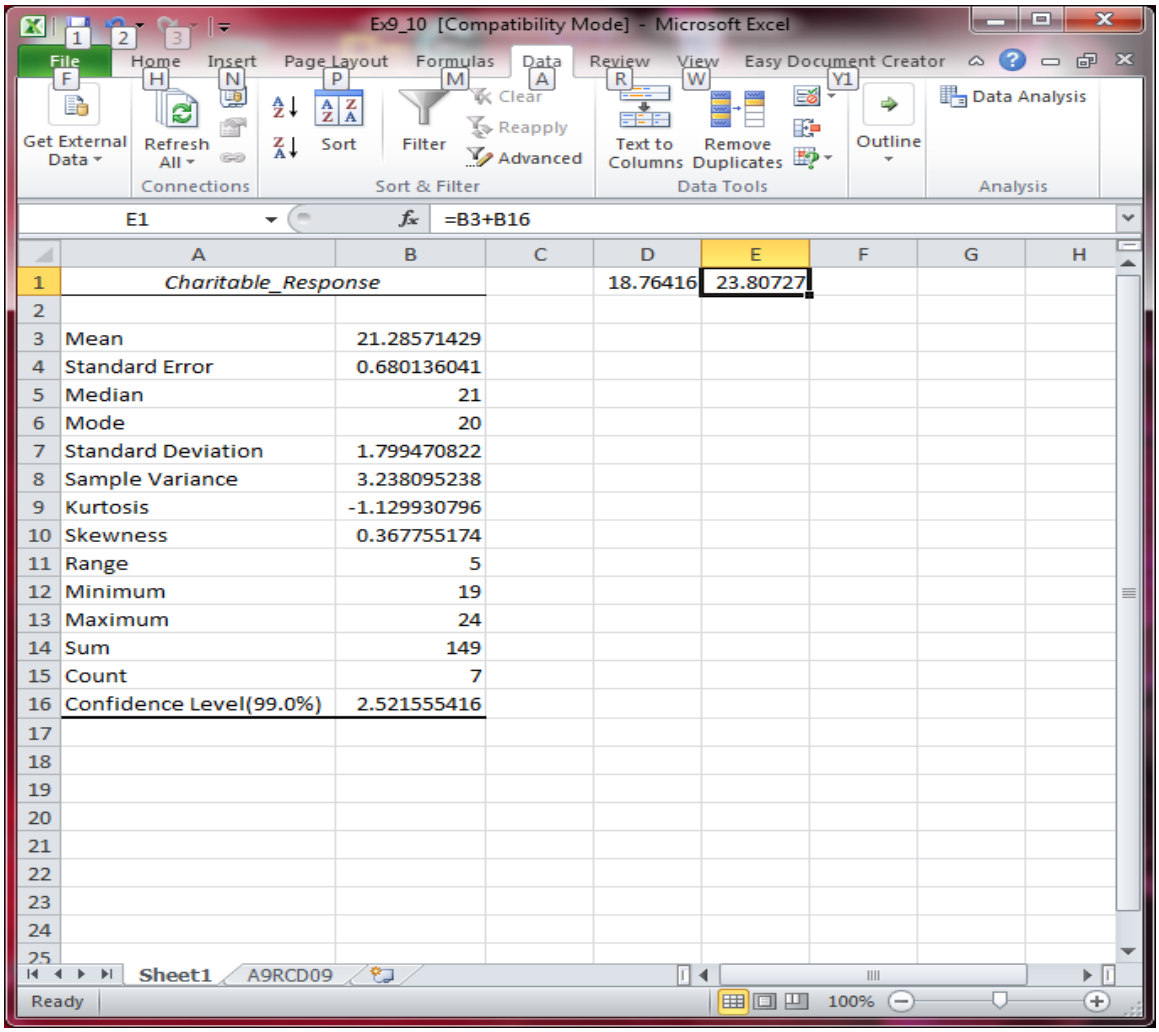
To begin, we open or enter the dataset in Excel.



To compute the confidence interval for the mean, we choose **Data>Data Analysis>Descriptive Statistics.** Input the data into Input Range and check the box for Labels in First Row and Summary Statistics. Check the box next to Confidence Level for Mean. Input 99 into the box for level. Click **OK.**



To find the confidence interval, we use two pieces of this display. Click in cell D1 and input =B3-B16, this creates the lower bound for the interval. Click in cell E1 and input =B3+B16, this creates the upper bound for the interval.. Results are shown below.


Ex9_10 [Compatibility Mode] - Microsoft Excel
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Easy Document Creator
Get External Data
Refresh All
Connections
Sort
Filter
Clear
Reapply
Advanced
Sort & Filter
Text to Columns
Remove Duplicates
Data Tools
Outline
Data Analysis
Analysis
E1
=B3+B16
Charitable_Response
18.76416
23.80727
Mean 21.28571429
Standard Error 0.680136041
Median 21
Mode 20
Standard Deviation 1.799470822
Sample Variance 3.238095238
Kurtosis -1.129930796
Skewness 0.367755174
Range 5
Minimum 19
Maximum 24
Sum 149
Count 7
Confidence Level(99.0%) 2.521555416
Sheet1
A9RCD09
Ready
100%

Chapter 10
Hypothesis Testing using a Single Sample

In the previous chapter, we considered one form of inference in the form of confidence intervals. In this chapter, we consider the second form of inference: inference testing. Now we use sample data to create a test for a set of hypotheses about a population parameter.

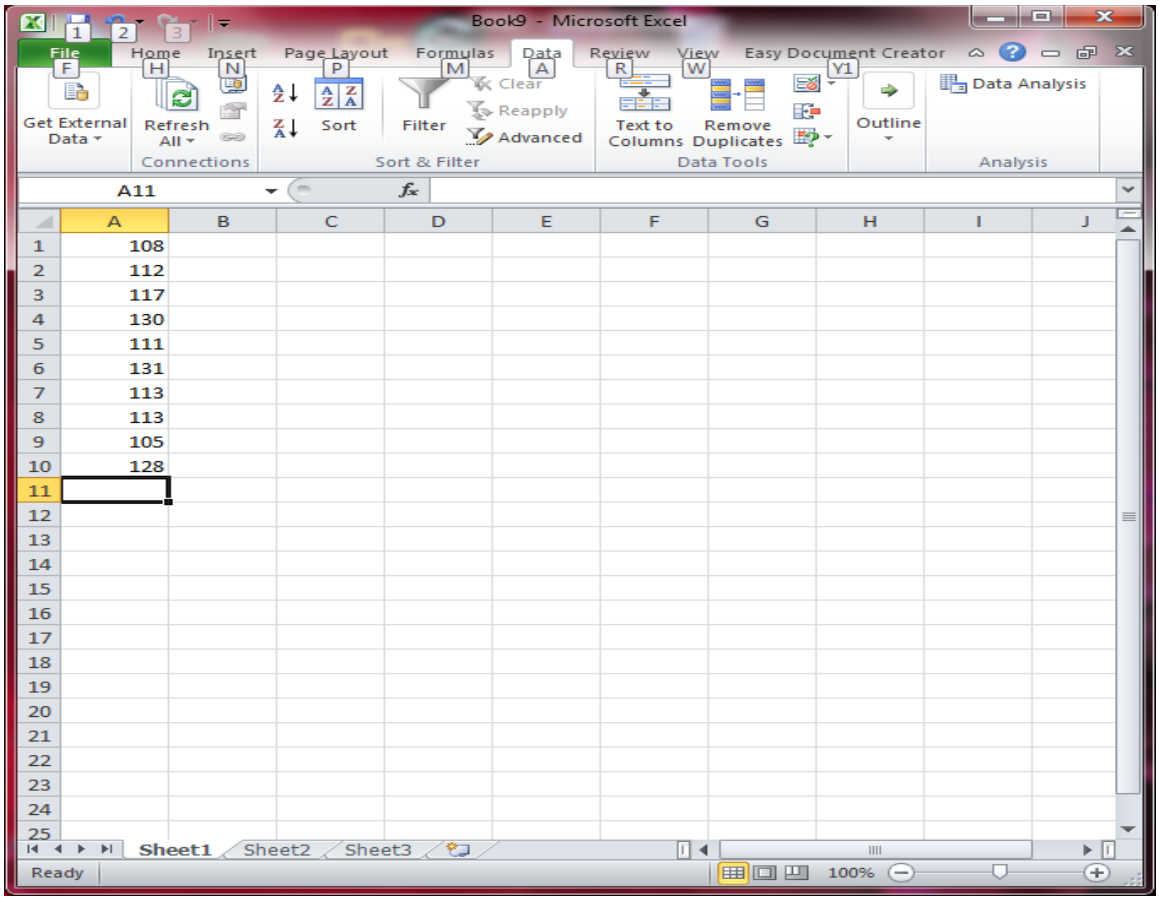
Excel does not support functionality to automatically produce tests for a single proportion or mean. Values for these tests must be computed by hand using descriptive statistics and NORM.DIST or T.DIST to compute p-values. We will illustrate this using a test for a single mean below.

Example 10.14 – Testing for a Single Mean

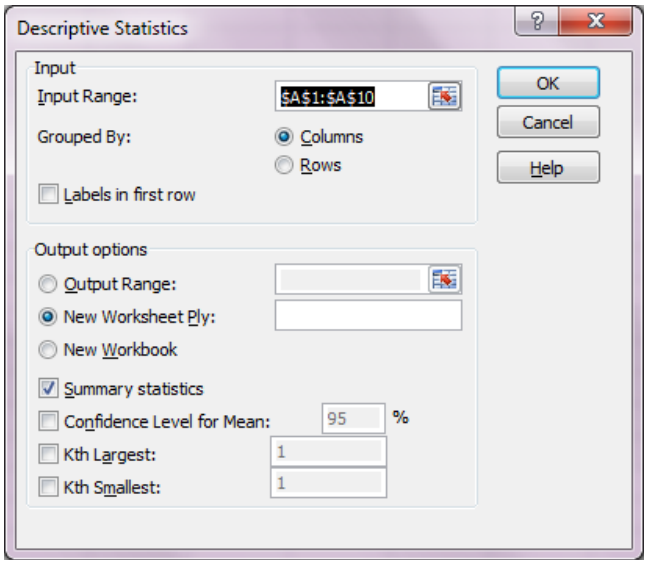
A growing concern of employers is time spent in activities like surfing the Internet and e-mailing friends during work hours. The *San Luis Obispo Tribune* summarized the findings from a survey of a large sample of workers in an article that ran under the headline "Who Goofs Off 2 Hours a Day? Most Workers, Survey Says" (August 3, 2006). Suppose that the CEO of a large company wants to determine whether the average amount of wasted time during an 8-hour work day for employees of her company is less than the reported 120 minutes. Each person in a random sample of 10 employees was contacted and asked about daily wasted time at work. (Participants would probably have to be guaranteed anonymity to obtain truthful responses!). The resulting data are the following:

108 112 117 130 111 131 113 113 105 128

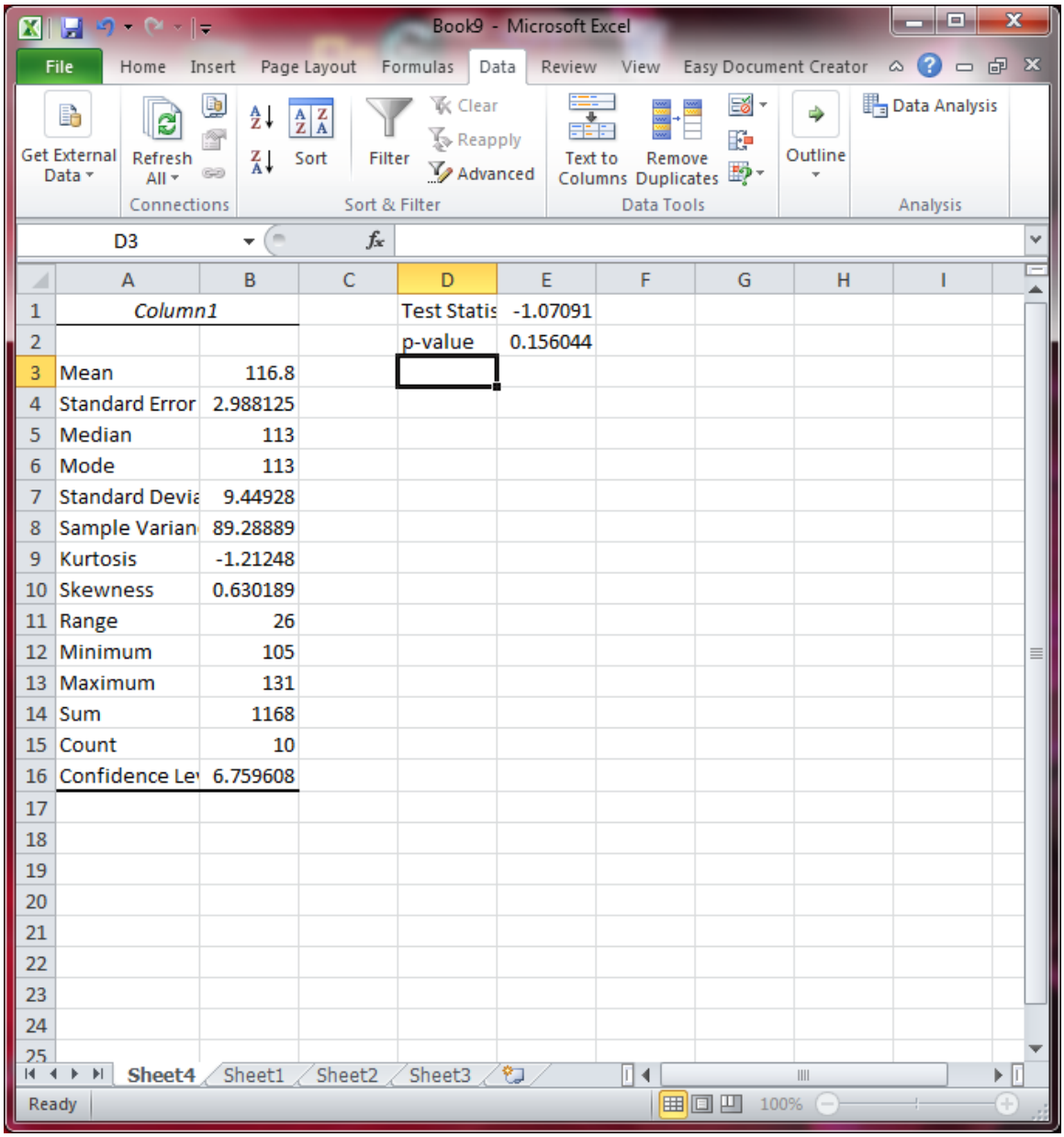
We begin by opening or entering the data in Excel.



To calculate the test statistic, begin by calculating the descriptive statistics as in Chapter 4.



To compute the test statistic, click in cell D1 and input Test Statistic. Click in cell E1 and input =(B3-120)/B4. To calculate the p-value, first input p-value in cell D2. Then in cell E1, input =T.DIST(E1, 9, TRUE). The results are shown below.



Chapter 11
Comparing Two Populations or Treatments

In many situations, we would like to compare two groups to determine if they behave in the same way. We may want to compare two groups to determine if they have the same mean value for a particular characteristic or to determine if they have the same success proportion for a particular characteristic.

