STATDISK 12 STUDENT LABORATORY MANUAL AND WORKBOOK

MARIO F. TRIOLA
Dutchess Community College

TRIOLA STATISTICS SERIES

TWELFTH EDITION

Mario F. Triola
Dutchess Community College

PEARSON

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto
Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Publishing as Pearson, 75 Arlington Street, Boston, MA 02116.

ISBN-13: 978-0-321-83381-5
ISBN-10: 0-321-83381-3

www.pearsonhighered.com

PEARSON

Statdisk 12

Student Laboratory Manual and Workbook

Version 1.0

Triola Statistics Series

Mario F. Triola

To accompany the Triola Statistics Series:

Elementary Statistics, Twelfth Edition
Essentials of Statistics, Fifth Edition
Elementary Statistics Using Excel, Fifth Edition
Elementary Statistics Using the TI-83/84 Plus Calculator, Third Edition

Statdisk

Statdisk 12

Student Laboratory Manual and Workbook

Version 1.0

The Triola Statistics Series:

Elementary Statistics, Twelfth Edition

Essentials of Statistics, Fifth Edition

Elementary Statistics Using Excel, Fifth Edition

Elementary Statistics Using the TI-83/84 Plus Calculator, Third Edition

Mario F. Triola

PEARSON

Statdisk

ISBN 0-321-83381-3

PEARSON www.pearsonhighered.com

Preface

The Statdisk 12 Student Laboratory Manual and Workbook and Statdisk version 12 and later are supplements to the Triola Statistics Series of textbooks:

- *Elementary Statistics*, Twelfth Edition
- *Essentials of Statistics*, Fifth Edition
- *Elementary Statistics Using Excel*, Fifth Edition
- *Elementary Statistics Using the TI-83/84 Plus Calculator*, Third Edition

Statdisk is also a supplement to Biostatistics for the Biological and Health Science by Marc. M. Triola, M.D. and Mario F. Triola, and Statistical Reasoning for Everyday Life, 3rd Edition, by Jeffrey O. Bennett, William L. Briggs, and Mario F. Triola.

Downloading Statdisk

- Statdisk is available as a free download at www.statdisk.org.

Objectives The major objectives of this manual/workbook and the Statdisk software include:

- Describe how Statdisk can be used for the methods of statistics presented in the textbook. Specific and detailed procedures for using Statdisk are included along with examples of Statdisk screen displays.
- Incorporate and important component of computer usage without using valuable class time required for concepts of statistics.
- Replace tedious calculations or manual construction of graphs with computer results.
- Apply alternative methods, such as simulations, that are possible with computer usage.
- Include topics, such as analysis of variance and multiple regression, that require calculations so complex that they realistically cannot be done without computer software.

Role It should be emphasized that this manual/workbook is designed to be a supplement to the Triola series of statistics textbooks; it is not designed to be a self-contained statistics textbook. It is assumed throughout this manual/workbook that the theory, assumptions, and procedures of statistics are described in the textbook that is used.

Format Chapter 1 of this supplement describes some important basics for using Statdisk. Chapters 2-14 in the manual/workbook correspond to Chapters 2-14 in *Elementary Statistics*, Twelfth Edition. However, individual chapter *sections* in this manual/workbook generally *do not* match the sections of the textbook. Each chapter includes a description of the Statdisk procedures relevant to the corresponding chapter in the textbook. The cross referencing makes it easy to use this supplement with the textbook. Chapters include a beginning section in which examples are illustrated with Statdisk. It would be helpful to follow the steps shown in these sections so the basic procedures will become familiar. You can compare your own computer display to the display given in this supplement and then verify that your procedure works correctly. You can then proceed to conduct the experiments that follow.

Data Sets Statdisk includes all data sets found in Appendix B of the textbook. All data sets referenced in this manual/workbook can be opened in Statdisk by selecting **Data Sets** on the top menu bar and then clicking on **Elementary Statistics 12th Edition**.

Thanks I thank Bill Flynn for the original STATDISK algorithms. I thank Russell F. Loane and Timothy C. Armstrong for their outstanding work on a previous version. I thank Justine Baker of Peirce College for contributing several Activities with STATDISK. The following beta testers have been extremely helpful: Gary Turner, Richard Dugan, Justine Baker, Robert Jackson, Caren McClure, Sr. Eileen Murphy, John Reeder, Carolyn Renier, Cheryl Slayden, Victor Strano, Henry Feldman, and others who we know only by their e-mail addresses. For this new version of STATDISK, I am very thankful to Marc Triola, MD, who had the talent and patience to completely rewrite thousands of lines of programming code, as well as create new features and updates. I extend special thanks to Scott Triola who was so instrumental in the creation of this new edition of the STATDISK Manual/Workbook. It is wonderful working with such competent and skilled professionals. Their dedication and talent are very apparent in this new version of STATDISK. Finally, I thank the Pearson Addison-Wesley staff for their enthusiastic support in this project. It is a genuine pleasure working with a publishing company committed to providing a product with the highest quality. I also thank the many instructors and students who took the time to provide many valuable suggestions.

Mario F. Triola
January, 2013

Contents

Click on chapter title to view

1 Statdisk Fundamentals

Statdisk

Statdisk is designed so that it uses many of the same features found in a wide variety of software applications, so tools introduced in this chapter have a universal usefulness that extends beyond Statdisk and statistics. For example, the **Copy** and **Paste** features of Statdisk are commonly included with many software programs. As you work with such features, you acquire or reinforce important and general computer skills that will help you with other applications.

1-1 Installing and Updating Statdisk

Installing Statdisk: Statdisk is available as a free download on www.statdisk.org for Windows XP/7/8 and Mac OS X. Installation instructions are provided on this website. In some cases, your instructor will have Statdisk installed on your college network.

Updating Statdisk: If using Statdisk on a computer with an Internet connection, you will be prompted when an update is available. You can also click on **Help** and select **Check for Updates**.

Statdisk automatically checks for updates by default. To enable/disable automatic updates follow these steps:

1. Click on **Edit** in the top menu bar of Statdisk and select **Preferences** from the dropdown menu.

2. On the **General** tab check the box labeled "*Automatically check for Statdisk updates*" to enable automatic updates. Uncheck the box to disable automatic updates.

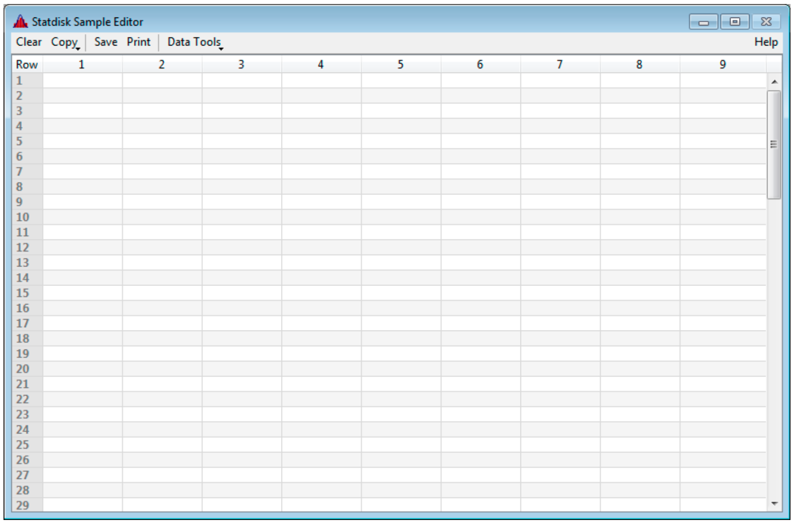
3. Click the **Check now** button to immediately check for a Statdisk update.

1-2 Entering Data

Your first use of Statdisk is likely to occur with topics from Chapter 2 of your Triola statistics textbook, and one of your first objectives is likely to be entering a set of sample data. (The data sets found in Appendix B of the Triola textbook are already stored with Statdisk, and they can be retrieved as described in Section 1-4 of this manual/workbook. It is not necessary to manually enter those data sets that are already included with Statdisk.) To manually enter a set of sample data, use the following procedure.

Statdisk Procedure for Entering Data

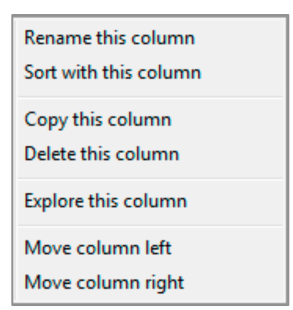
1. After opening Statdisk, you should see the Sample Editor window shown.



2. Click on the cell in row 1 of column 1 and type in your first data value, and then press the **Enter** key. Next, type the second data value and press the **Enter** key again. Continue to enter all of your sample values. You can use your mouse or the keyboard arrow keys to navigate to and select individual cells in the Sample Editor (press **ESC** to navigate using the keyboard arrows if you are editing a cell) .
   Note: If you see that you have made a mistake, simply click on the wrong value and make the correction.

There are a few other important features that are available by right clicking on the **Column Name**. If you right click on a column name (e.g. column "*1*") you can select from the following options:

Rename this column: Allows you to enter or modify the *names* of the columns of data



Sort with this column: Allows you to sort data in one or all columns in ascending or descending order. See Section 2-6 for detailed sorting instructions.

Copy this column: Copies the data contained in the column into the clipboard.

Delete this column: Deletes the column of data.

Explore this column: Allows you to obtain one screen that displays key statistics and graphs for data in a specific column.

Move column left/right: Moves all data in the column one column to the left/right. If there is data in the adjacent column, this data will be moved into the currently selected column.

1-3 Saving Data

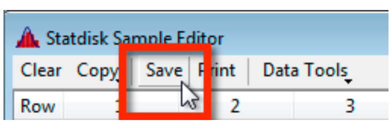
After entering a set of data as described above, you can save it for the future by using the following procedure.

Statdisk Procedure for Saving Data

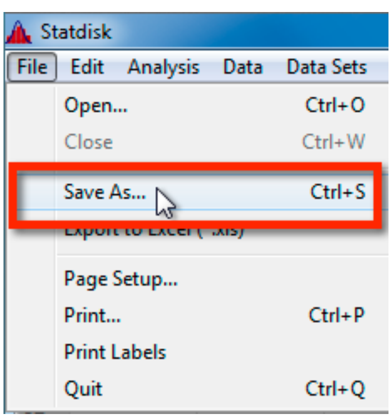
1. After entering all of the values in the Sample Editor window, use the mouse to click on **Save** located in the Statdisk Sample Editor menu.
   or
   Click on **File** in the top menu and select the item **Save As....** in the dropdown.