We continue to use the inference procedures of interval estimation and inference testing for these situations. In this chapter, we perform both types of inference for comparing two groups' population means and proportions. We will use the commands

Data>Data Analysis>t-Test: Two-Sample Assuming Unequal Variances

to perform inference for two population means and

Data>Data Analysis>t-Test: Paired Two Sample for Means

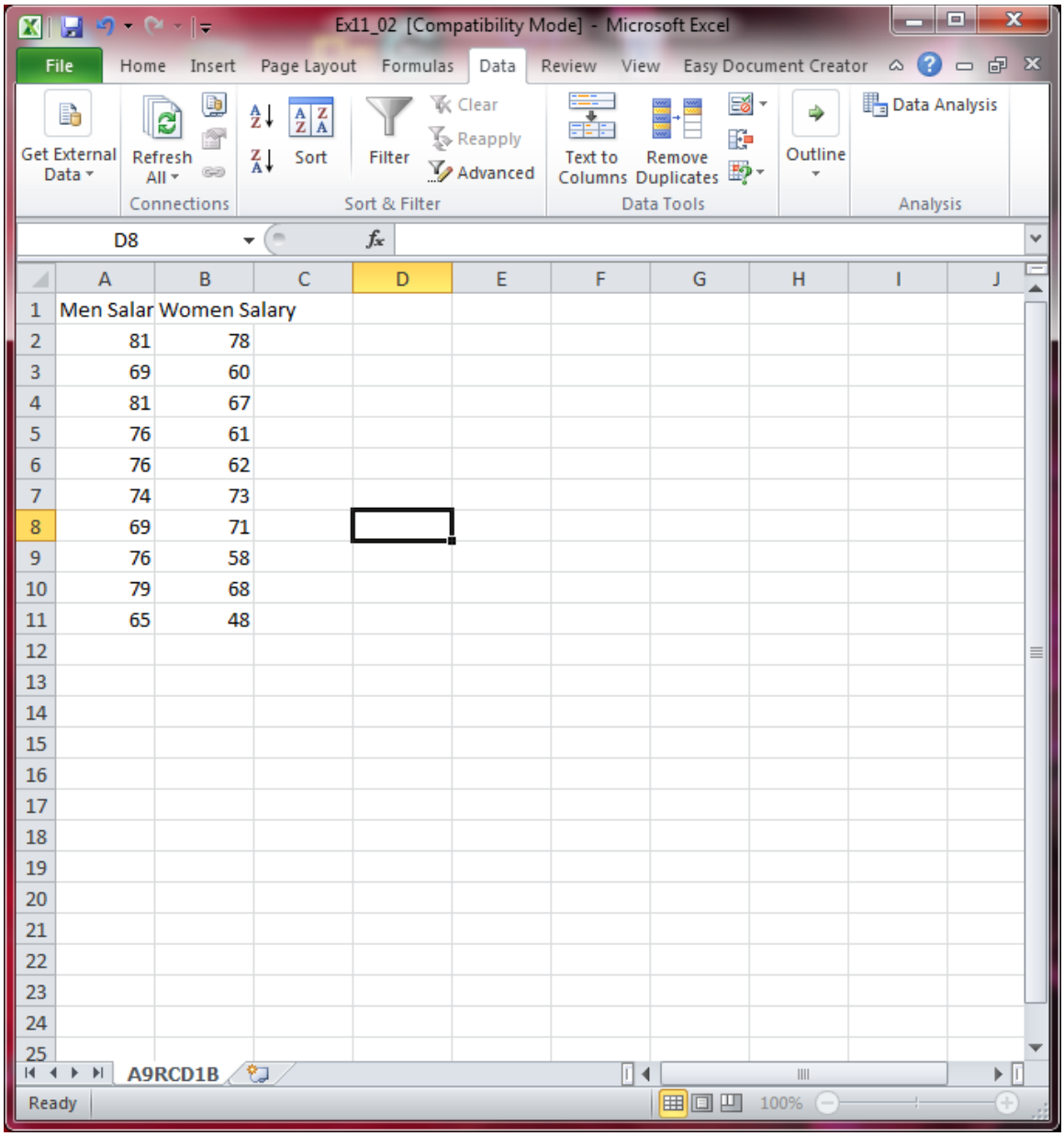
for inference for population means when the two groups are dependent.

Excel does not provide the functionality for a two-sample proportion test.

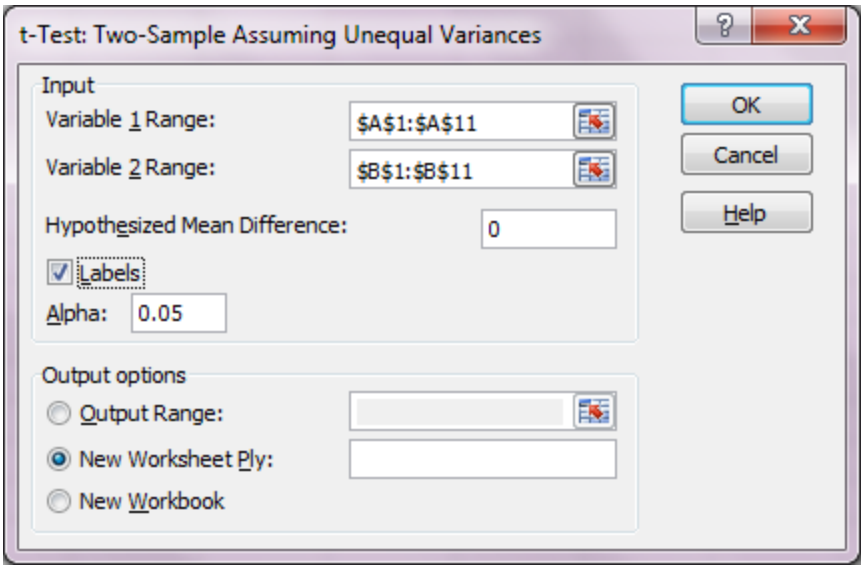
Example 11.2 – Confidence Interval and Inference Test for Two Population Means

Are women still paid less than men for comparable work? The authors of the paper "Sex and Salary: A Survey of Purchasing and Supply Professionals" (*Journals of Purchasing and Supply Management* [2008]: 112-124) carried out a study in which salary data was collected from a random sample of men and a random sample of women who worked as purchasing managers and who were subscribers to *Purchasing* magazine. Salary data consistent with summary quantities given in the paper appeal below (the actual sample size for the study were much larger).

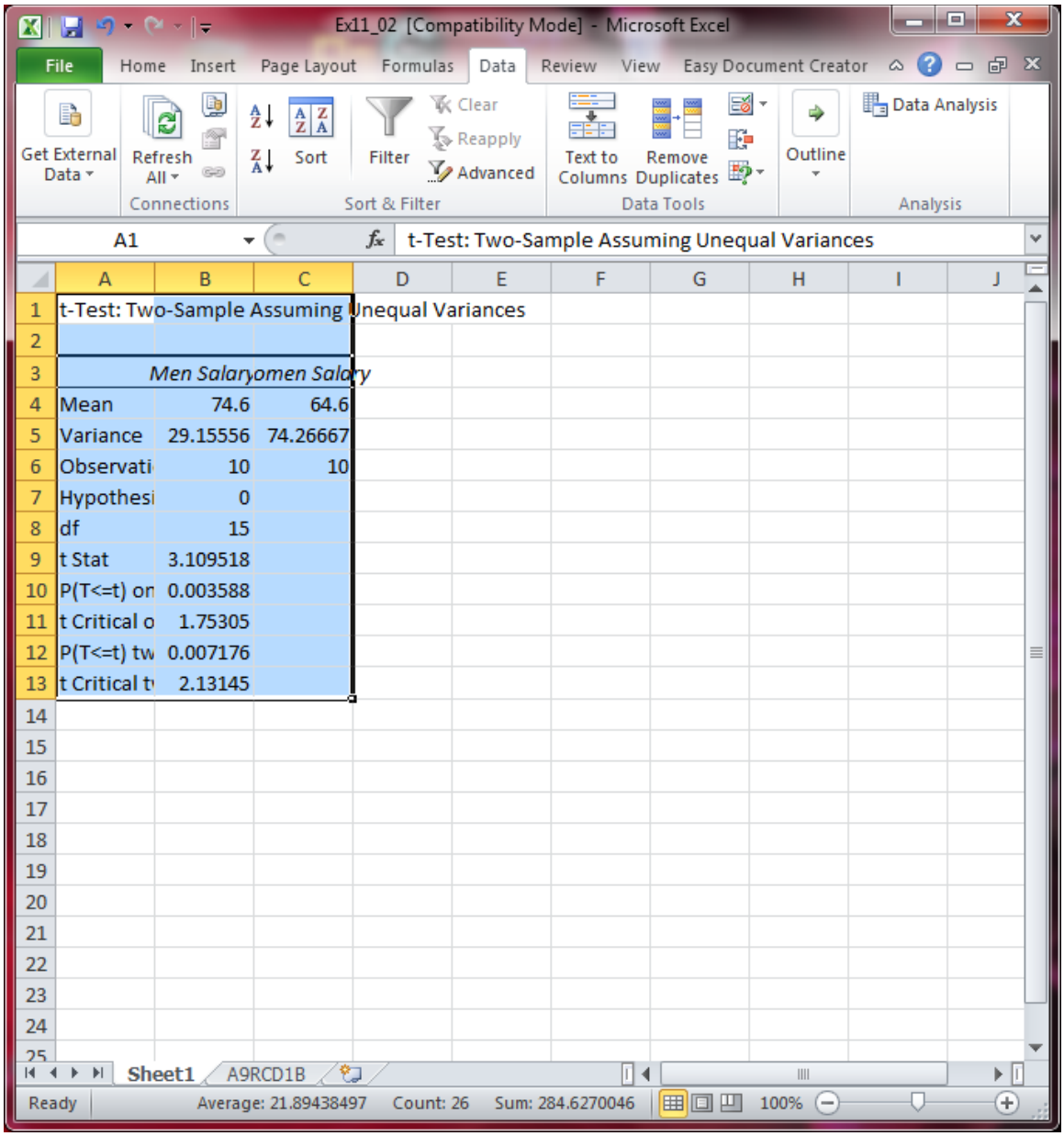
Begin by opening or entering the data in Excel.



Click **Data>Data Analysis>t-Test: Two Sample Assuming Unequal Variances**. Input Men's salary for Variable 1 Range and Women's salaries for Variable 2 Range. Check the box next to Labels. Input 0 into Hypothesized Mean Difference box. Click **OK.**



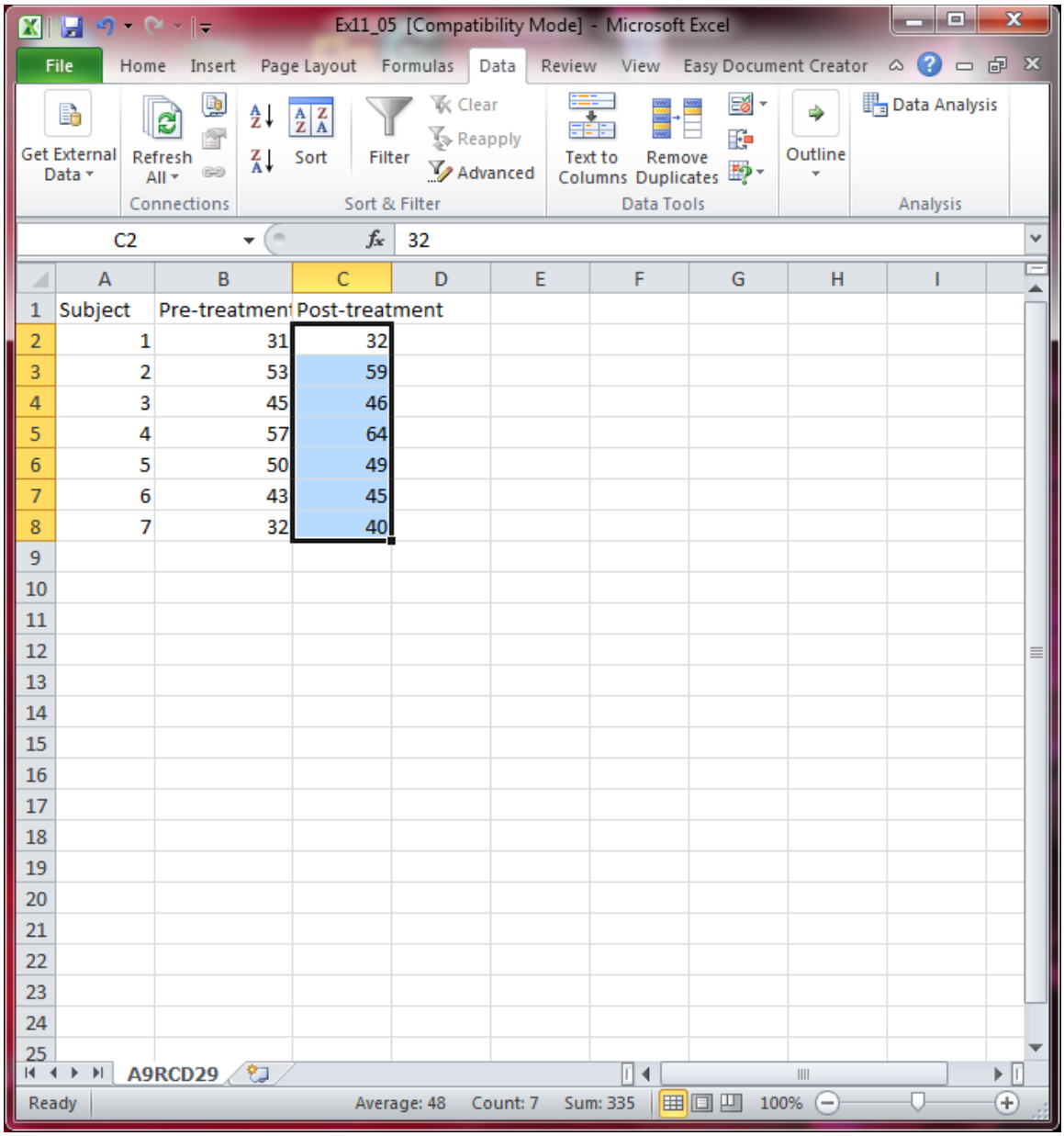
The results are shown below.



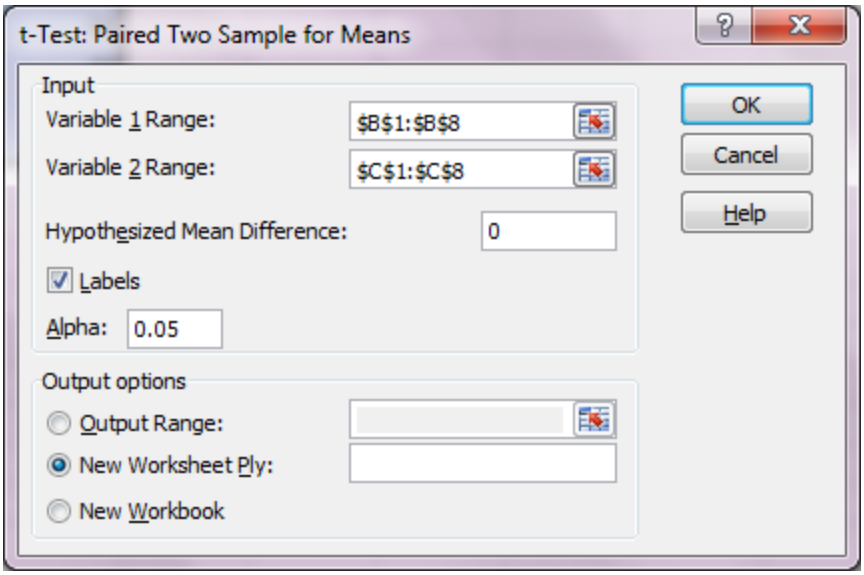
Example 11.8 – Confidence Interval and Inference Test for Paired Means

Ultrasound is often used in the treatment of soft tissue injuries. In an experiment to investigate the effect of an ultrasound and stretch therapy on knee extension, range of motion was measured both before and after treatment for a sample of physical therapy patients. A subset of the data appearing in the paper "Location of Ultrasound Does Not Enhance Range of Motion Benefits of Ultrasound and Stretch Treatment" (University of Virginia Thesis, Trae Tashiro, 2003) is given in the text.

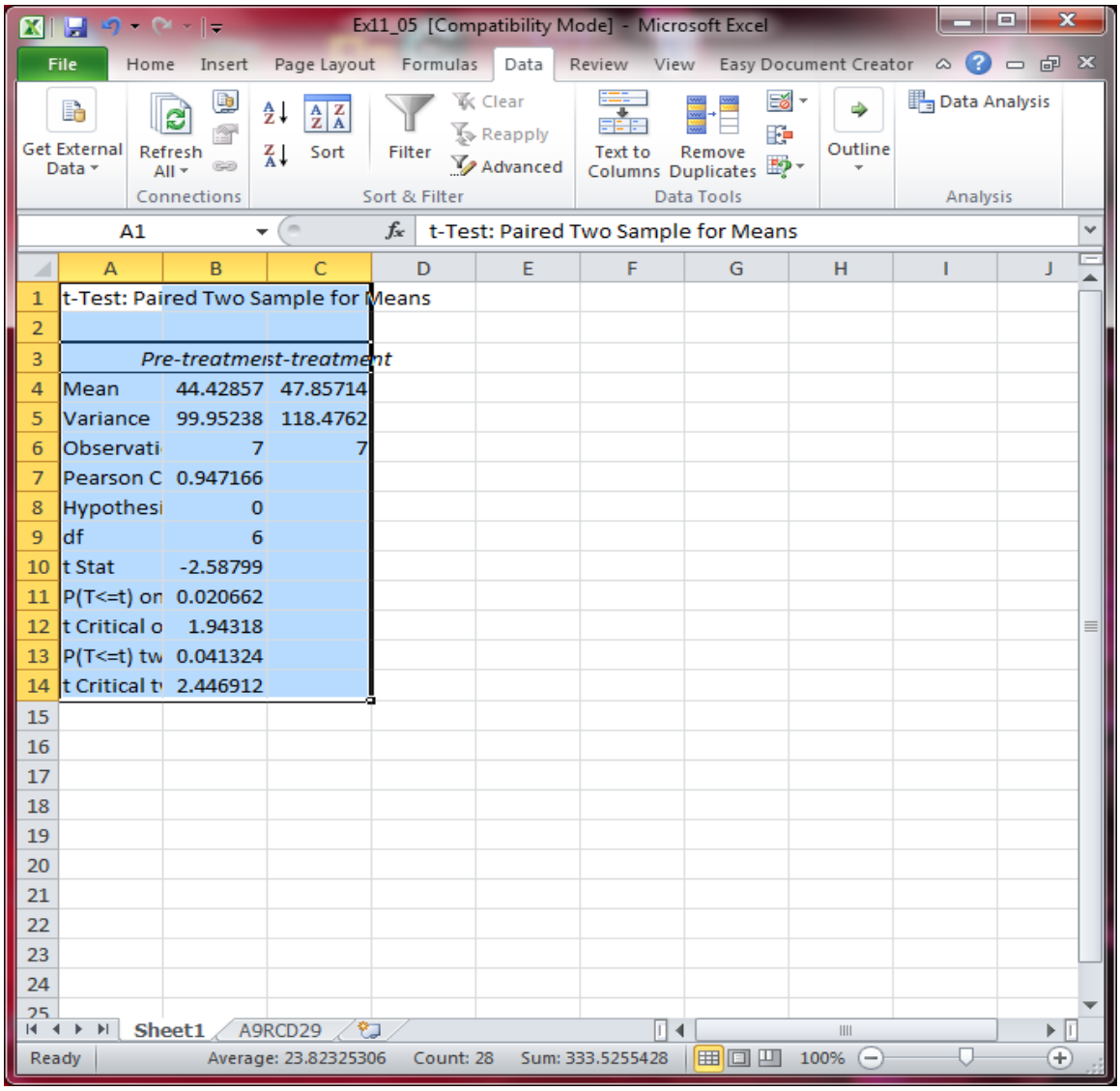
Begin by opening or entering the data into Excel as below.



Click **Data>Data Analysis>t-Test: Paired Two Sample for Means**. Input Pre-treatment for Variable 1 Range and Post-treatment for Variable 2 range. Check the box next to Labels. Input 0 for Hypothesized Mean Difference. Click **OK**.



The results are shown below.


Ex11_05 [Compatibility Mode] - Microsoft Excel
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Easy Document Creator
Get External Data
Refresh All
Connections
Sort
Filter
Clear
Reapply
Advanced
Sort & Filter
Text to Columns
Remove Duplicates
Data Tools
Outline
Data Analysis
Analysis
A1
t-Test: Paired Two Sample for Means
t-Test: Paired Two Sample for Means
Pre-treatment
t-treatment
Mean 44.42857 47.85714
Variance 99.95238 118.4762
Observati 7 7
Pearson C 0.947166
Hypothesi 0
df 6
t Stat -2.58799
P(T<=t) on 0.020662
t Critical o 1.94318
P(T<=t) tw 0.041324
t Critical t 2.446912
Sheet1
A9RCD29
Ready
Average: 23.82325306
Count: 28
Sum: 333.5255428
100%

Chapter 12
The Analysis of Categorical Data and Goodness-of-Fit Tests

The information in this chapter is specifically designed to explore the relationship between two categorical variables. Excel does not easily compute goodness-of-fit tests, so these are not addressed in this manual.

However, Excel does compute the Chi-squared Test for Independence. To compute this, we will need to calculate the expected values first then use the formula

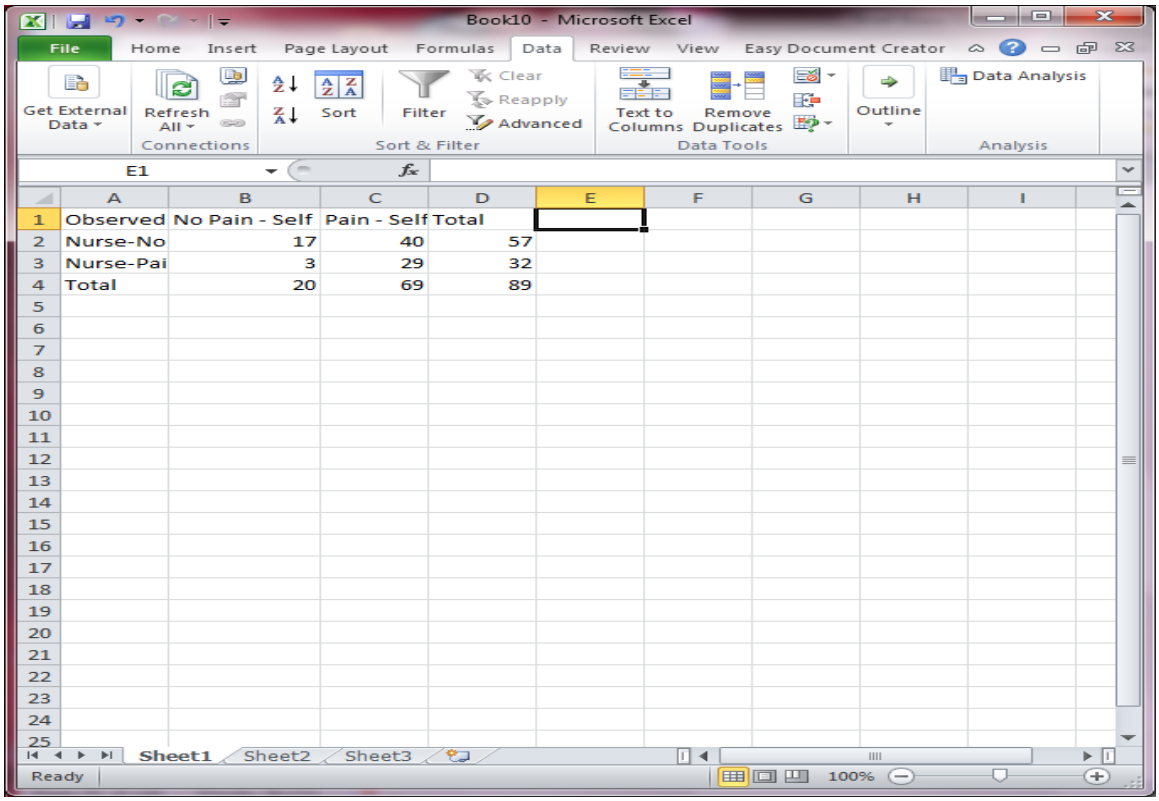
CHITEST

to complete analysis of this type.

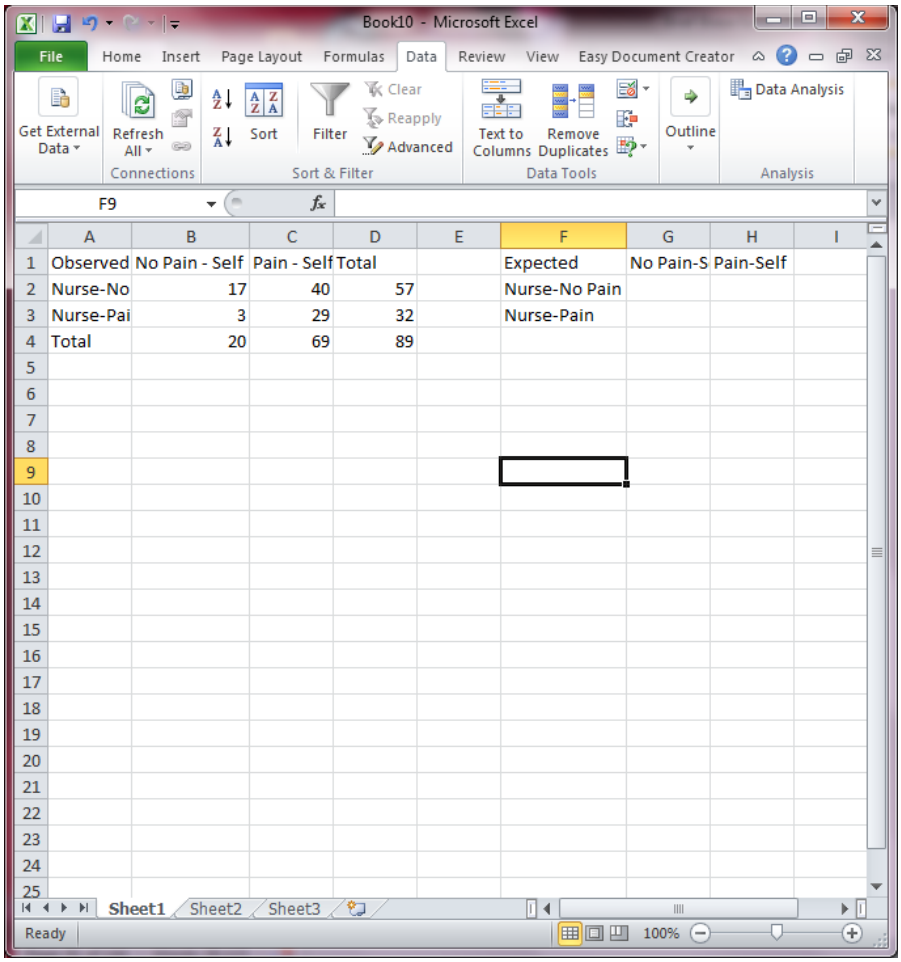
Example 12.7 – Chi-squared Test for Independence

The paper "Facial Expression of Pain in Elderly Adults with Dementia" (*Journal of Undergraduate Research* [2006]) examined the relationship between a nurse's assessment of a patient's facial expression and his or her self-reported level of pain. Data for 89 patients are summarized in the text. We use this data to test for independence.

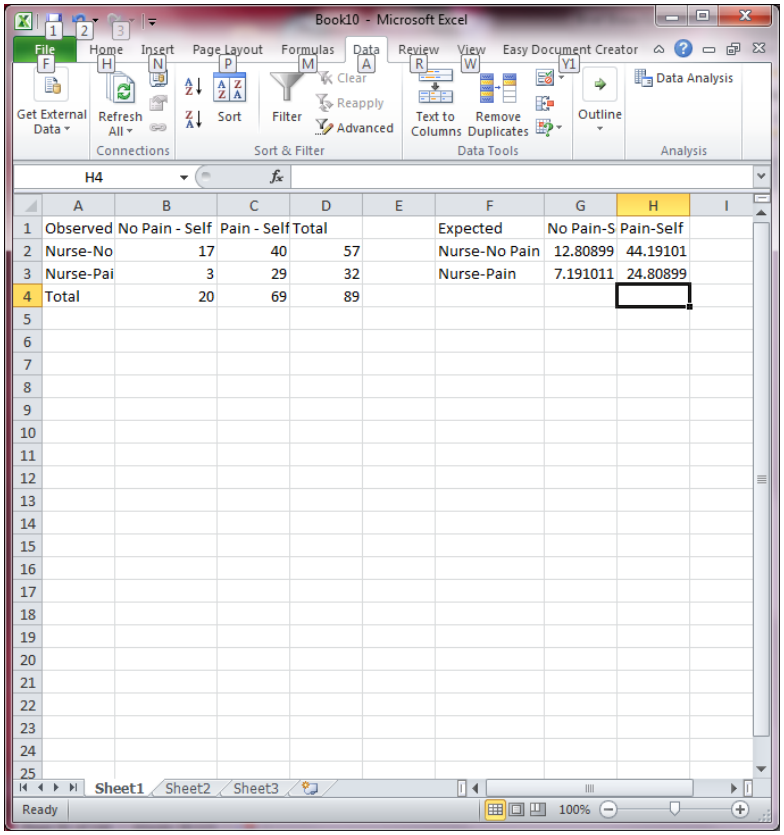
Begin by entering the two-way table below.



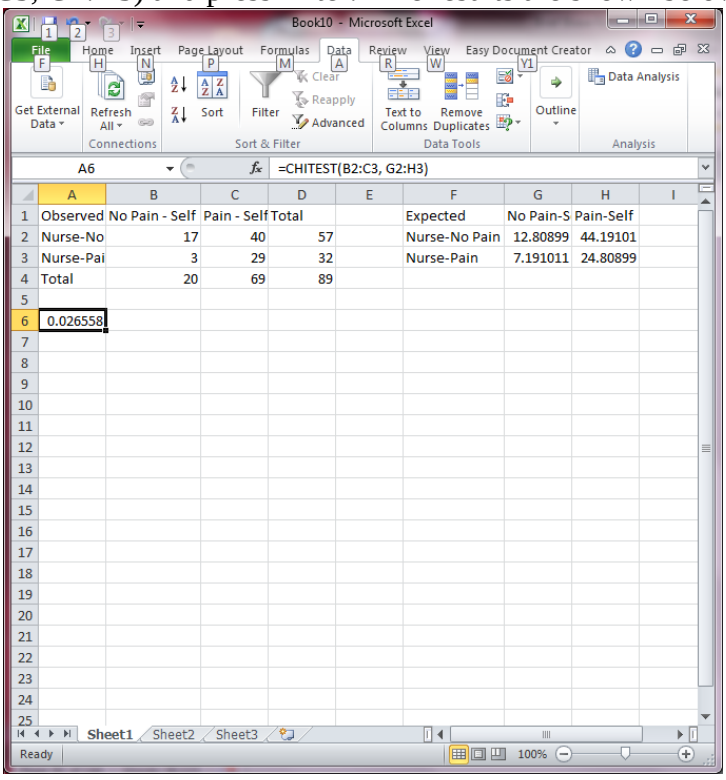
Input a table for Expected values as shown below.



To fill the table, use the formula from the text for each cell. For example, for the first cell, input =D2*B4/D4. The table is shown below.



In cell A6, we will input the p-value for the Chi-squared test. Click in cell A6 and input =CHITEST(B2:C3, G2:H3) and press Enter. The results are shown below.



Chapter 13
Simple Linear Regression and Correlation: Inferential Methods

Regression and correlation were introduced in Chapter 5 as techniques for describing and summarizing bivariate data. We continue this discussion in this chapter, moving into prediction and inferential methods for bivariate data.