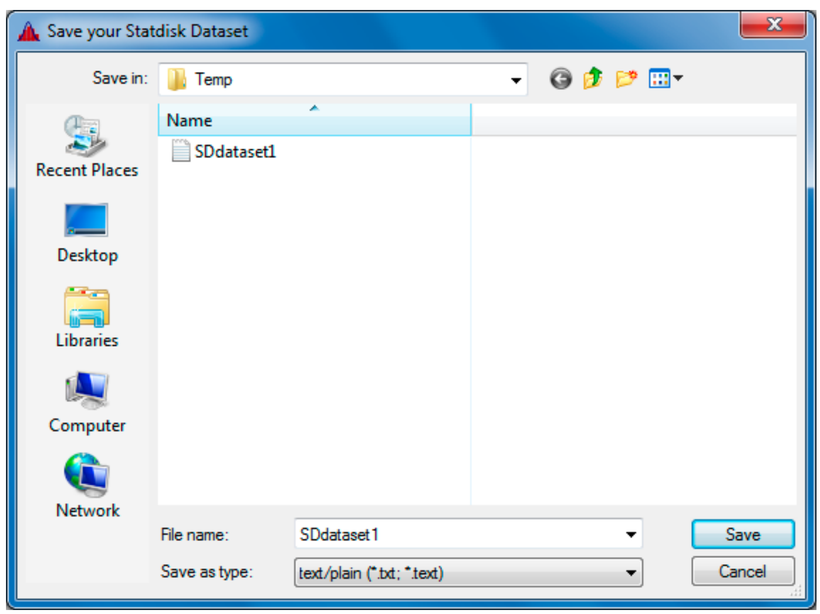
2. You will now see a dialog box with the title of "*Save your Statdisk Dataset*." See the *Save in* box for the location to be used for storing the data set. *Important:* If you want to save your data set in a location different from the default location already shown, you can change the drive and folder as you desire. For example, if you want to save a data set named *"Heights"* in a folder named *"Temp"*, do the following:

   a. Enter "*Heights*" in the box labeled "*File name*."
   b. Click on the "*Save in*" box and select the location to be used for saving the data set.



or





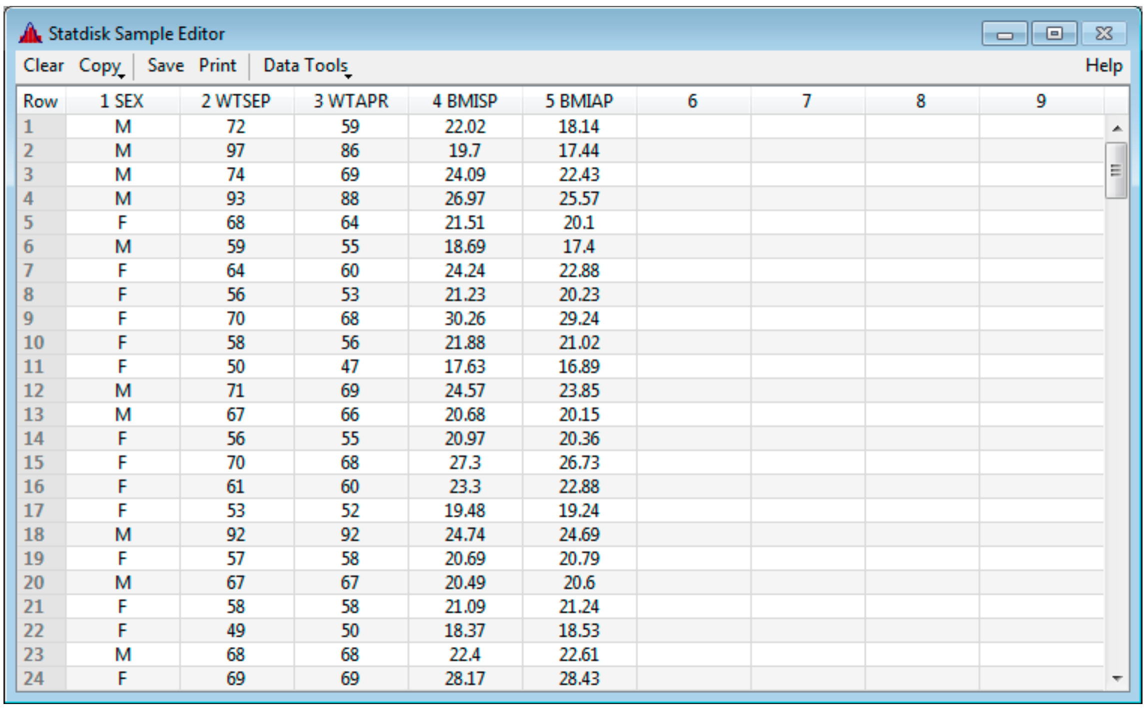
1-4 Retrieving Data

Appendix B Data sets: Statdisk includes the Appendix B data sets from the Triola Statistics Series. Many Exercises in the textbook require that you use these data sets. To retrieve one of the stored Appendix B data sets, follow these steps.

1. Click on the top menu item of **Data Sets**.

2. Select the Triola textbook that you are using.

3. Click on the name of the desired data set to open. For example, to select the Freshman 15 data set, scroll down to the name of **4 - Freshman 15 Study Data**, and then click on that name. The data set will be inserted in the Sample Editor window, as shown below. The full data has 67 rows, and you can see the additional rows by scrolling down.



Row	1 SEX	2 WTSEP	3 WTAPR	4 BMISP	5 BMIAP	6	7	8	9
1	M	72	59	22.02	18.14				
2	M	97	86	19.7	17.44				
3	M	74	69	24.09	22.43				
4	M	93	88	26.97	25.57				
5	F	68	64	21.51	20.1				
6	M	59	55	18.69	17.4				
7	F	64	60	24.24	22.88				
8	F	56	53	21.23	20.23				
9	F	70	68	30.26	29.24				
10	F	58	56	21.88	21.02				
11	F	50	47	17.63	16.89				
12	M	71	69	24.57	23.85				
13	M	67	66	20.68	20.15				
14	F	56	55	20.97	20.36				
15	F	70	68	27.3	26.73				
16	F	61	60	23.3	22.88				
17	F	53	52	19.48	19.24				
18	M	92	92	24.74	24.69				
19	F	57	58	20.69	20.79				
20	M	67	67	20.49	20.6				
21	F	58	58	21.09	21.24				
22	F	49	50	18.37	18.53				
23	M	68	68	22.4	22.61				
24	F	69	69	28.17	28.43				

Retrieving Your Own Dataset: If you want to open a data set that you have previously saved, click on **File**, select the menu item of **Open...**, and select the location of the data set. Click on the name of the data set and then click on **Open**.

1-5 Copy and Paste Data

The Copy and Paste feature is used in many different software applications, including word processors and spreadsheets. You should clearly understand the following.

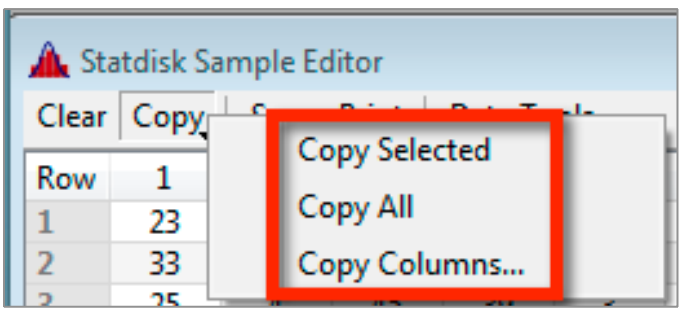
- **After entering or retrieving a data set and using the *Copy* command, the data set will remain available for use in the "clipboard" until you use *Copy* for a new data set, or until you exit the program.**

- **After using the *Copy* command, go to the program where you want to use the data set, then click on the *Paste* button in the window for that application.**

See Section 1-9 for procedures allowing you to copy data sets between Statdisk and other applications, such as Minitab or Excel or Microsoft Word. The procedure below illustrates the usefulness of the Copy/Paste feature within Statdisk.

Statdisk Procedure for Using Copy and Paste

There are three ways to copy data and all are easily accessed by clicking on **Copy** on the Statdisk Sample Editor menu bar and selecting the option you prefer (described below). Note that only data can be copied in Statdisk, column names cannot be copied.



Copy Selected – Copies only the data you select.

1. Click on the first cell you want to copy, keep the mouse/trackpad button depressed and drag the pointer to select all of the data you want to copy. The selected data will be highlighted.

2. Select **Copy** from the Sample Editor bar and then select **Copy Selected**. The data selection has now been copied.

3. To paste the copied data, click on the first cell in the column where you want to insert the data. Select **Edit** from the top menu bar and then select **Paste**.

Copy All – Copies all the data in the Sample Editor.

1. Select **Copy** from the Sample Editor bar and then select **Copy All**. All data in the Sample Editor has now been copied.

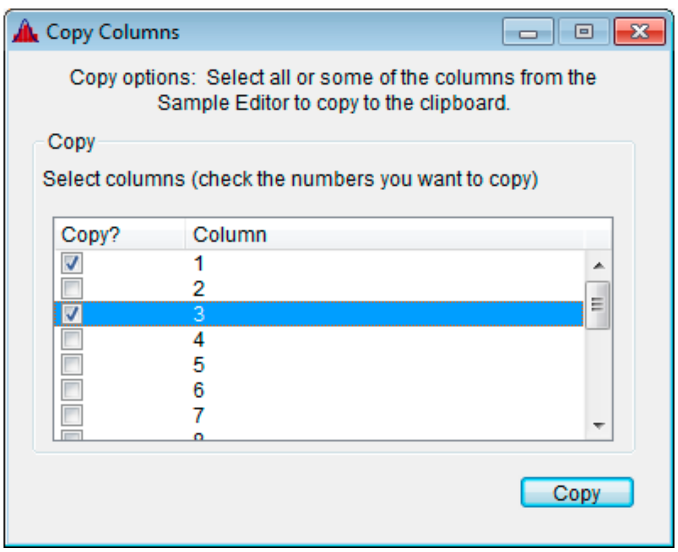
2. To paste the copied data, click on the first cell in the column where you want to insert the data. Select **Edit** from the top menu bar and then select **Paste**.