We will use the commands

Data>Data Analysis>Regression

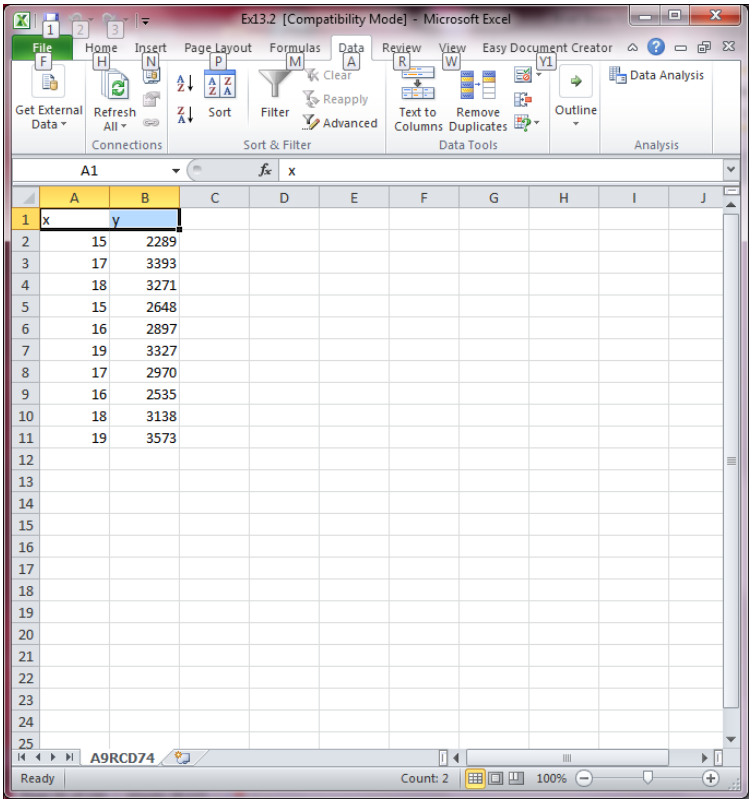
in this chapter.

Example 13.2 – Prediction (Regression)

Medical researchers have noted that adolescent females are much more likely to deliver low-birth-weight babies than are adult females. Because low-birth-weight babies have higher mortality rates, a number of studies have examined the relationship between birth weight and mother's age for babies born to young mothers.

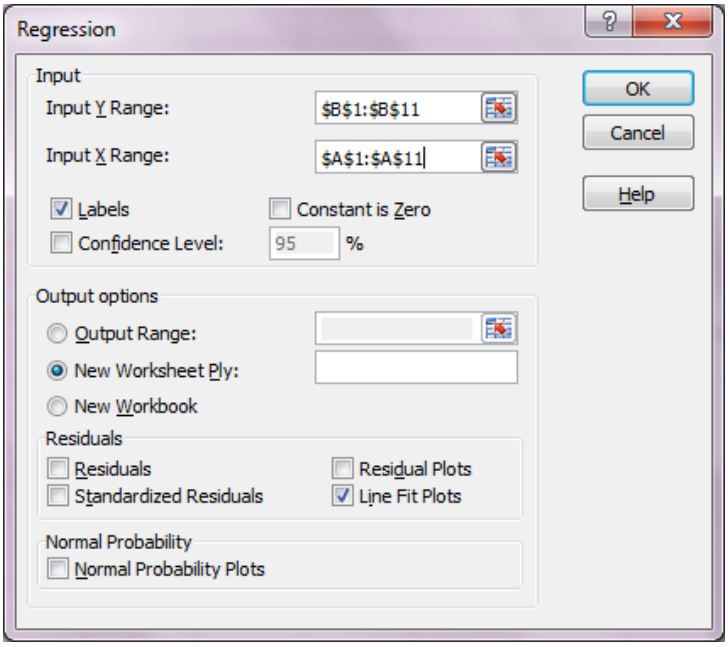
One such study is described in the article "Body Size and Intelligence in 6-Year-Olds: Are Offspring of Teenage Mothers at Risk?" (*Maternal and Child Health Journal* [2009]: 847-856). The data in the text on x = maternal age (in years) and y = birth weight of baby (in grams) are consistent with the summary values given in the referenced article and also with data published by the National Center for Health Statistics.

To predict baby weight for an 18-year-old mother, begin by opening the dataset in Excel.



	A	B
1	x	y
2	15	2289
3	17	3393
4	18	3271
5	15	2648
6	16	2897
7	19	3327
8	17	2970
9	16	2535
10	18	3138
11	19	3573

Click **Data>Data Analysis>Regression.** Input the data from column B in the Input Y Range and column A in the Input X Range. Check the box next to Labels. Click **OK**



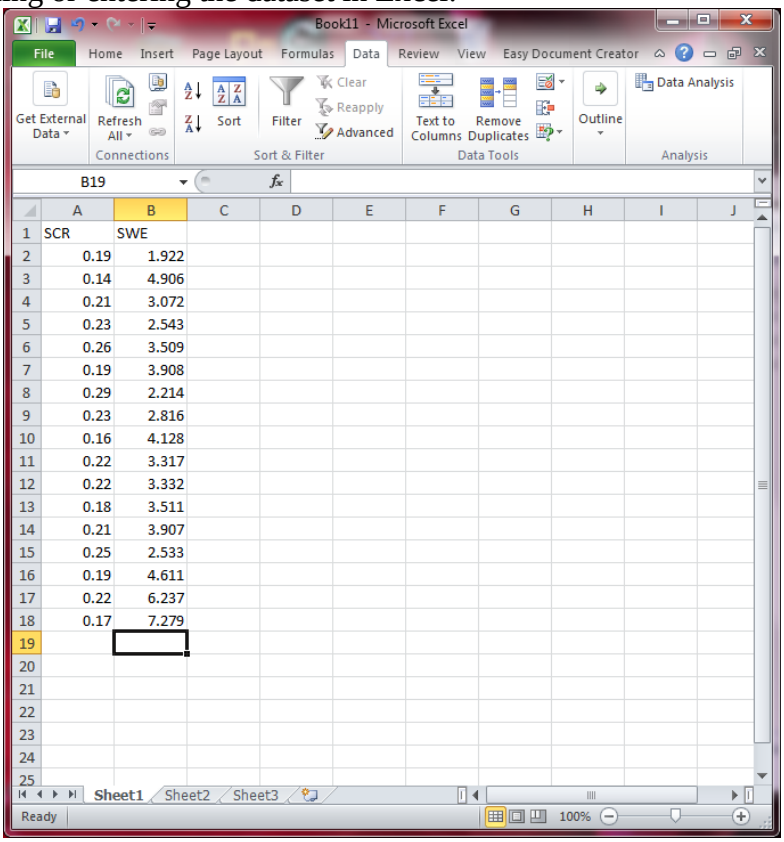
To find the prediction for x=18, input this value into the predicted line from the output: -1163.45+245.15*18 = 3249.25.

Example 13.4 – Confidence Interval for the Slope

The dedicated work of conservationists for over 100 years has brought the bison in Yellowstone National Park from near extinction to a herd of over 3000 animals. This recovery is a mixed blessing. Many bison have been exposed to the bacteria that cause brucellosis, a disease that infects domestic cattle, and there are many domestic herds near Yellowstone. Because of concerns that free-ranging bison can infect nearby cattle, it is important to monitor and manage the size of the bison population, and if possible, keep bison from transmitting this bacteria to ranch cattle.

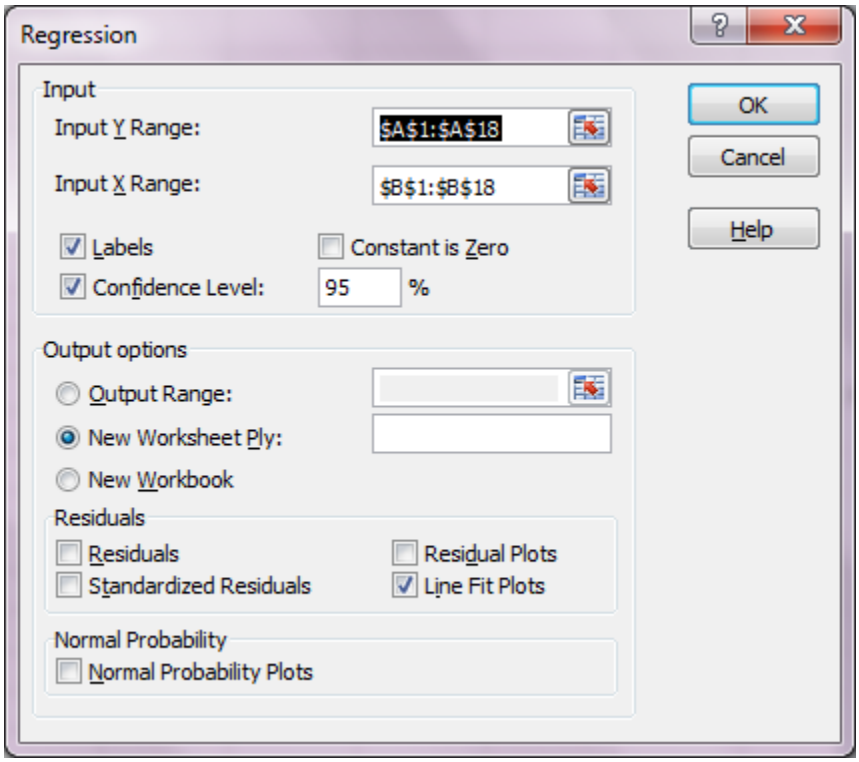
The article, "Reproduction and Survival of Yellowstone Bison" (*The Journal of Wildlife Management* [2007]: 2365-2372) described a large multiyear study of the factors that influence bison movement and herd size. The researchers studied a number of environmental factors to better understand the relationship between bison reproduction and the environment. One factor thought to influence reproduction is stress due to accumulated snow, making foraging more difficult for the pregnant bison. Data from 1981-1997 on y = spring calf ration (SCR) and x = previous fall snow-water equivalent (SWE) are shown in the text.

Begin by opening or entering the dataset in Excel.

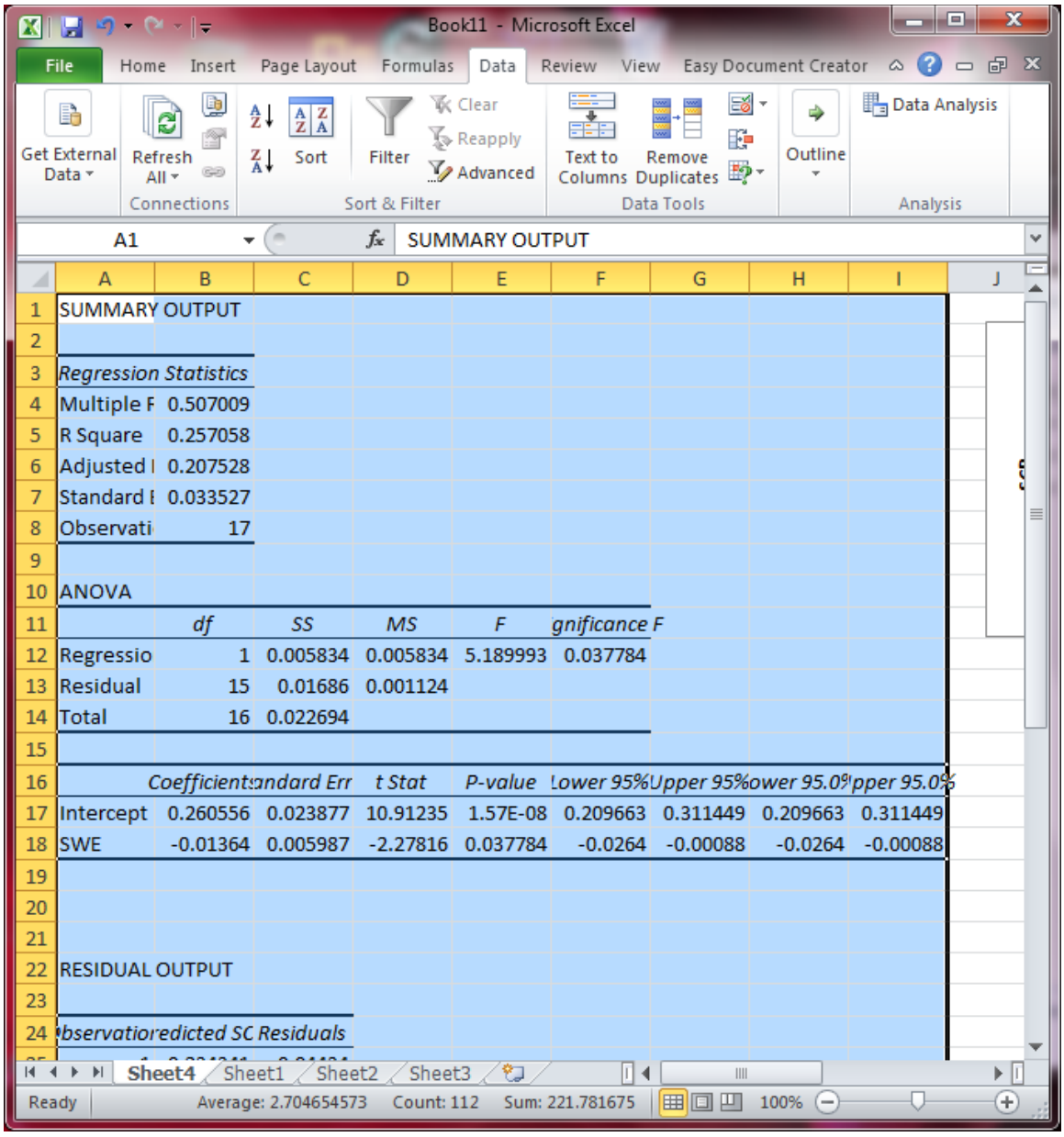


	A	B
1	SCR	SWE
2	0.19	1.922
3	0.14	4.906
4	0.21	3.072
5	0.23	2.543
6	0.26	3.509
7	0.19	3.908
8	0.29	2.214
9	0.23	2.816
10	0.16	4.128
11	0.22	3.317
12	0.22	3.332
13	0.18	3.511
14	0.21	3.907
15	0.25	2.533
16	0.19	4.611
17	0.22	6.237
18	0.17	7.279
19		

Click **Data>Data Analysis>Regression**. Input the SCR values into Input Y Range and the SWE values into the Input X Range. Check the box next to Labels and the box next to Confidence Level. Click **OK.**



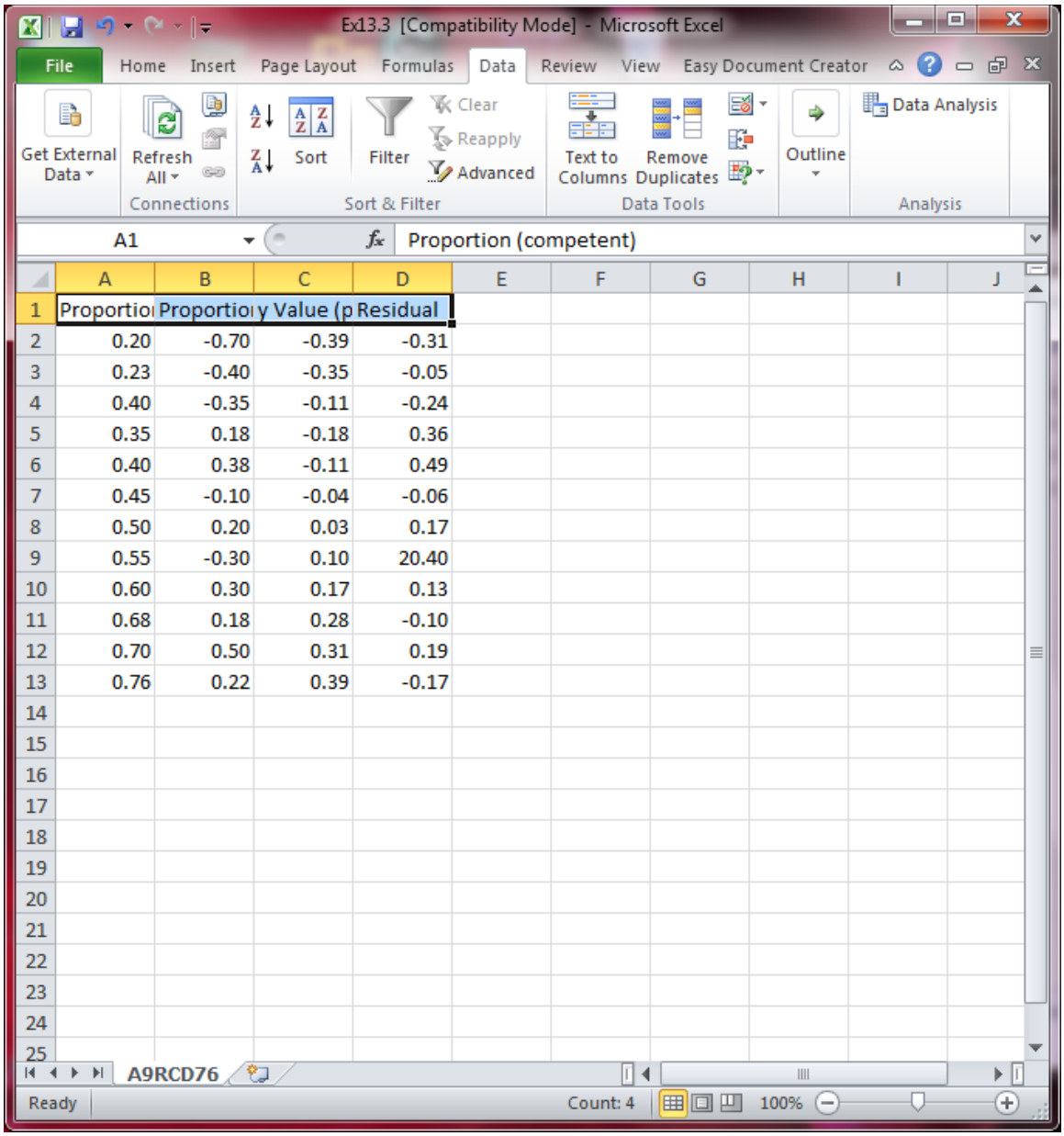
The results are shown below.



Example 13.6 – Computing residuals and displaying QQ plots

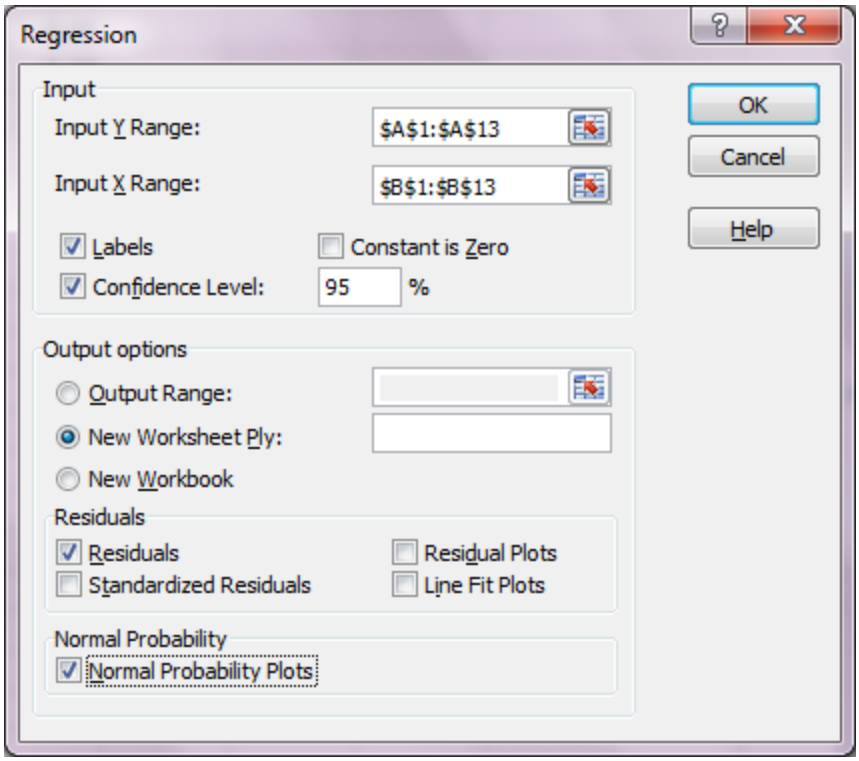
The authors of the paper "Inferences of Competence from Faces Predict Election Outcomes" (*Science* [2005]: 1623-1623) found that they could successfully predict the outcome of a U.S. congressional election substantially more than half the time based on the facial appearance of the candidates. In the study described in the paper, participants were shown photos of two candidates for a U.S. Senate or House of Representative election. Each participant was asked to look at the photos and then indicated which candidate he or she thought was more competent.

The data is listed in the text. Begin by opening or entering the dataset.

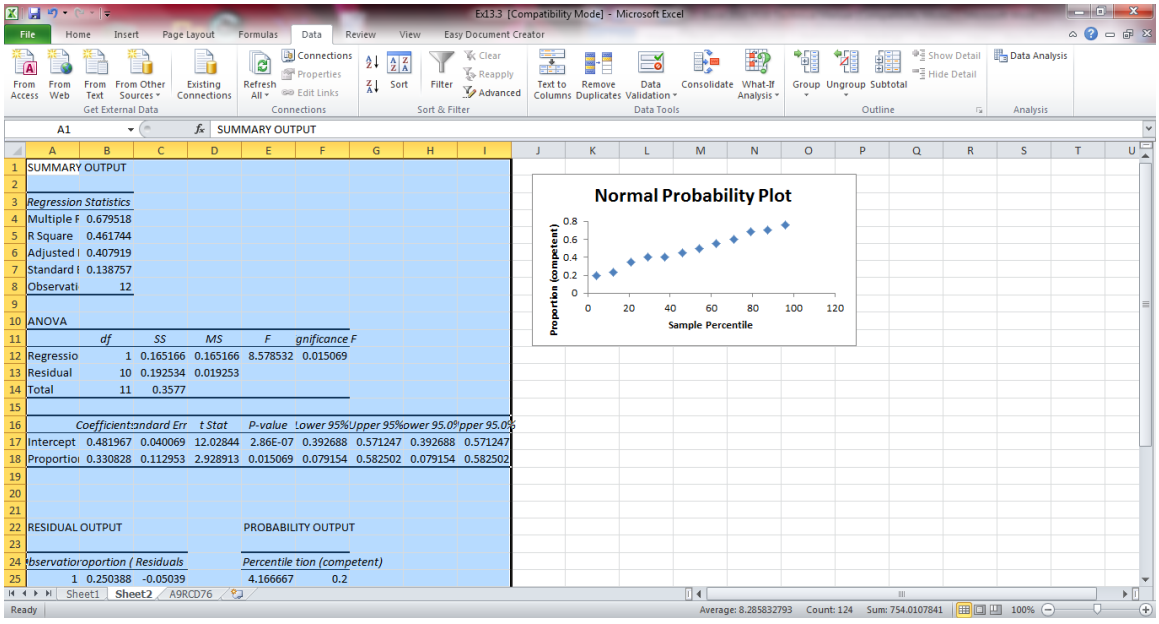


	A	B	C	D
1	Proportio	Proportio	y Value (p	Residual
2	0.20	-0.70	-0.39	-0.31
3	0.23	-0.40	-0.35	-0.05
4	0.40	-0.35	-0.11	-0.24
5	0.35	0.18	-0.18	0.36
6	0.40	0.38	-0.11	0.49
7	0.45	-0.10	-0.04	-0.06
8	0.50	0.20	0.03	0.17
9	0.55	-0.30	0.10	20.40
10	0.60	0.30	0.17	0.13
11	0.68	0.18	0.28	-0.10
12	0.70	0.50	0.31	0.19
13	0.76	0.22	0.39	-0.17

Click **Data>Data Analysis>Regression**. Input values from column A for Input Y Range and values from column B for Input X Range. Check the box next to Labels and Residuals and Normal Probability Plots. Click **OK.**



The results follow.

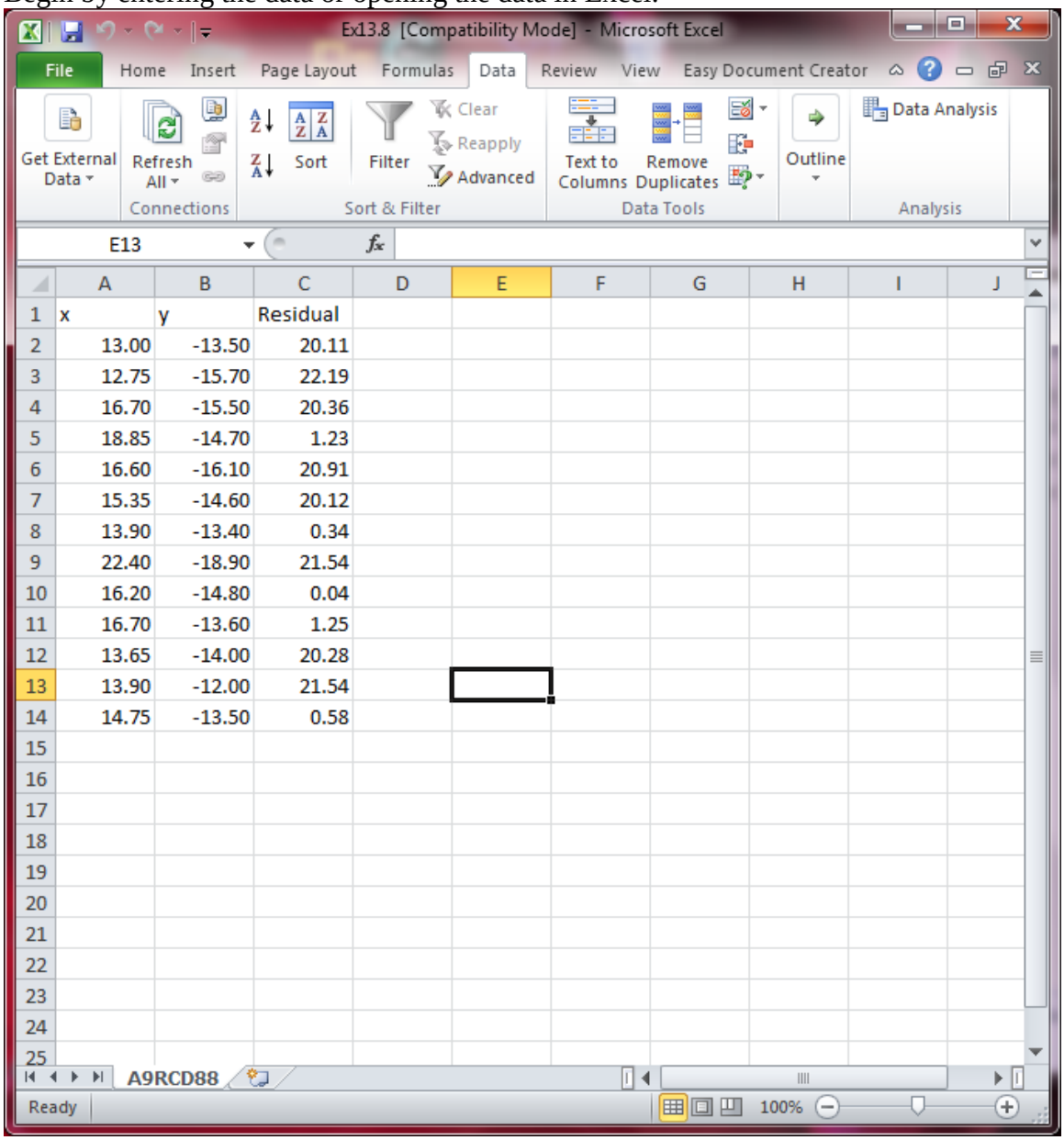


Example 13.9– Other Residual Plots

The article “Snow Cover and Temperature Relationships in North America and Eurasia”

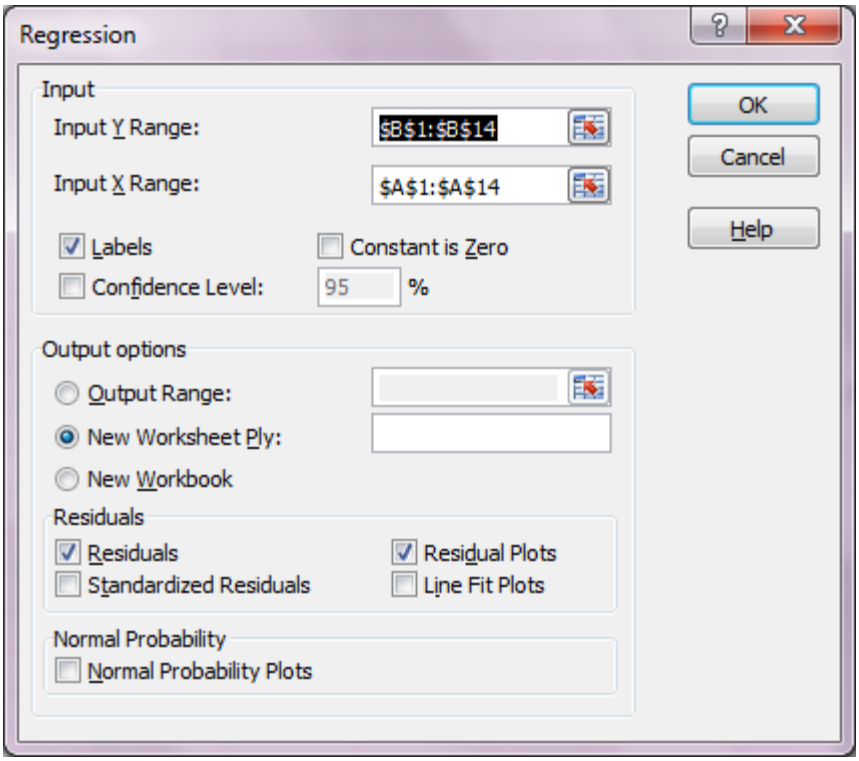
(*Journal of Climate and Applied Meteorology* [1983]: 460-469) explored the relationship between October-November continental snow cover (*x*, in millions of square kilometers) and December-February temperature (*y*, in °C). The data in the next are for Eurasia. We use this data to produce more probability plots.

Begin by entering the data or opening the data in Excel.