Copy Columns… - Copies data in selected columns.

1. Select **Copy** from the Sample Editor bar and then select **Copy Columns**. The Copy Columns dialog will appear as shown below.

2. Select the column(s) of data you want to copy by clicking on the "*Copy?*" box. A checkmark indicates the column will be copied. You can select multiple columns of data.

3. To paste the copied data, click on the first cell in the column where you want to insert the data. Select **Edit** from the top menu bar and then select **Paste**.



Here are some options for using the copied columns of data within Statdisk:

Isolate desired columns of data: Clear the Sample Editor window (select **Edit** from the top menu and then click on **Clear Data**), then click the **Paste** button to insert only those columns that were previously copied.

Combine desired columns of data: Open another data set in the Sample Editor window, then click the **Paste** button to insert the copied columns in the column(s) you select. For example, suppose that you want to combine columns of data from different data sets, such as the weights of females and the weights of males. Those weights are found in two different data sets: *Body Measurements Female* and *Body Measurements Male*. The Copy/Paste feature can be used to combine the weights from those two data sets. Simply copy the weights from one of the data sets and paste it into the other data set, so the weights of females and males are included in the same Sample Editor window. Column names are not copied with the data, so you may want to change the column name after you paste the data.

1-6 Editing and Transforming Data

It is easy to edit a data set in the Sample Editor window.

- *Delete* an entry by clicking on the cell and using the **Delete** key to remove it.
- *Insert* an entry by typing it into the first empty cell in the desired column.
- *Move* columns left or right by right clicking on the column name and selecting **Move column left/right**.
- *Sort* data in ascending or descending order by clicking on **Data Tools** in the Sample Editor menu bar and selecting **Sort Data**. See Section 2-6 for detailed sorting instructions.

Data may also be *transformed* with operations such as adding a constant, multiplying by a constant, or using the functions of adding, subtracting, multiplying, dividing, raising to a power, or stripping away the decimal part of data values. For example, if you have a data set consisting of temperatures on the Fahrenheit scale (such as the *Body Temperature* data set in Appendix B of the textbook) and you want to transform the values to the Celsius scale, you can use the equation

$$C = \frac{5}{9}(F - 32)$$

Statdisk Procedure for Transforming Data

1. First enter the data in one or more columns of the Sample Editor window.

2. Click on the top menu item of **Data**.

3. Click on **Sample Transformations** to get a window like the one shown.

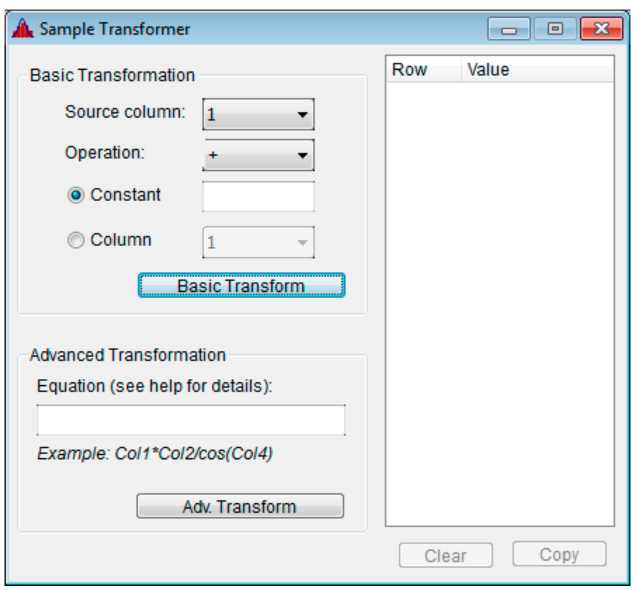
Basic Transformation

4. For the *Source column*, select the column of data to be transformed.

5. For *Operation*, select the desired operation from the list of available options.

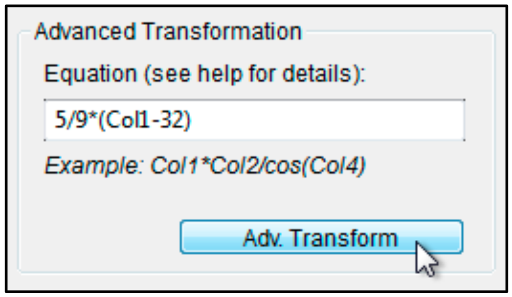
6. Select "*Constant*" to perform the operation using the same value for all rows, or select "*Column*" to perform the operation with corresponding row values in two different columns (such as adding two columns).

7. Click on **Basic Transform** and the transformed values will appear in the column at the extreme right. The result can be copied to the Sample Editor window.



Advanced Transformation

The "*Advanced Transformation*" feature can be used for more advanced transformations, such as those involving absolute values, logarithms, or the sine function. For example, to convert to Fahrenheit to Celsius, enter the following equation in the **Equation** box and then click the **Adv. Transform** button.



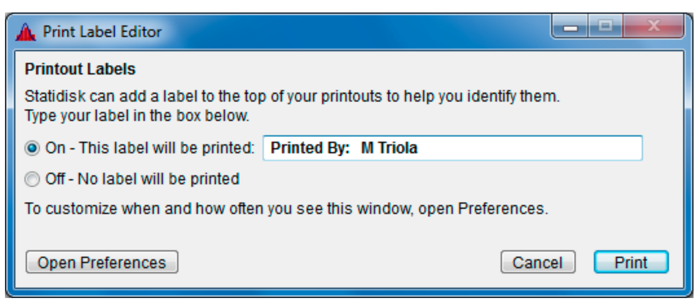
For more details of the procedure for transforming data, click on the **Help** button.

1-7 Printing Screens and Data Sets

After you have obtained results from Statdisk, such as a graph or a listing of statistics, you can print those results. Simply select **File** from the top menu and then select **Print**. (You could also click on the **Copy** button to copy the results that can then be pasted into some other application, such as a word processing document you are using for writing a report.)

Including a Name With the Printout

Statdisk allows you to enter a label or name so that printed results can be identified when there are multiple students printing at the same time in a computer lab. When clicking on the Print button for the first time in a session, you will see the **Print Label Editor** dialog shown.



If you want to have your name included with printed results, click the **On** button and enter your name. The time and date can also be included or excluded by clicking on **Open Preferences,** then clicking on the **Printing** tab, and checking the top box to include the time and date.

The Print Label Editor dialog only appears the first time you click on the **Print** button in each session. You can access the Print Label Editor at any time by selecting **File** form the top menu and then selecting **Print Labels** and clicking the **Printing** tab.

Statdisk will print all of the data, not just those visible on the screen. *Caution:* If you have a large data set that you would like to print, it is recommended that you move it to a word processor, such as MS Word, where you can reconfigure the data for more efficient printing. Use the following procedure.

Printing a Large Dataset in a Word Processor

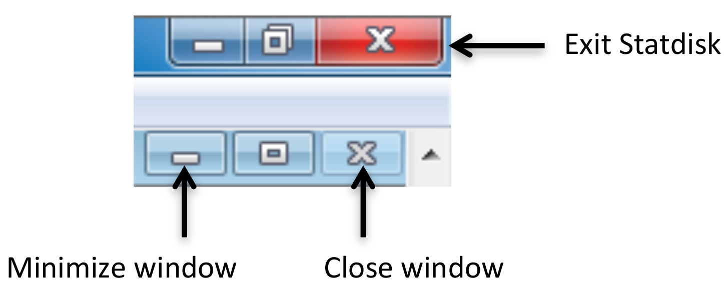
1. With the data set listed in the Sample Editor window, select all of the data and click on the **Edit - Copy** button.

2. In your word processor, click on **Edit**, then **Paste**. The entire list of values will be inserted into your document where you can configure and reformat them as you please. Instead of printing a data set of 1000 values with one column covering many pages, use many values in each row so that fewer pages will be printed. Environmentalists will be very thankful.

1-8 Closing Windows and Exiting

Closing Windows: To minimize clutter, *close* windows after they are no longer needed. Close windows by clicking on the small box labeled "*x*" that is located in the upper right corner of the Statdisk window. (Clicking on the box with the symbol "**-**" will cause the window to be minimized / hidden, but it continues to remain open and available for recall.)

Exiting Statdisk: Had enough for now? To exit or quit the Statdisk program, click on the × located in the extreme upper right corner. Another way to exit Statdisk is to click on **File**, then click on **Quit**.



1-9 Exchanging Data with Other Applications

There may be times when you want to move data from Statdisk to another application (such as Excel or Minitab or Word) or to move data from another application to Statdisk. Instead of manually retyping all of the data values, you can usually transfer the data set directly. Given below are two ways to accomplish this.

Method 1: Use Copy and Paste

Using the methods described in Section 1-5 of this manual/workbook, use **Copy** and **Paste** to copy the columns of data, then paste them directly into the other application.

Method 2: Use Text Files

1. In the software program containing the original set of data, create a text file of the data. (Statdisk files are already saved as text files.)

2. Statdisk and most other major applications allow you to import the text file that was created. (To import a text file into Statdisk, select **File**, then **Open**. Enter the location of the text file in the "*Look in*" box, then click the **Open** button. Use the *Open Saved Data* dialog options and preview window to ensure the data is correctly imported into Statdisk.)

CHAPTER 1 EXPERIMENTS: Statdisk Fundamentals

1-1. ***Entering Sample Data*** When first experimenting with procedures for using Statdisk, it's a good strategy to use a small data set instead of one that is large. If a small data set is lost, you can easily enter it a second time. In this experiment, we will enter a small data set, save it, retrieve it, and print it. *Dataset 3: Body Temperatures of Healthy Adults* in Appendix B of the Triola textbook includes these body temperatures, along with others:

98.6 98.6 98.0 98.0 99.0

a. Open Statdisk and enter the above sample temperatures. (See the procedure described in Section 1-2 of this manual/workbook.)
b. Save the data set using the file name of TEMP. See the procedure described in Section 1-3 of this manual/ workbook.
c. Print the data set.
d. Exit Statdisk, then reload it and retrieve the file named TEMP. Save another copy of the same data set using the file name of TEMP2. Print TEMP2 and include the title at the top. (That is, use a title of TEMP2 instead of the default.)

1-2. ***Retrieving Data*** *Dataset 7: Bear Measurements* is already stored in Statdisk. It contains nine columns with 54 values in each column. Open this data set (see Section 1-2 of this manual/workbook) and print the data.

1-3. ***Using Copy/Paste*** Enter the sample values 1, 2, 3, 10, 20 and use **Data/Sample Transformations** to add 5 to each value, then use Copy/Paste to copy the results to the second column of the Sample Editor window. Then click **Data**, select **Descriptive Statistics**, select column 2, and click **Evaluate** and obtain a printed copy of the resulting screen display. The resulting statistics will be described in the textbook.

1-4. ***Using Copy/Paste*** The data sets *Body Measurements – Males* and the data set *Body Measurements - Females* are already stored in Statdisk. Use Copy/Paste to create a Sample Editor window that includes only these two columns: (1) heights ("*ht*") of the males and (2) heights ("*ht*") of the females. Obtain a printed copy of the Sample Editor window.

1-5. ***Editing Data*** *Dataset 7: Bears* includes measurements from 54 anesthetized wild bears.
a. Open the data set Bears, then find the value of the *mean* of the weights by selecting **Data,** then **Descriptive Statistics.** Enter the mean weight. ________
b. Go back to Sample Editor window and change the weight of 34 lb to 3400 lb. Repeat part (a) and record the new value of the mean.__________
c. Did the mean change much when 34 lb was changed to 3400 lb?

____

1-6. ***Generating Random Data*** In addition to entering or retrieving data, Statdisk can also *generate* data sets. In this experiment, we will use Statdisk to simulate the rolling of a pair of dice 500 times. Select **Data** from the top menu bar, then select **Dice Generator**. For the sample size, enter 500 (for 500 rolls), enter 2 for the number of dice, enter 6 for the number of sides, then click on **Generate**. Examine the displayed totals and count the number of times that 7 occurs. Record the result here: __________

1-7. ***Transforming Data*** Experiment 1-1 results in saving a sample of body temperatures in degrees Fahrenheit. Retrieve that data set, then proceed to transform the temperatures to the Celsius scale. (See Section 1-6 in this manual/workbook.) After obtaining the Celsius temperatures, select **Data**, then select **Descriptive Statistics** and proceed to find the value of the *mean*. Enter the mean here:_______________

1-8. ***Retrieving and Transforming Data*** Open the Statdisk *Dataset 7: Bears*, which includes the weights (in pounds) of a sample of bears. To convert the weights to kilograms, multiply them by 0.4536. Use Statdisk to convert the weights from pounds to kilograms. In the space below, write the weights (in kilograms) of the first five bears.

____

2 Summarizing and Graphing Data

Statdisk

Important note: The topics of this chapter require that you use Statdisk to enter data, retrieve data, save files, and print results. These functions are covered in Chapter 1 of this workbook. Be sure to understand these functions before beginning this chapter.

Section 2-2 in the textbook describes the construction of a table representing the *frequency distribution* for a set of data. Shown below is Table 2-1 from the textbook which contains the full IQ scores of subjects from the low lead group. (The listed IQ scores are part of Data Set 5 in Appendix B of the textbook, and the data are included in the Statdisk data set *Elementary Statistics 12th Edition – Data Set 5: IQ and Lead Exposure*. The "*Low Lead*" group values are in contained in rows 1-78 of the data set and are designated by the value "*1*" in the "*1 Lead*" column.)

Table 2-1 Full IQ Scores of Low Lead Group and High Lead Group

Low Lead Level (Group 1)															
70	85	86	76	84	96	94	56	115	97	77	128	99	80	118	86
141	88	96	96	107	86	80	107	101	91	125	96	99	99	115	106
105	96	50	99	85	88	120	93	87	98	78	100	105	87	94	89
80	111	104	85	94	75	73	76	107	88	89	96	72	97	76	107
104	85	76	95	86	89	76	96	101	108	102	77	74	92		

Chapter 2 of the textbook begins with the construction of frequency distributions, followed by the construction of histograms. In this manual/workbook, we begin with histograms, and then we see how to obtain frequency distributions from the histograms.

2-1 Histograms

Statdisk can be used to automatically generate a histogram. The basic approach is to get the data listed in the Sample Editor window, and then use the *Histogram* module to generate the histogram. When using Statdisk's *Histogram* function, you have the option of simply accepting the default settings, or you can select your own desired class width and starting point. If you choose to set your own limits, you must understand the definition of *class width*. In the textbook, we define class width as follows:

> ***Class width*** is the difference between two consecutive lower class limits or two consecutive lower class boundaries.

As an example, see Table 2-2 which is a frequency summarizing the listed IQ scores from Table 2-1. The *class width is 20* (the difference between the consecutive lower class limits of 50 and 70).

Statdisk

Table 2-2
IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1

Procedure for Generating a Histogram

1. Enter or retrieve a set of sample data using one of these procedures:

    - **Manual entry of data:** Values can be entered in the Sample Editor window.
    - **Retrieve a data set from those included in Appendix B:** Click on the top menu item of **data sets** and proceed to select one of the listed data sets.
    - **Retrieve a data set that you created:** Use **File** - **Open** as described in Section 1-4 of this manual/workbook.

2. Click on **Data** in the main menu bar at the top.

3. Click on **Histogram**.

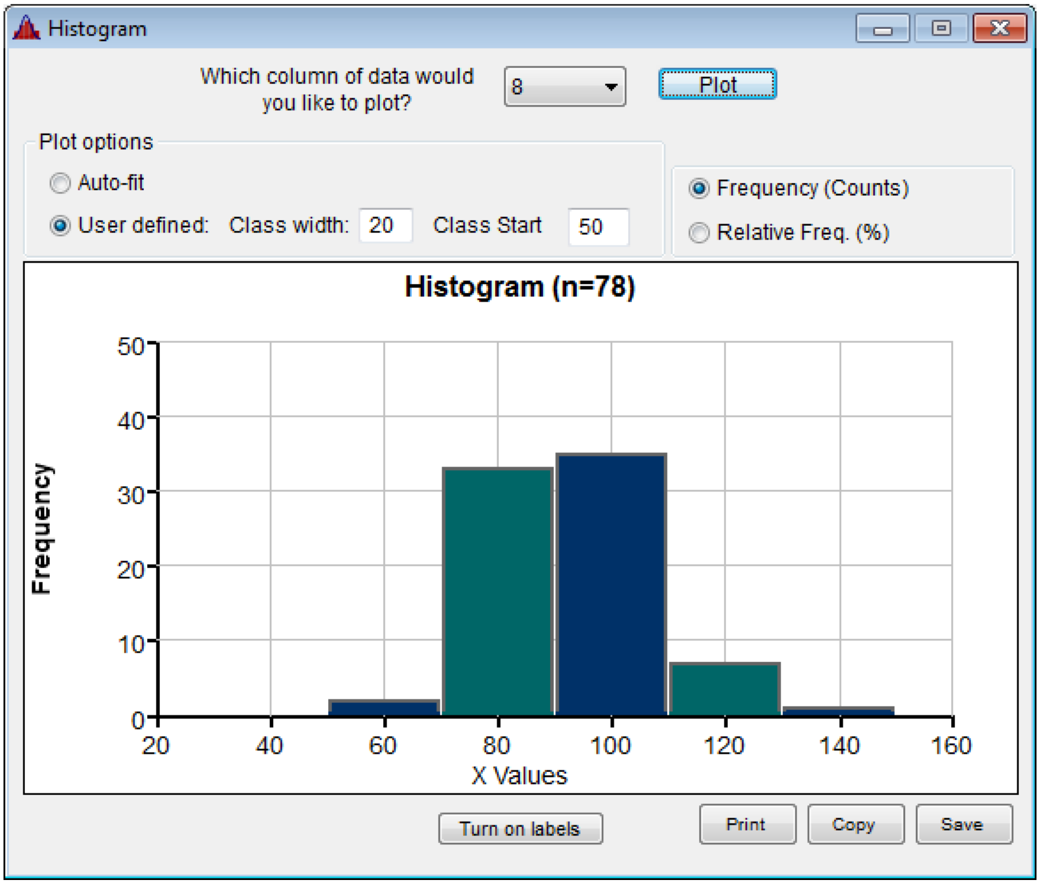
4. In the Histogram window, first select the column to be used. The default is column 1, and it can be changed to any column number.

5. **Using the default settings**: Click **Plot** to allow Statdisk to automatically generate a histogram using default settings.

    Using your own settings: Click the **User Defined** button and proceed to enter your desired class width and the starting value of the first class.

6. Click on **Plot**.

As an example, see the Statdisk display on the following page. The histogram is constructed using the list of IQ scores from Table 2-1. Instead of using the default settings, we clicked on the **User defined** button and entered 20 for the class width and 50 as the starting point. These entries correspond to the frequency table shown in Table 2-2. (See the preceding frequency distribution table and verify that the class width is 20 and the lower limit of the first class is 50.)



The histogram gives us insight into the nature of the *distribution*. In later chapters, we must often determine whether sample data appear to come from a population with a normal distribution. For now, we can consider a normal distribution to be a distribution with a histogram that is roughly bell–shaped. Simply examine the histogram and make a judgment about whether it appears to be approximately bell-shaped.

If we examine the Statdisk histogram shown above, we can see that the distribution does appear to be bell–shaped, so that requirement of a normal distribution does appear to be satisfied for this data set. (Statdisk includes a feature of *Assessing Normality* that provides more information. That feature is discussed in Chapter 6 of this manual/workbook.)

2-2 Frequency Distributions

Statdisk does not include a specific menu item for generating a frequency distribution from a list of data, but frequency distributions can be obtained by using Statdisk's ability to generate histograms. If you want to use Statdisk to construct a frequency distribution, enter your own starting point and class width (based on the range of values and the minimum value).