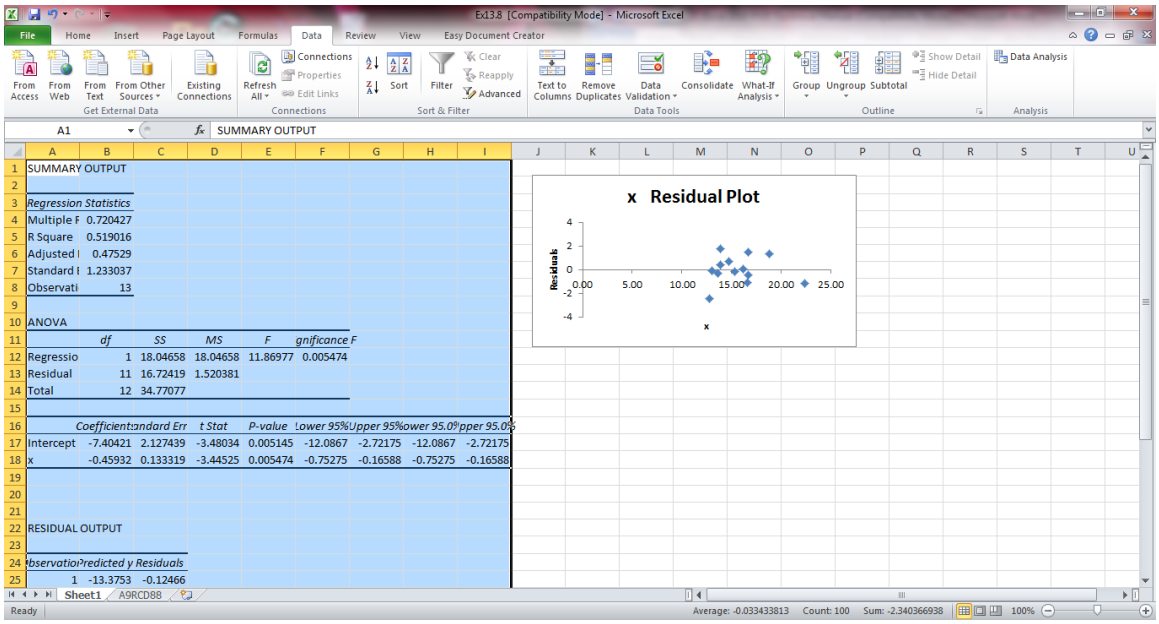


	A	B	C
1	x	y	Residual
2	13.00	-13.50	20.11
3	12.75	-15.70	22.19
4	16.70	-15.50	20.36
5	18.85	-14.70	1.23
6	16.60	-16.10	20.91
7	15.35	-14.60	20.12
8	13.90	-13.40	0.34
9	22.40	-18.90	21.54
10	16.20	-14.80	0.04
11	16.70	-13.60	1.25
12	13.65	-14.00	20.28
13	13.90	-12.00	21.54
14	14.75	-13.50	0.58

Click **Data>Data Analysis>Regression**. Input values from column B for Input Y Range and values from column A for Input X Range. Check the box next to Labels and Residuals and Residual Plots. Click **OK.**



The results follow.



Chapter 14
Multiple Regression Analysis

In this chapter, we consider multiple predictors in regression analysis. There are many similarities between multiple regression analysis and simple regression analysis. There is also a great deal more to consider when multiple predictors are involved.

In this chapter, we use Excel to estimate regression coefficients. Excel does not provide capabilities to consider the problem of model selection.

We use the commands

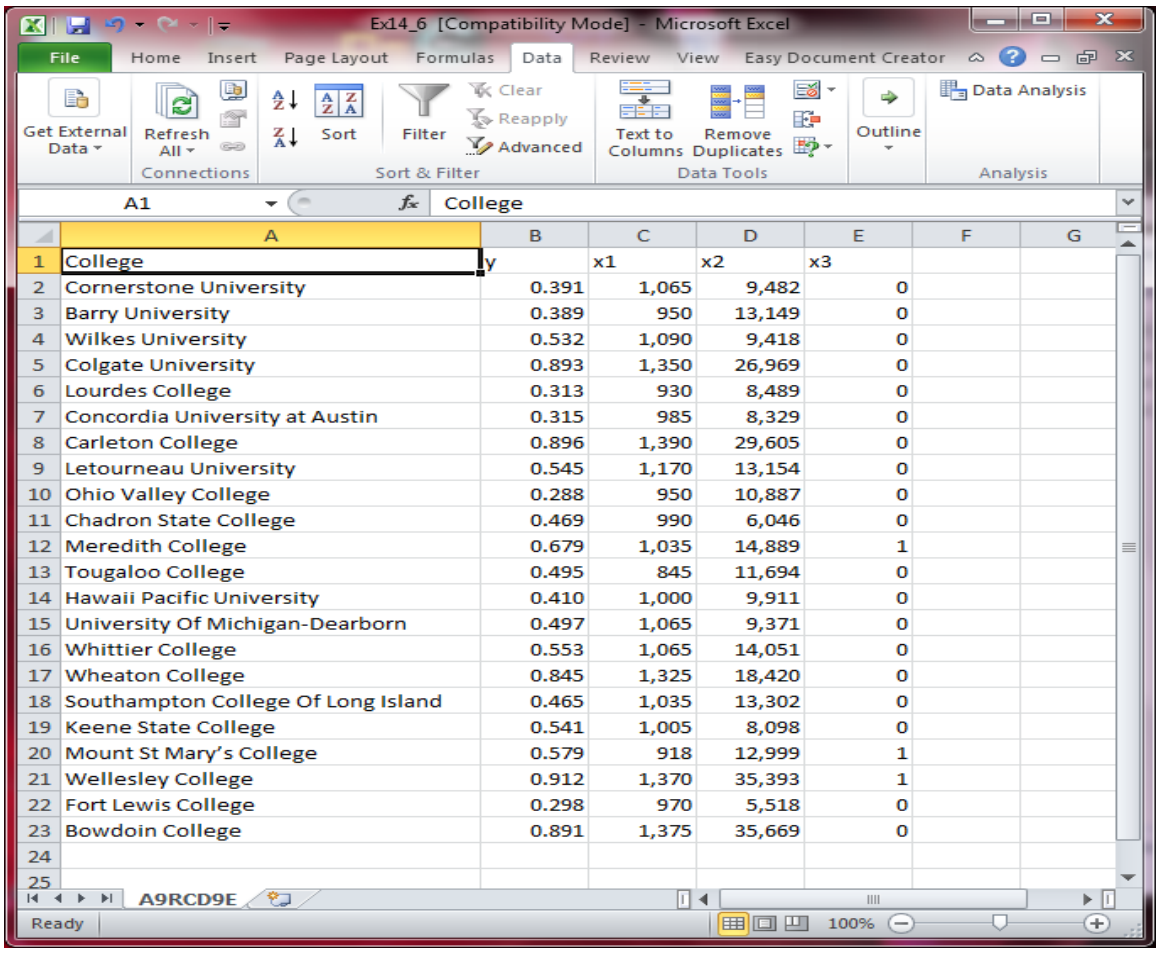
Data>Data Analysis>Regression.

Example 14.6 – Estimating Regression Coefficients

One way colleges measure success is by graduate rates. The Education Trust publishes 6-year graduation rates along with other college characteristics on its web site (www.collegeresults.org). We will consider the following variables:

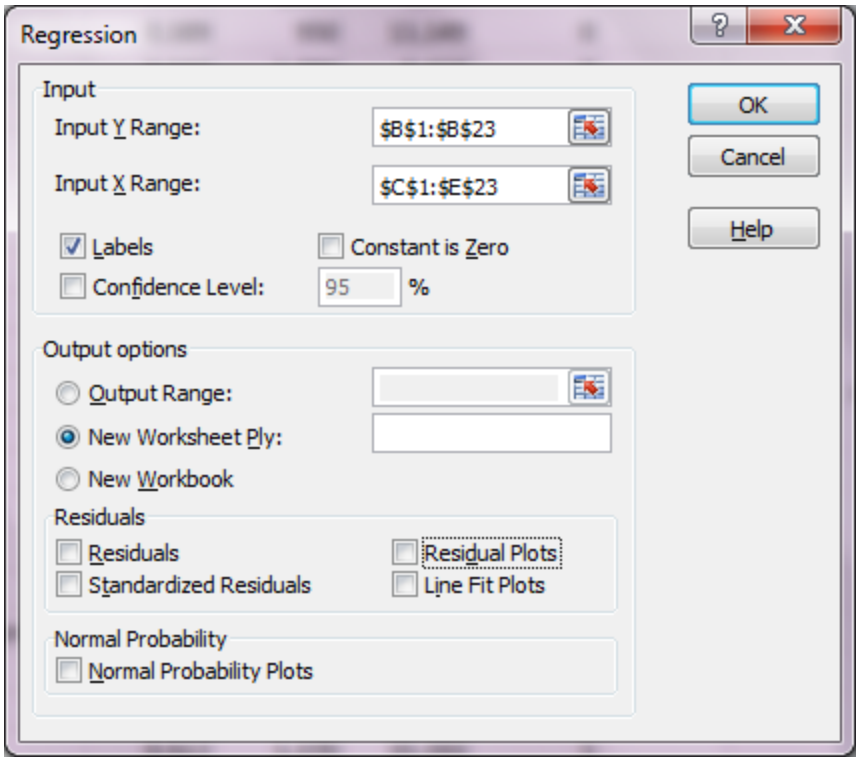
y = 6-year graduation rate
x_1 = median SAT score of students accepted to the college
x_2 = student-related expense per full-time student (in dollars)
x_3 = 1 if college has only female students or only male students; 0 if college has both male and female students

Begin by entering or opening the dataset in Excel.

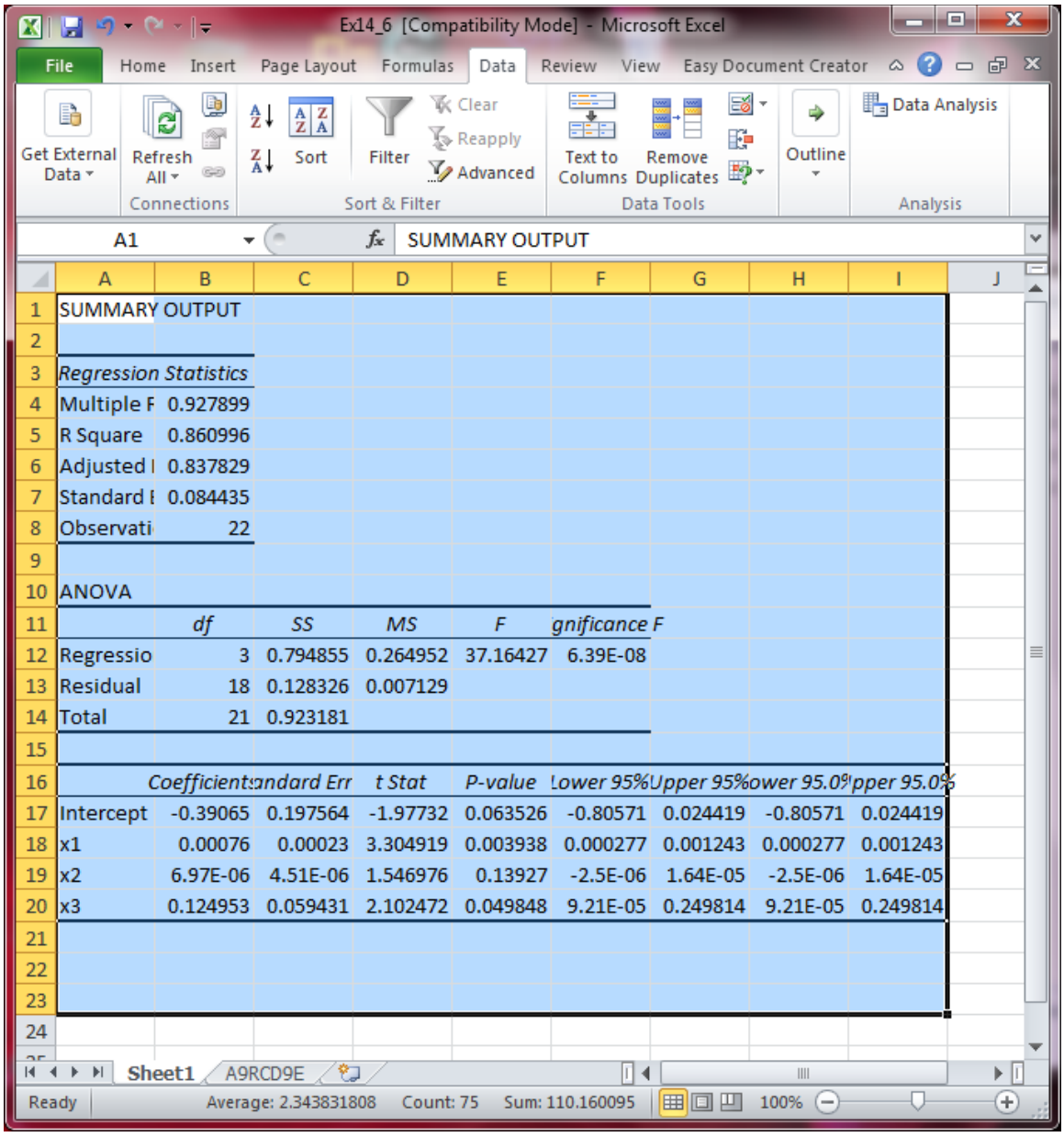


	A	B	C	D	E	F	G
1	College	y	x1	x2	x3		
2	Cornerstone University	0.391	1,065	9,482	0		
3	Barry University	0.389	950	13,149	0		
4	Wilkes University	0.532	1,090	9,418	0		
5	Colgate University	0.893	1,350	26,969	0		
6	Lourdes College	0.313	930	8,489	0		
7	Concordia University at Austin	0.315	985	8,329	0		
8	Carleton College	0.896	1,390	29,605	0		
9	Letourneau University	0.545	1,170	13,154	0		
10	Ohio Valley College	0.288	950	10,887	0		
11	Chadron State College	0.469	990	6,046	0		
12	Meredith College	0.679	1,035	14,889	1		
13	Tougaloo College	0.495	845	11,694	0		
14	Hawaii Pacific University	0.410	1,000	9,911	0		
15	University Of Michigan-Dearborn	0.497	1,065	9,371	0		
16	Whittier College	0.553	1,065	14,051	0		
17	Wheaton College	0.845	1,325	18,420	0		
18	Southampton College Of Long Island	0.465	1,035	13,302	0		
19	Keene State College	0.541	1,005	8,098	0		
20	Mount St Mary's College	0.579	918	12,999	1		
21	Wellesley College	0.912	1,370	35,393	1		
22	Fort Lewis College	0.298	970	5,518	0		
23	Bowdoin College	0.891	1,375	35,669	0		
24							
25							

Select **Data>Data Analysis>Regression.** Input the data in column B for Input Y Range. Input all data from column C, D and E for Input X range. Check the box next to Labels. Click **OK**.



The results follow.



	A	B	C	D	E	F	G	H	I	J
1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple F	0.927899								
5	R Square	0.860996								
6	Adjusted I	0.837829								
7	Standard E	0.084435								
8	Observati	22								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*gnificance F*				
12	Regressio	3	0.794855	0.264952	37.16427	6.39E-08				
13	Residual	18	0.128326	0.007129						
14	Total	21	0.923181							
15										
16		*Coefficients*	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*	
17	Intercept	-0.39065	0.197564	-1.97732	0.063526	-0.80571	0.024419	-0.80571	0.024419	
18	x1	0.00076	0.00023	3.304919	0.003938	0.000277	0.001243	0.000277	0.001243	
19	x2	6.97E-06	4.51E-06	1.546976	0.13927	-2.5E-06	1.64E-05	-2.5E-06	1.64E-05	
20	x3	0.124953	0.059431	2.102472	0.049848	9.21E-05	0.249814	9.21E-05	0.249814	
21										
22										
23										
24										

Example 14.17 – Model Selection

Model selection cannot be performed using Excel.

Example 14.18 – Model Selection

Model selection cannot be performed using Excel.

Chapter 15
Analysis of Variance

In previous chapters, we have discussed comparison of means for two groups. We now consider comparison of means for more than two groups. Analysis of variance provides a method to compare means amongst multiple groups.