See the above Statdisk *Histogram* display at the end of Section 2-1 in this manual/workbook and note that there is a **Turn on labels** button. If you click that button, you get the labels consisting of frequency counts above the bars of the histogram as shown below. Knowing that the class width is 20 and the lower limit of the first class is 50, we can see that the first class is 50-69 and it has a frequency of 2. Similarly, the next class is 70-89 and it has a frequency of 33. The third class of 90-109 has a frequency of 35, and so on. This information can be listed in a frequency distribution as shown below in Table 2-2 from the textbook.

Technical note: Statdisk is designed so that a sample value falls into a particular class if it is equal to or greater than the lower class limit and less than the upper class limit.

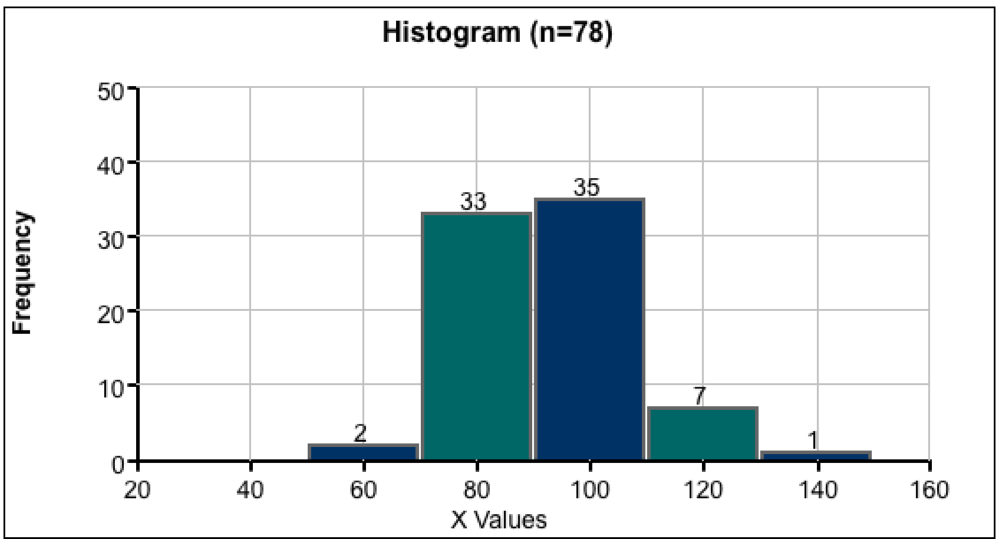


Table 2-2
IQ Scores of Low Lead Group

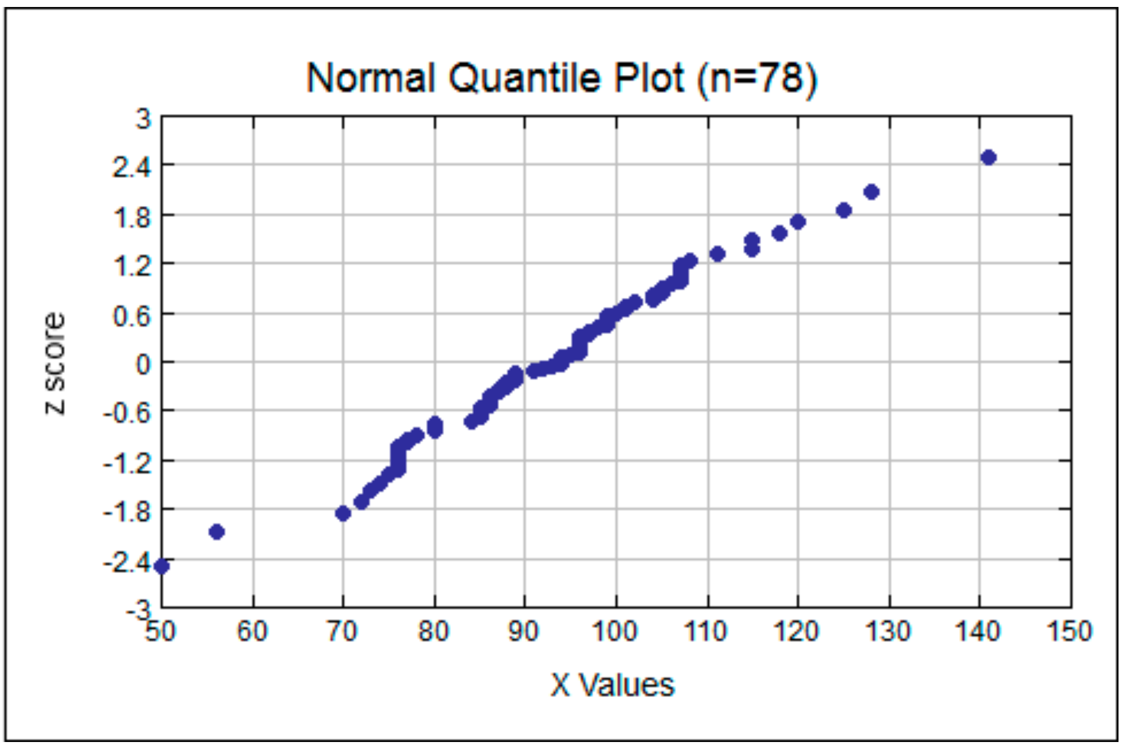
IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1

Statdisk

2-3 Normal Quantile Plots

The textbook points out that *normal quantile plots* are helpful in determining whether sample data appear to be from a population having a normal distribution. Section 6-6 in the textbook discusses methods for *assessing normality*, and Section 2-3 in the textbook includes a brief discussion of normal quantile plots. Statdisk generates normal quantile plots like the one shown here. This normal quantile plot was generated using the same IQ score data used to generate the histogram and frequency distributions earlier in this chapter (Statdisk data set *Elementary Statistics 12th Edition – "Data Set 5: IQ and Lead Exposure"*.)

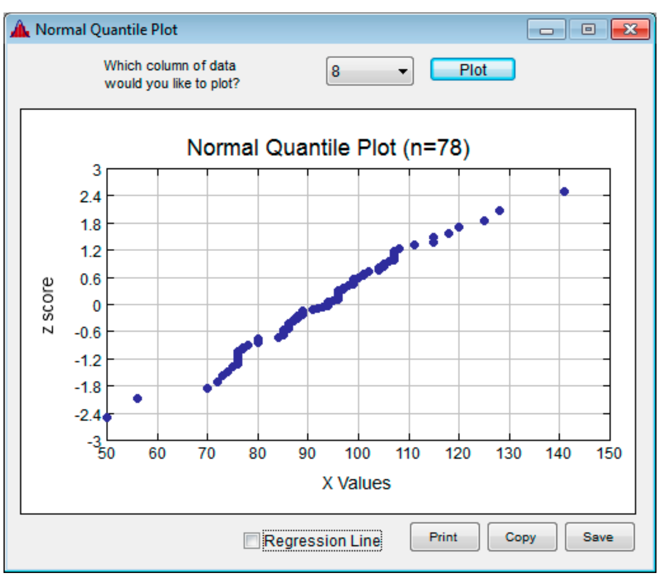
In later chapters, we must often determine whether sample data appear to come from a population with a normal distribution. For now, we can consider a normal distribution to be a normal quantile plot that shows a pattern of points that is reasonably close to a straight line pattern. Simply examine the normal quantile plot and make a judgment about whether it appears to be approximately a straight line pattern.



Procedure for Generating Normal Quantile Plots

1. Enter or retrieve a set of sample data using one of these procedures:
   - **Manual entry of data:** Values can be entered in the Sample Editor window.
   - **Retrieve a data set from those included in Appendix B:** Click on the main menu item of **data sets** and proceed to select one of the listed data sets.
   - **Retrieve a data set that you created:** Use **File** - **Open** as described in Section 1-4 of this manual/workbook.
2. Click on **Data** in the main menu bar at the top.
3. Click on **Normal Quantile Plot**.
4. Select the data column to be used for the normal quantile plot. The default columns can be changed as desired.
   a. If the "Regression line" box is left checked, the graph will include a straight line that best fits the points. In the display shown below, the box is not checked so the line is not included.
5. Click on the **Plot** button.

As an example, see the following Statdisk display. The histogram is constructed using the list of IQ scores from Table 2-1 earlier in this chapter.

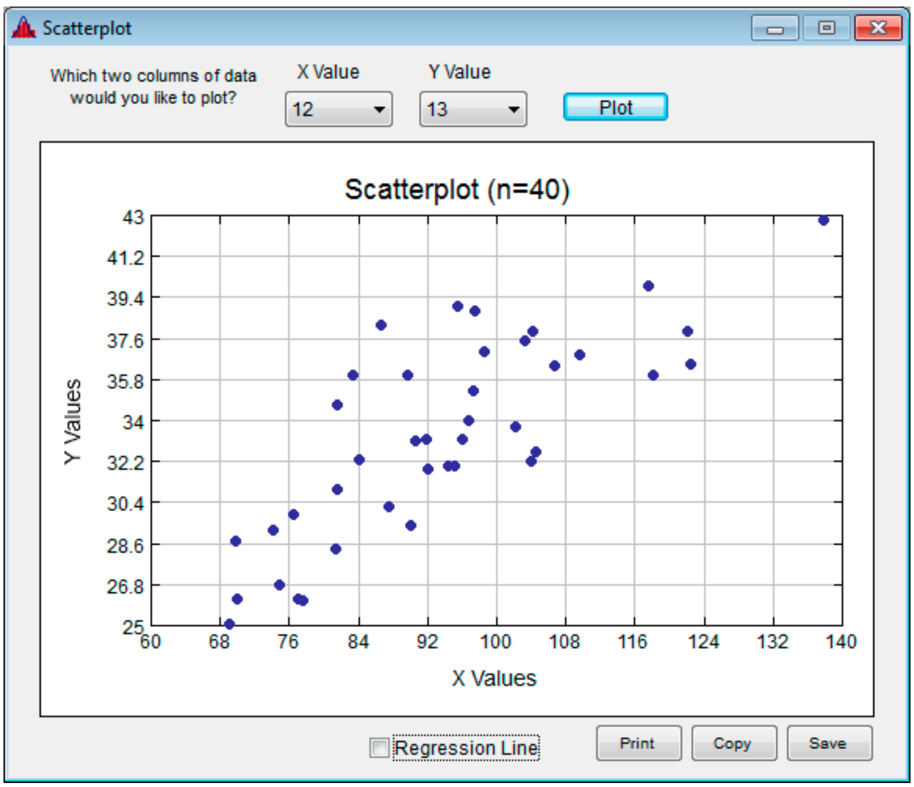


Statdisk

2-4 Scatterplots

The textbook describes a scatterplot (or scatter diagram) as a plot of paired (*x, y*) data with a horizontal *x*-axis and a vertical *y*-axis. The data are paired in a way that matches each value from one data set with a corresponding value from a second data set. A scatterplot can be very helpful in seeing a relationship between two variables.