In this chapter, we use Excel to produce ANOVA results using the commands

Data>Data Analysis>Anova: Single Factor

and

Data>Data Analysis>Anova: Two-Factor Without Replication

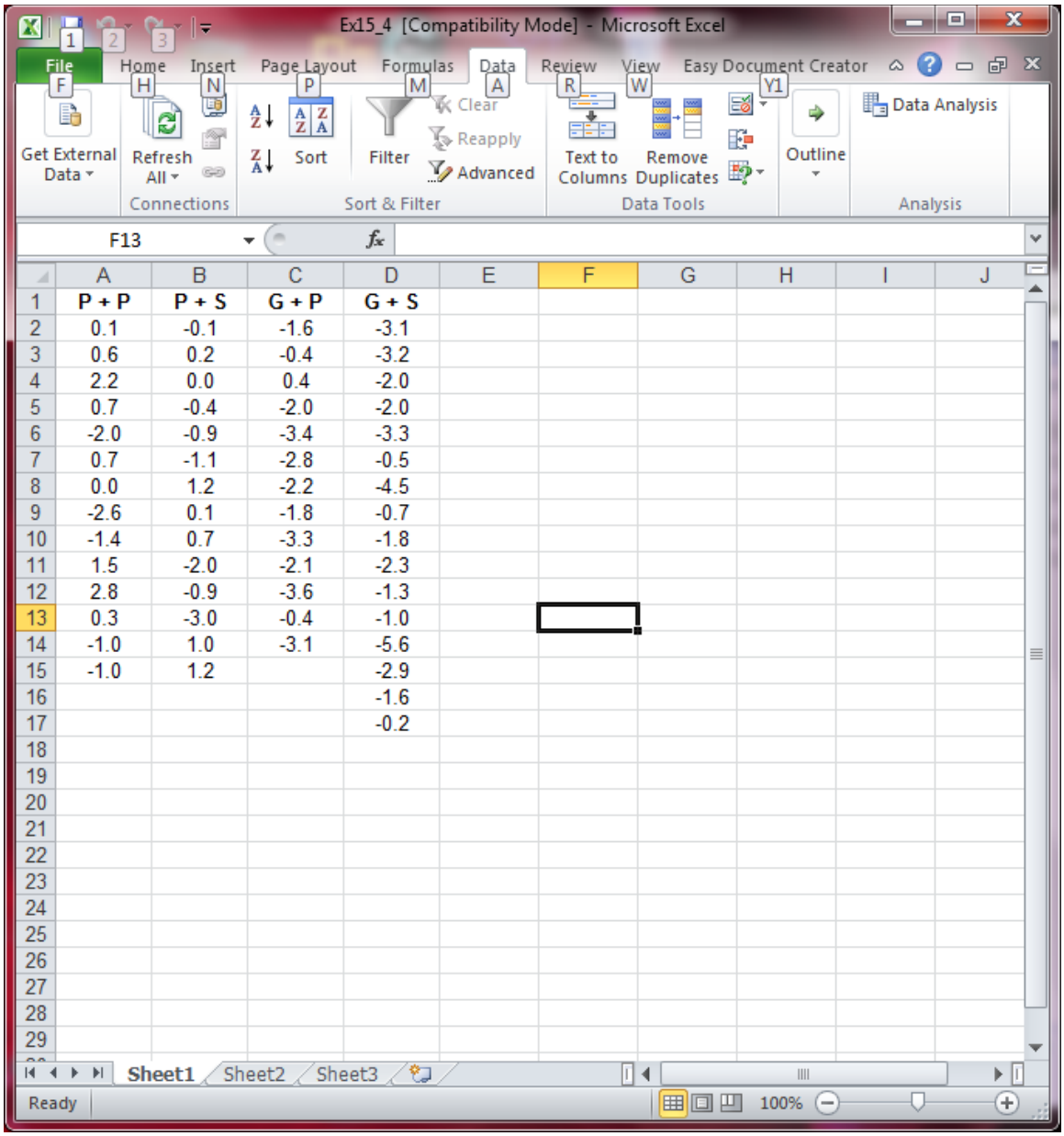
Example 15.4 – ANOVA

The article "Growth Hormone and Sex Steroid Administration in Healthy Aged Women and Men" (*Journal of the American Medical Association* [2002]: 2282-2292) described an experiment to investigate the effect of four treatments on various body characteristics. In this double-blind experiment, each of 57 female subjects age 65 or older was assigned at random to one of the following four treatments:

1. Placebo "growth hormone" and placebo "steroid" (denoted by P + P);
2. Placebo "growth hormone" and the steroid estradoil (denoted by P + S);
3. Growth hormone and placebo "steroid" (denoted G + P); and
4. Growth hormone and the steroid estradoil (denoted by G + S).

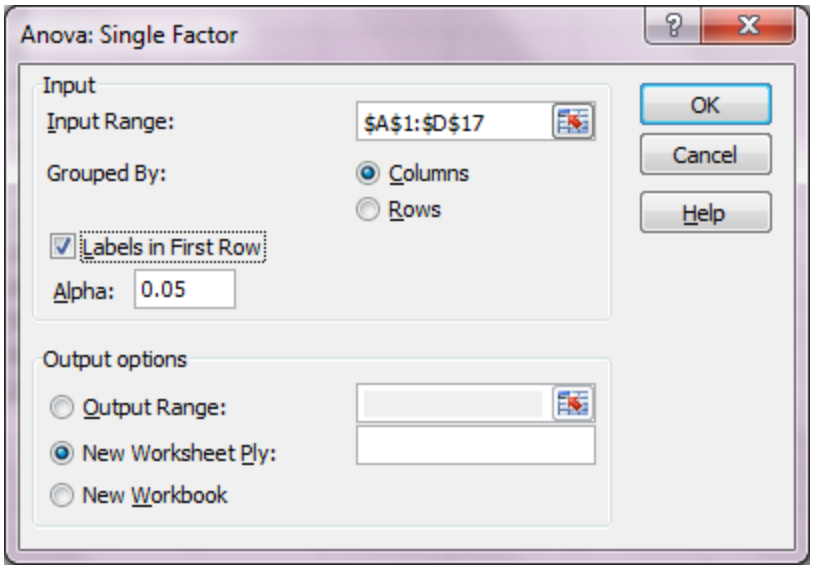
We perform an ANOVA to determine if the mean change in body fat mass differs for the four treatments.

Begin by opening or inputting the data in Excel.

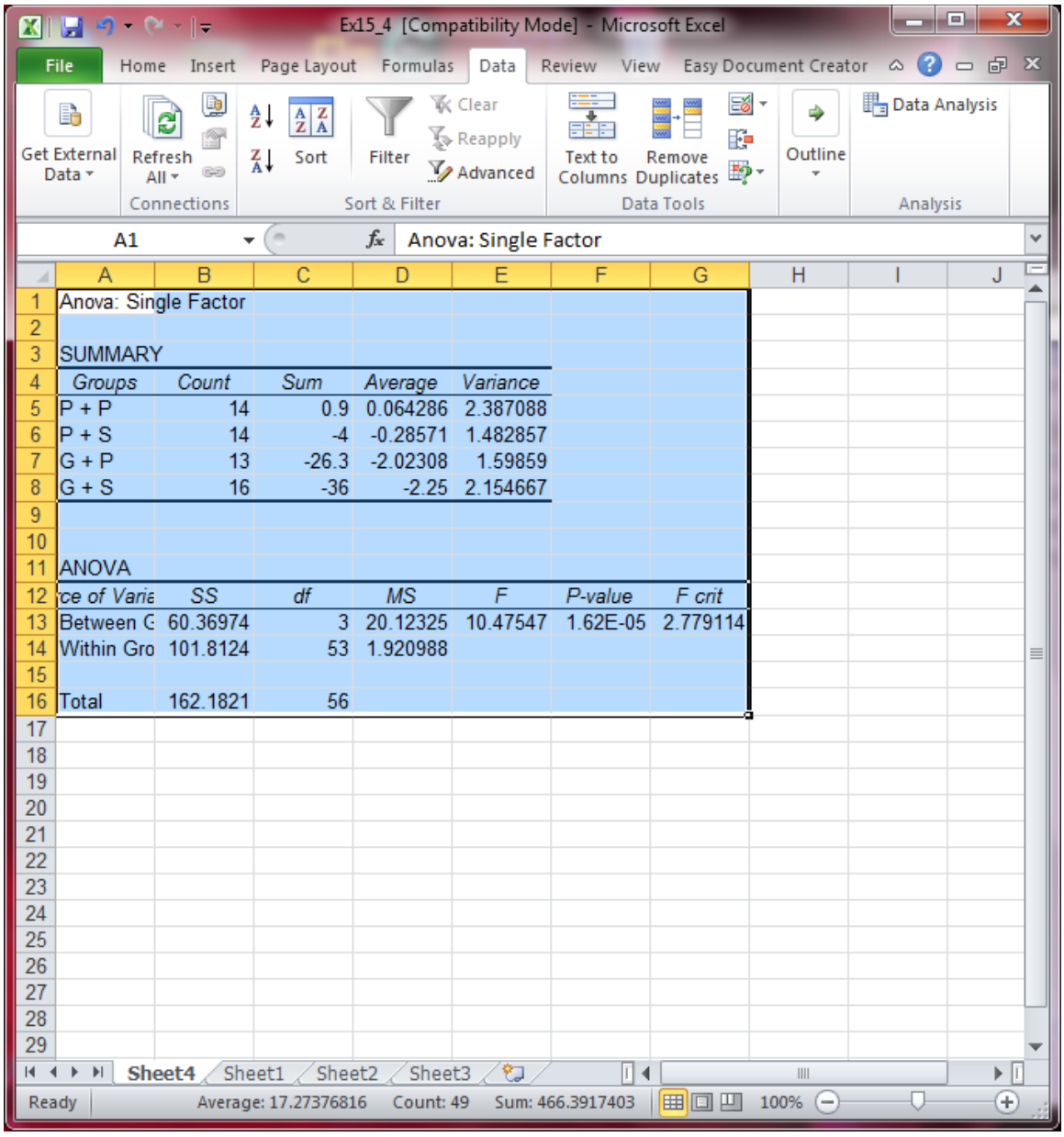


	A	B	C	D
1	**P + P**	**P + S**	**G + P**	**G + S**
2	0.1	-0.1	-1.6	-3.1
3	0.6	0.2	-0.4	-3.2
4	2.2	0.0	0.4	-2.0
5	0.7	-0.4	-2.0	-2.0
6	-2.0	-0.9	-3.4	-3.3
7	0.7	-1.1	-2.8	-0.5
8	0.0	1.2	-2.2	-4.5
9	-2.6	0.1	-1.8	-0.7
10	-1.4	0.7	-3.3	-1.8
11	1.5	-2.0	-2.1	-2.3
12	2.8	-0.9	-3.6	-1.3
13	0.3	-3.0	-0.4	-1.0
14	-1.0	1.0	-3.1	-5.6
15	-1.0	1.2		-2.9
16				-1.6
17				-0.2

Select **Data>Data Analysis>Anova: Single Factor.** Select all data for Input Range and check Labels in First Row. Click **OK**.



The results follow.



Example 15.6 – Multiple Comparisons

Excel does not have the capability to produce multiple comparisons.

Example 15.10 – Block Design

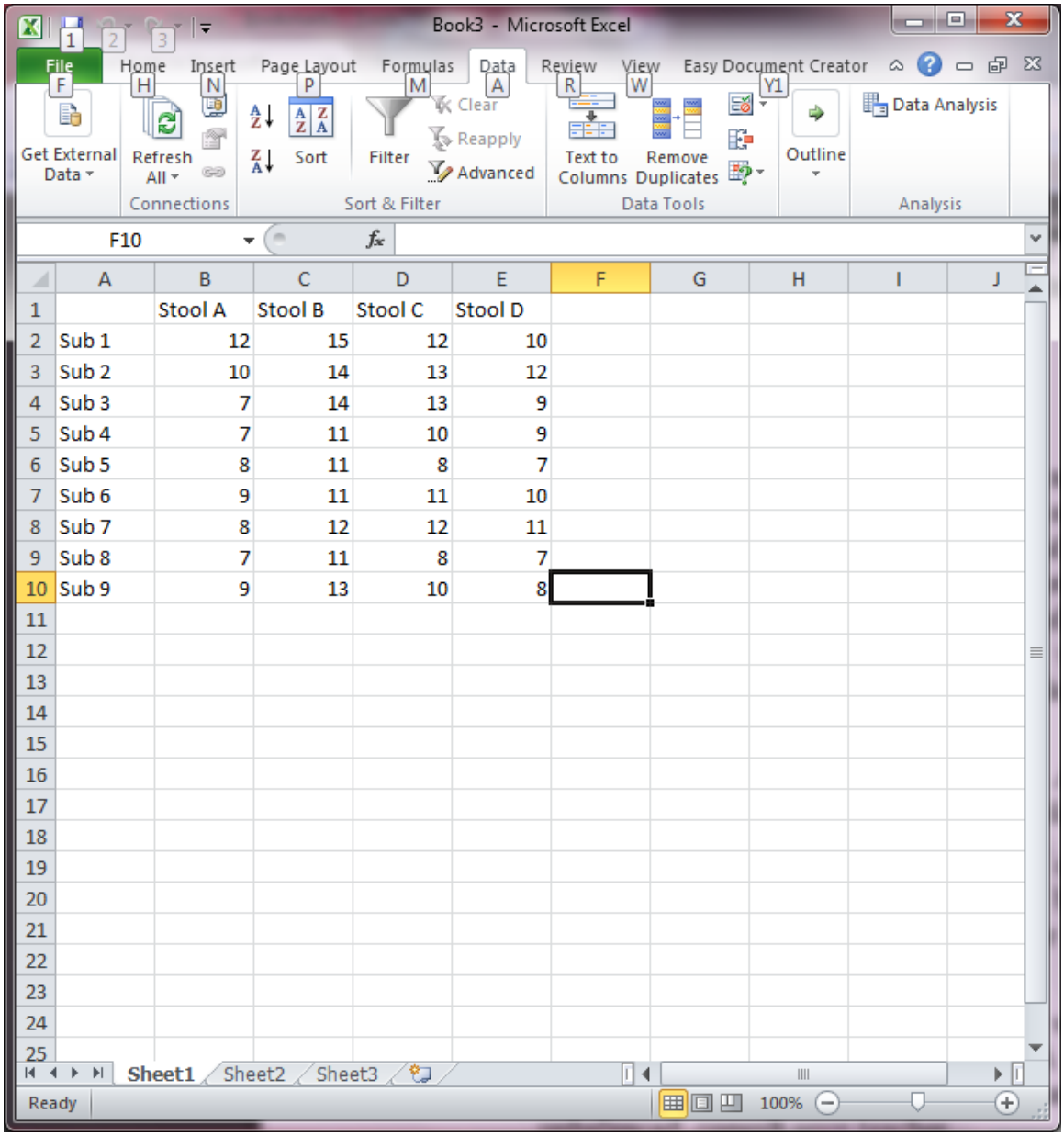
In the article "The Effects of a Pneumatic Stool and a One-Legged Stool on Lower Limb Joint Load and Muscular Activity During Sitting and Rising" (*Ergonomics* [1993]: 519-535), the accompanying data were given on the effort required by a subject to rise from a

sitting position for each of the four different stools. Because it was suspected that different people could exhibit large differences in effort, even for the same type of stool, a sample of nine people was selected and each person was tested on all four stools. Results are displayed in the text.

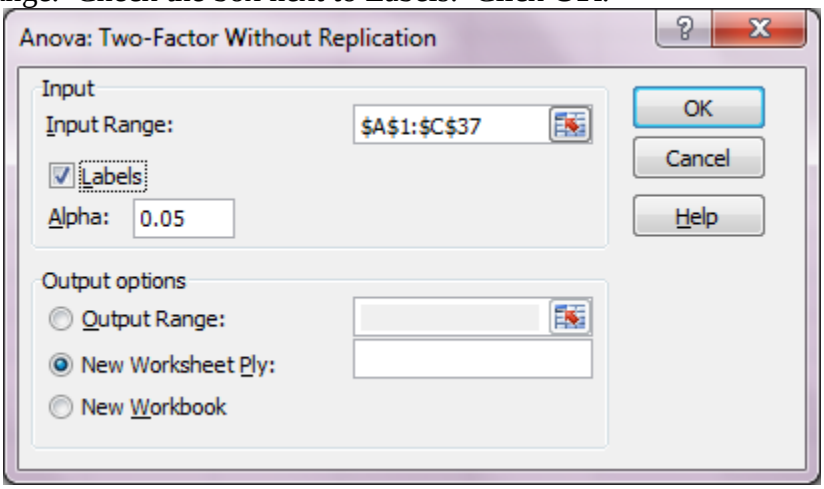
For each person, the order in which the stools were testing was randomized. This is a randomized block experiment, with subjects playing the role of blocks.