The Statdisk scatterplot shown below results from paired data consisting of waist circumference (cm) and arm circumferences (cm) of randomly selected mails. This data is in the Statdisk data set *Elementary Statistics – Data Set 1: Body Measurements Male*. This scatterplot shows that as the waist circumference (*x* axis) increase, the corresponding arm circumference (*y* axis) tends to be higher.



Procedure for Generating a Scatterplot

To use Statdisk for generating a scatterplot, you must have a collection of *paired* data listed in the Sample Editor window.

1. Enter or retrieve a set of sample data using one of these procedures:

    - **Manual entry of data:** Values can be entered in the Sample Editor window.

- **Retrieve a data set from those included in Appendix B:** Click on the main menu item of **Data Sets** and proceed to select one of the listed data sets.

- **Retrieve a data set that you created:** Use **File** - **Open** as described in Section 1-4 of this manual/workbook.

2. Click on **Data** in the main menu bar at the top.

3. Click on **Scatterplot**.

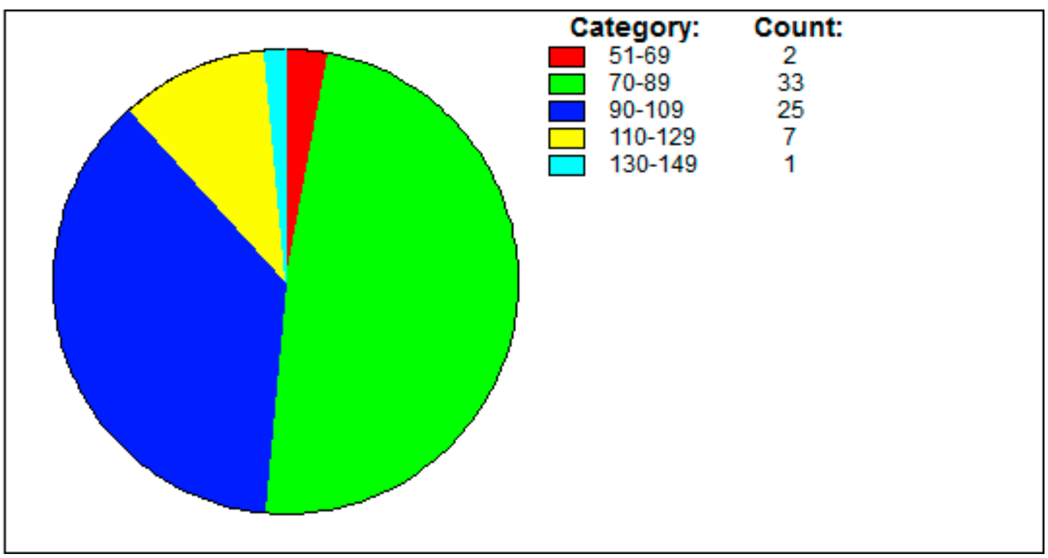
4. Select the two columns to be used for the scatterplot. The default columns can be changed as desired.

5. If the "Regression line" box is left checked, the graph will include a straight line that best fits the points. In the Scatterplot display shown in this section, the box is not checked so the line is not included.

6. Click on the **Plot** button.

2-5 Pie Charts

The textbook makes the point that pie charts are generally poor as effective graphs for depicting data. Statdisk does, however, include an option for generating pie charts. Shown here is a pie chart showing the frequency distribution used in section 2-2.



Procedure for Generating a Pie Chart

1. Begin by entering the category names as text in column 1 of the Statdisk data window.

2. Next, enter the corresponding frequency counts in column 2.

3. Click on **Data** from the main menu at the top,

4. Click on **Pie Chart**.

5. Click on the **Chart** button and the pie chart will be displayed.

2-6 Sorting Data

To *sort* data is to arrange them in order. There are several cases in which it becomes necessary to rearrange a data set so that the values are in order, ascending from low to high or vice versa. Statdisk 12 features a new sorting tool to make sorting data fast and easy.

Procedure for Sorting Data

1. Enter or retrieve a set of sample data so that the sample values are listed in the Sample Editor window.

2. Click on **Data** from the main menu at the top,

3. Click on **Sort Data** and the following toolbar will appear above the Sample Editor:

4. By selecting **Sort - All Columns**, you can sort a single column while rearranging the other columns so that data in the same row will stay in the same row.

   By selecting **Sort - One column**, you can sort a single column while leaving the other columns unchanged.

5. Select which column you want to sort by.

6. Select the order in which you want to sort. **A to Z** will sort data in ascending order, **Z to A** will sort data in descending order.

7. Click **Sort**.

Using Sort to Identify Outliers The sort feature is useful for identifying outliers. When analyzing data, it is important to identify outliers because they can have a dramatic effect on many results. It is usually difficult to recognize an extreme value when it is buried in the middle of a long list of values arranged in a random order, but *outliers become much easier to recognize with sorted data, because they will be found either at the beginning or end*. To identify outliers, simply sort the data, and then examine the lowest and highest values to determine whether they are dramatically far from almost all of the other sample values.

CHAPTER 2 EXPERIMENTS: Graphing Data

Histograms. *In Experiments 2-1 through 2-8, use the Elementary Statistics 12th Edition data sets included in Statdisk to construct a histogram. These data sets are the same data sets included in Appendix B of the textbook. It is not necessary to use a specific class width or class boundaries.*

2-1. **Pulse Rates of Males** Use the pulse rates of males contained in *Data Set 1: Body Measurements Male*. Construct a histogram. Does the histogram appear to depict data having a normal distribution? Why or why not?

2-2. **Pulse Rates of Females** Use the pulse rates of females listed in *Data Set 1: Body Measurements Female*. Construct a histogram. Does the histogram appear to depict data having a normal distribution? Why or why not?

2-3. **Earthquake Magnitudes** Use the earthquake magnitudes listed in *Data Set 16: Earthquake Measurements*. Construct a histogram. Using a loose interpretation of the requirements for a normal distribution, do the magnitudes appear to be normally distributed? Why or why not?

2-4. **Earthquake Depths** Use the earthquake depths listed in *Data Set 16: Earthquake Measurements*. Construct a histogram. Using a loose interpretation of the requirements for a normal distribution, do the depths appear to be normally distributed? Why or why not?

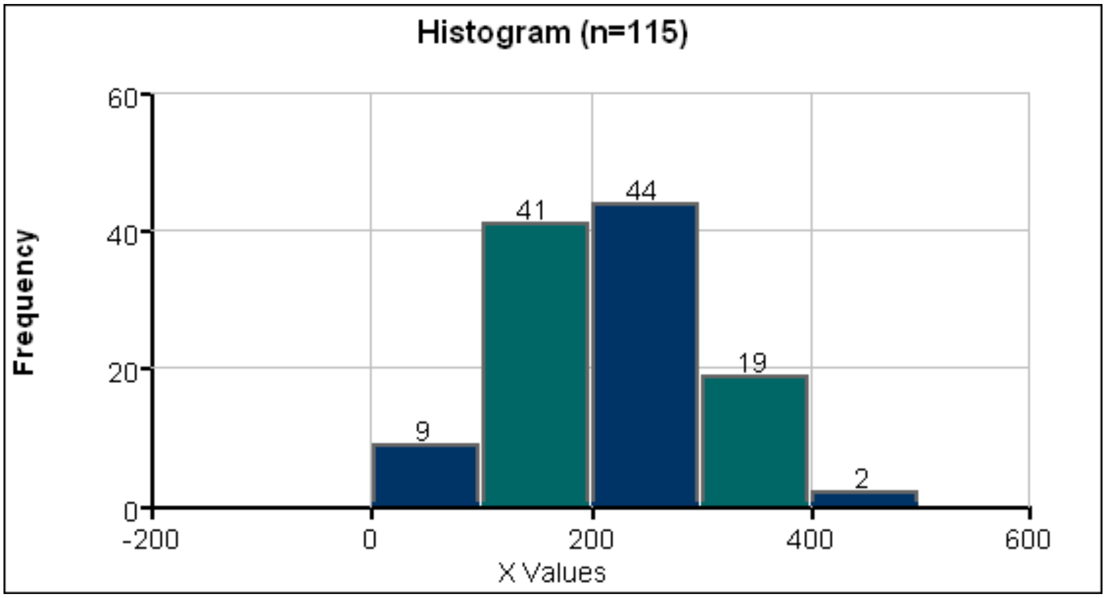
2-5. **Male Red Blood Cell Counts** Use the red blood cell counts of males listed in *Data Set 1: Body Measurements Male*. Construct a histogram. Using a very loose interpretation of the requirements for a normal distribution, do the red blood cell counts appear to be normally distributed? Why or why not?

2-6. **Female Red Blood Cell Counts** Use the red blood cell counts of females listed in *Data Set 1: Body Measurements Female.* Construct a histogram. Using a very loose interpretation of the requirements for a normal distribution, do the red blood cell counts appear to be normally distributed? Why or why not?

2-7. **Flight Arrival Times** Use the flight arrival times listed in *Data Set 15 – Flight Data*. Construct a histogram. Which part of the histogram depicts flights that arrived early, and which part depicts flights that arrived late?

2-8. **Flight Taxi-Out Times** Use the flight taxi-out times listed in *Data Set 15 – Flight Data*. Construct a histogram. If the quality of air traffic procedures is improved so that the taxi-out times vary much less, would the histogram be affected?

2–9. **Frequency Distribution** Shown below is a Statdisk histogram representing the durations (in hours) of flights of NASA space shuttles. Use the displayed histogram to construct a table representing the frequency distribution, and identify the value of the class width.



Class width: ___________ Enter the frequency distribution below:

Scatterplots. *In Experiments 2-10 through 2-13, use the given paired data from the Elementary Statistics 12th Edition data sets included in Statdisk to construct a scatterplot. These data sets are the same data sets included in Appendix B of the textbook.*

2-10. **President's Heights** Refer to *Data Set 12: POTUS* and use the heights of U. S. Presidents and the heights of their main opponents in the election campaign. Does there appear to be a correlation? (*Hint:* Because there are some missing entries, Statdisk will not construct the scatterplot unless the rows with missing entries are deleted.)

2-11. **Brain Volume and IQ** Refer to *Data Set 6: IQ and Brain Size* and use the brain volumes (cm3) and IQ scores. A simple hypothesis is that people with larger brains are more intelligent and they have higher IQ scores. Does the scatterplot support that hypothesis?

2-12. **Bear Chest Size and Weight** Refer to *Data Set 7: Bear Measurements* and use the measured chest sizes and weights of bears. Does there appear to be a correlation between those two variables?

2-13. **Coke Volume and Weight** Refer to *Data Set 19: Cola Weights and Volumes* and use the volumes and weights of regular Coke. Does there appear to be a correlation between volume and weight? What else is notable about the arrangement of the points, and how can it be explained?

2–14. **Effect of Outlier** In Experiment 2–3 we obtained a histogram of 50 earthquake magnitudes . Change the first magnitude value from 0.70 to the outlier of 7.0, then obtain the histogram and print it. How is the histogram affected by the presence of the outlier? Does the outlier disguise the true nature of the distribution of the data?

____

____

____

2–15. **Sorting Data** Open *Data Set 11: Ages of Oscar Winners* and *sort* the actresses ages. Identify any outliers and explain why they were selected.

____

____

2–16. **Working with Your Own Data**
Through observation or experimentation, collect your own set of sample values. Obtain at least 40 values and try to select data from an interesting population. Use Statdisk to generate graphs suitable for describing the distribution of the data. Describe the data and any notable or important characteristics.

____

____

____

____

____

____

3 Statistics for Describing, Exploring, and Comparing Data

Statdisk

The topics of this chapter require that you use Statdisk to enter data, retrieve data, save files, and print results. These functions are covered in Chapter 1 of this manual/workbook. Be sure to understand these functions before beginning this chapter.

Important Characteristics of Data

When describing, exploring, and comparing data sets, the following characteristics are usually extremely important:

1. **Center:** Measure of center, which is a representative or average value that gives us an indication of where the middle of the data set is located.

2. **Variation:** A measure of the amount that the values vary among themselves.

3. **Distribution:** The nature or shape of the distribution of the data, such as bell-shaped, uniform, or skewed.

4. **Outliers:** Sample values that are very far away from the vast majority of the other sample values.

5. **Time:** Changing characteristics of the data over time.

In Chapter 2 of this manual/workbook we addressed the characteristics of distribution and outliers by using histograms and sorted lists of data. In this chapter we address measures of center and variation. We will use the Numbers of Chocolate Chips in Different Brands of Cookies shown in Table 3-1 in this chapter. (Although the listed chocolate chip counts are not included as a data set in Appendix B of the textbook, this data set is included among the Statdisk data sets for *Elementary Statistics* 12th Edition.)

Table 3-1 Numbers of Chocolate Chips in Different Brands of Cookies

Chips Ahoy (regular)																			
22	22	26	24	23	27	25	20	24	26	25	25	19	24	20	22	24	25	25	20
23	30	26	20	25	28	19	26	26	23	25	23	23	23	22	26	27	23	28	24
Chips Ahoy (chewy)																			
21	20	16	17	16	17	20	22	14	20	19	17	20	21	21	18				
20	20	21	19	22	20	20	19	16	19	16	15	24	23	14	24				
Chips Ahoy (reduced fat)																			
13	24	18	16	21	20	14	20	18	12	24	23	28	18	18	19	22	21	22	16
13	20	20	23	24	20	17	20	19	21	27	16	24	19	23	25	14	18	15	19
Keebler																			
29	31	25	32	27	31	30	29	31	26	32	33	32	30	33	29	30			
28	32	35	37	31	24	30	30	34	29	27	24	38	37	32	26	30			
Hannaford																			
13	15	16	21	15	14	14	15	13	13	16	11								
14	12	13	12	14	12	16	17	14	16	14	15								

3-1 Measures of Center and Variation

Important measures of center and variation can be obtained by using Statdisk's "*Explore Data*" or "*Descriptive Statistics*" functions. The *Explore Data* function is recommended because it provides much more information. To explore a list of data, follow these steps.

1. Enter or retrieve a set of sample data using one of these procedures:

   - **Manual entry of data:** Values can be entered in the Statdisk data window.

   - **Retrieve a data set from those included in Appendix B:** Click on the main menu item of **Data Sets** and proceed to select one of the listed data sets.

   - **Retrieve a data set that you created:** Use **File** - **Open** as described in Section 1-4 of this manual/workbook.

2. Click on **Data** in the main menu bar at the top.

3. Click on **Explore Data**.

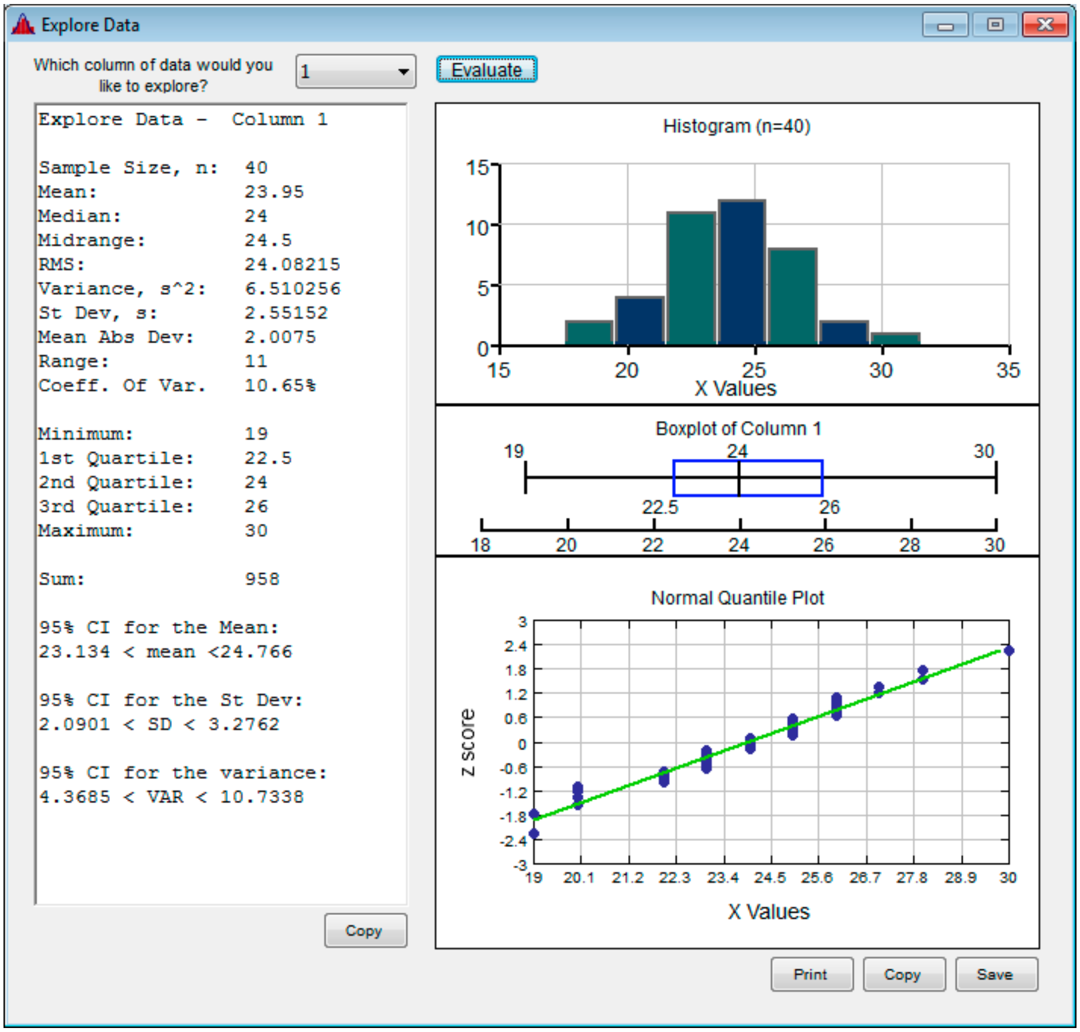
4. Select the column to be used for the calculations and graphs.

5. Click on the **Evaluate** button.

As an example, consider the number of chocolate chips in *Chips Ahoy (regular)*. Those numbers are listed in Table 3-1 and included in the *Elementary Statistics 12th Edition* data set *Chocolate Chips in Cookies*. The result will be as shown on the following page.

From the display on the next page we see that there are n = 40 sample values, the sample mean is 23.95 chocolate chips, the median is 24 chocolate chips, the midrange is 24.5 chocolate chips. The value of "RMS" is the value of the *root mean square* (or quadratic mean) described in the textbook. The variance is s^2 = 6.51 chocolate chips^2 (rounded), and the standard deviation is 2.55 chocolate chips (rounded). The value listed as "Mean Abs. Dev" is the mean absolute deviation described in the textbook. Also see that a histogram is displayed. In addition, there are other results that will be discussed later.

Descriptive statistics, including the measures of center and variation, can also be found by selecting **Data**, then **Descriptive Statistics**, but using the **Explore Data** feature is much better in the sense that more results are provided. Disadvantages of using Explore Data are (1) it is based on a single list of data values, and (2) the histogram uses default settings. If you prefer to use boxplots to compare two or more data sets, use **Data - Boxplot** instead of **Data - Explore Data**. If you prefer to generate a histogram using your own class width and starting point instead of the default settings, use **Data - Histogram** instead of **Data - Explore Data**.