We create an ANOVA using block design.

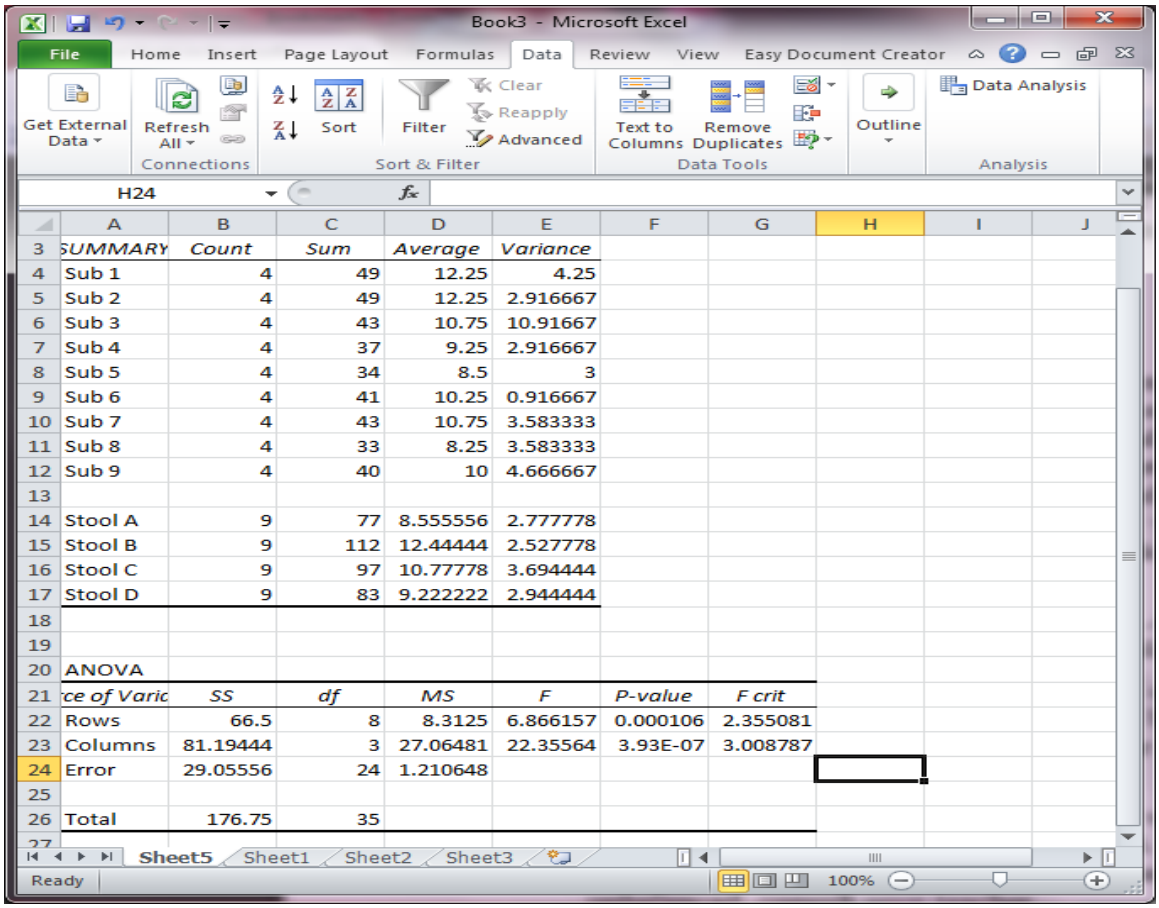
Begin by inputting or opening the data in Excel.



Select **Data>Data Analysis>Anova: Two-Factor Without Replication**. Input the data into Input Range. Check the box next to Labels. Click **OK**.



The results are shown copied below.

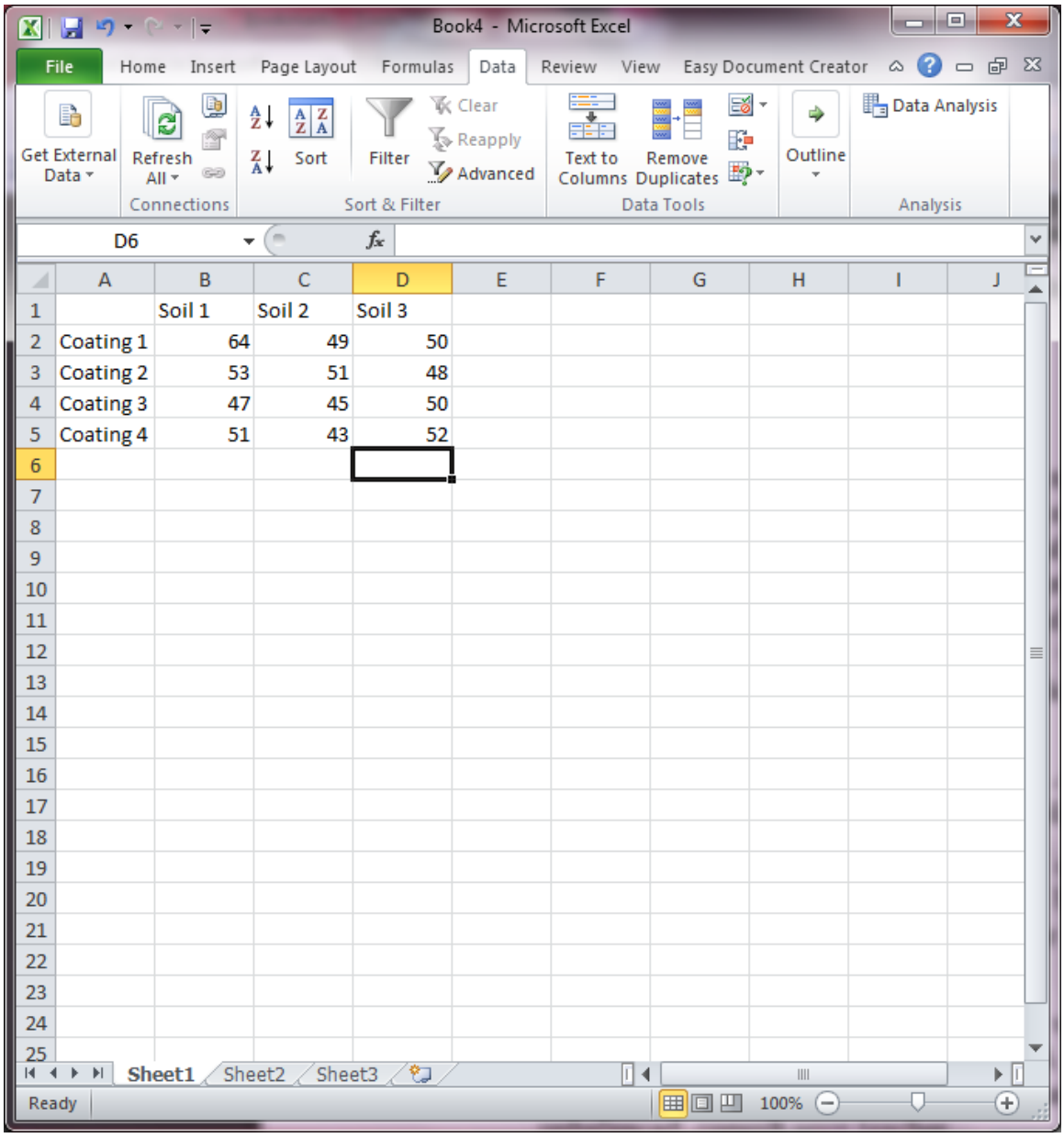


	A	B	C	D	E	F	G	H	I	J
3	SUMMARY	Count	Sum	Average	Variance					
4	Sub 1	4	49	12.25	4.25					
5	Sub 2	4	49	12.25	2.916667					
6	Sub 3	4	43	10.75	10.91667					
7	Sub 4	4	37	9.25	2.916667					
8	Sub 5	4	34	8.5	3					
9	Sub 6	4	41	10.25	0.916667					
10	Sub 7	4	43	10.75	3.583333					
11	Sub 8	4	33	8.25	3.583333					
12	Sub 9	4	40	10	4.666667					
13										
14	Stool A	9	77	8.555556	2.777778					
15	Stool B	9	112	12.44444	2.527778					
16	Stool C	9	97	10.77778	3.694444					
17	Stool D	9	83	9.222222	2.944444					
18										
19										
20	ANOVA									
21	ce of Varic	SS	df	MS	F	P-value	F crit			
22	Rows	66.5	8	8.3125	6.866157	0.000106	2.355081			
23	Columns	81.19444	3	27.06481	22.35564	3.93E-07	3.008787			
24	Error	29.05556	24	1.210648						
25										
26	Total	176.75	35							

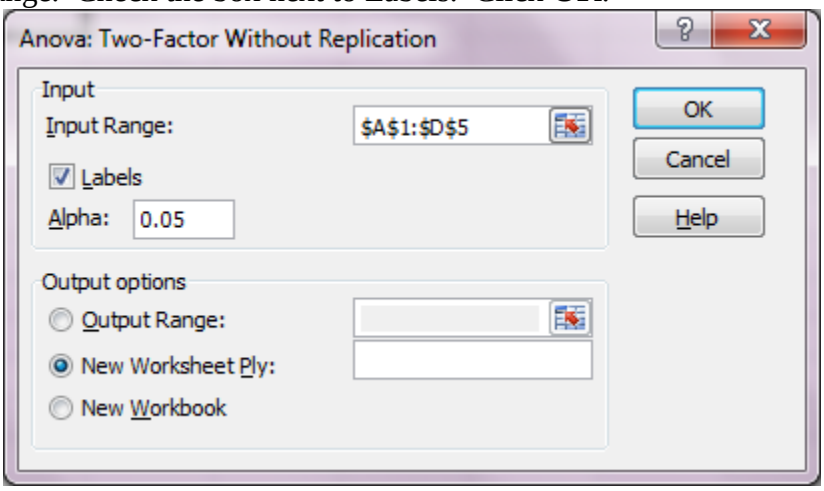
Example 15.14 – Two Way ANOVA

When metal pipe is buried in soil, it is desirable to apply a coating to retard corrosion. Four different coatings are under consideration for use with pipe that will ultimately be buried in three types of soil. An experiment to investigate the effects of these coatings and soils was carried out by first selecting 12 pipe segments and applying each coating to 3 segments. The segments were then buried in soil for a specified period in such a way that each soil type received on piece with each coating. The resulting table is shown in the text. Assuming there is no interaction, test for the presence of separate coating and soil effects.

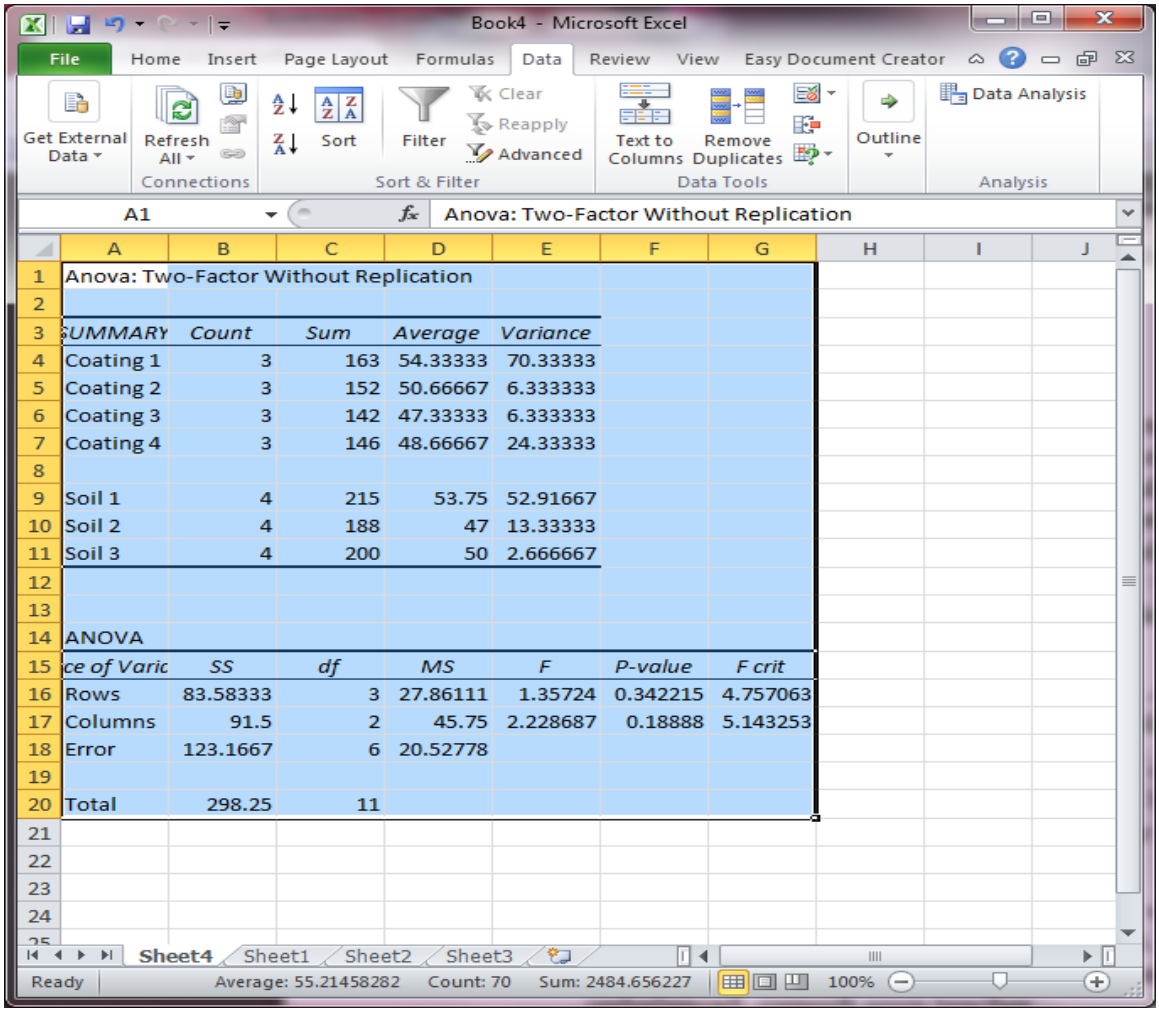
Begin by inputting or opening the dataset in Excel.



Select **Data>Data Analysis>Anova: Two-Factor Without Replication**. Input the data into Input Range. Check the box next to Labels. Click **OK**.



Results are shown below.



	A	B	C	D	E	F	G
1	Anova: Two-Factor Without Replication						
2							
3	SUMMARY	Count	Sum	Average	Variance		
4	Coating 1	3	163	54.33333	70.33333		
5	Coating 2	3	152	50.66667	6.333333		
6	Coating 3	3	142	47.33333	6.333333		
7	Coating 4	3	146	48.66667	24.33333		
8							
9	Soil 1	4	215	53.75	52.91667		
10	Soil 2	4	188	47	13.33333		
11	Soil 3	4	200	50	2.666667		
12							
13							
14	ANOVA						
15	ce of Varic	SS	df	MS	F	P-value	F crit
16	Rows	83.58333	3	27.86111	1.35724	0.342215	4.757063
17	Columns	91.5	2	45.75	2.228687	0.18888	5.143253
18	Error	123.1667	6	20.52778			
19							
20	Total	298.25	11				
21							
22							
23							
24							

Chapter 16
Nonparametric (Distribution-Free) Statistical Methods

There are no examples or material from Chapter 16.