Explore Data
Which column of data would you like to explore?
1
Evaluate
Explore Data - Column 1
Sample Size, n: 40
Mean: 23.95
Median: 24
Midrange: 24.5
RMS: 24.08215
Variance, s^2: 6.510256
St Dev, s: 2.55152
Mean Abs Dev: 2.0075
Range: 11
Coeff. Of Var. 10.65%
Minimum: 19
1st Quartile: 22.5
2nd Quartile: 24
3rd Quartile: 26
Maximum: 30
Sum: 958
95% CI for the Mean:
23.134 < mean <24.766
95% CI for the St Dev:
2.0901 < SD < 3.2762
95% CI for the variance:
4.3685 < VAR < 10.7338
Copy
Histogram (n=40)
X Values
Boxplot of Column 1
Normal Quantile Plot
z score
X Values
Print
Copy
Save

Statdisk

3-2 Quartiles and 5-Number Summary

The textbook includes the definition of a "5-number summary (minimum, 1st quartile, 2nd quartile, 3rd quartile, maximum), and that summary is included with the above Statdisk results. Here is the 5-number summary:

Minimum:	19 chocolate chips
1st Quartile Q_1:	22.5 chocolate chips
2nd Quartile Q_2:	24 chocolate chips
3rd Qaurtile Q_3:	26 chocolate chips
Maximum:	30 chocolate chips

Important Note: The textbook states that there is not universal agreement on a single procedure for calculating quartiles, and different computer programs might yield different results

3-3 Boxplots

The textbook describes the construction of boxplots. They are based on the 5-number summary consisting of the minimum, first quartile, second quartile, third quartile, and maximum. A boxplot is included among the results when Statdisk's *Explore Data* feature is used, but when using boxplots to compare two or more data sets, it is better to use the following procedure that allows you to generate two or more boxplots in the same display so that comparisons become easier.

Procedure for Generating a Boxplot

1. Enter or retrieve a set of sample data using one of these procedures:

   - **Manual entry of data:** Values can be entered in the Statdisk data window.

   - **Retrieve a data set from those included in Appendix B:** Click on the main menu item of **Data Sets** and proceed to select one of the listed data sets.

   - **Retrieve a data set that you created:** Use **File** - **Open** as described in Section 1-4 of this manual/workbook.

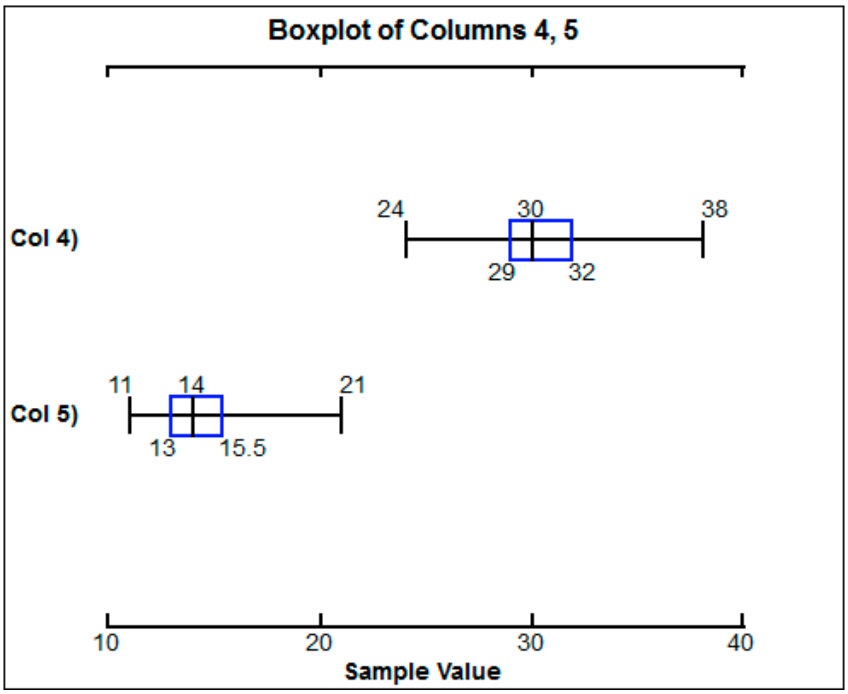
2. Click on **Data** in the main menu bar at the top.

3. Click on **Boxplot**.

4. Select the column(s) to be used for the creation of one or more boxplots. Click a box to insert a check mark or to remove a check mark.

5. Click on the **Boxplot** button.

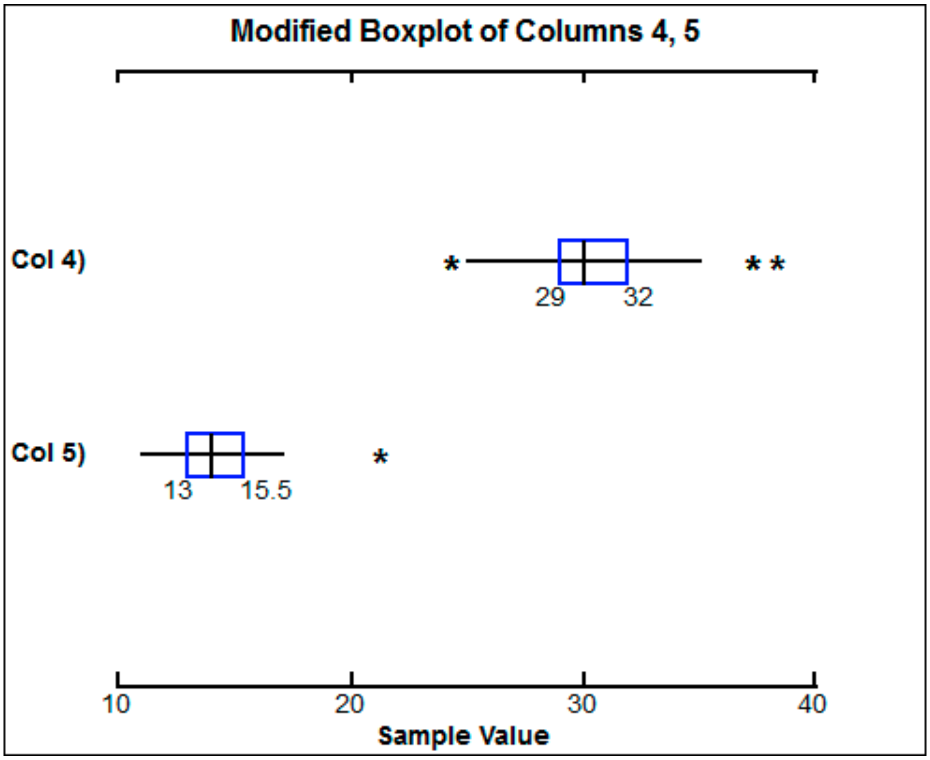
One important advantage of boxplots is that they are very useful in comparing data sets. Shown below is the Statdisk display showing the two boxplots representing the number of chocolate chips in *Keebler* brand cookies and *Hannaford* brand cookies provided in Table 3-1. In the display shown below, the top boxplot depicts the number of chocolate chips in the Keebler brand and the bottom boxplot represents the number of chocolate chips in the Hannaford brand. Because the two boxplots are constructed on the same scale, a comparison becomes easier. The boxplots suggests that counts from the Keebler and Hannaford cookies appear to be very different; the boxplots show that there isn't any overlap, and all of the Hannaford cookies have lower counts than any of the Keebler cookies. It might seem that the Hannaford brand is stingy with its chocolate chips, but the Hannaford brand had many chips that were substantially larger than those in any of the other brands.



Important note: Statdisk generates boxplots based on the minimum, maximum, and three quartiles. Statdisk determines the values of the quartiles by following the same procedure described in the textbook, but other programs may use different procedures, so there may be some differences in boxplot results.

3-4 Modified Boxplots

Part 2 of Section 3-4 in the textbook describes the construction of modified boxplots. The procedure for generating modified boxplots is the same as the procedure for generating boxplots, except that **Modified Boxplot** is selected in Step 5. That is, select **Data** from the main menu, then select **Boxplot**. Now select the column(s) to be used, and click on **Modified Boxplot**. The modified boxplots for chocolate chip counts in Keebler and Hannaford brands are shown below.



3-5 Outliers

Statdisk's sort feature is useful for identifying outliers. When analyzing data, it is important to identify outliers because they can have a dramatic effect on many results. It is usually difficult to recognize an extreme value when it is buried in the middle of a long list of values arranged in a random order, but *outliers become much easier to recognize with sorted data, because they will be found either at the beginning or end*. To identify outliers, simply sort the data, then examine the lowest and highest values to determine whether they are dramatically far from almost all of the other sample values.

Procedure for Sorting Data

1. Enter or retrieve a set of sample data so that the sample values are listed in the Sample Editor window.

2. Click on **Data** in the top menu bar or **Data Tools** in the Sample Editor menu bar.

3. Click on **Sort Data** and the following toolbar will appear above the Sample Editor:

4. By selecting **Sort - All Columns**, you can sort a single column while rearranging the other columns so that data in the same row will stay in the same row.

   By selecting **Sort - One column**, you can sort a single column while leaving the other columns unchanged.

5. Select which column you want to sort by.

6. Select the order in which you want to sort. **A to Z** will sort data in ascending order, **Z to A** will sort data in descending order.

7. Click **Sort**.

3-6 Statistics from a Frequency Distribution

If sample data are summarized in the form of a table representing a frequency distribution such as the one shown below, Statdisk can be used to obtain the important measures of center and variation. The basic idea is to use Statdisk's *Frequency Table Generator* to generate a list of sample values based on the table. Given the table below, for example, Statdisk can generate a list containing 30 values of 25.5 (the midpoint of the first class), 30 values of 35.5 (the midpoint of the second class), 12 values of 45.5 (the midpoint of the third class), and so on. The result will be a list of 79 sample values that correspond to the frequency distribution.

Number of Chocolate Chips - Chips Ahoy (regular)	Frequency
17-20	2
20-23	8
23-26	19
26-29	10
29-32	1

Procedure for Obtaining Statistics From a Frequency Distribution

1. First identify the frequency distribution to be used. (For example, see the above table.)

2. Click on **Data** in the main menu bar at the top.

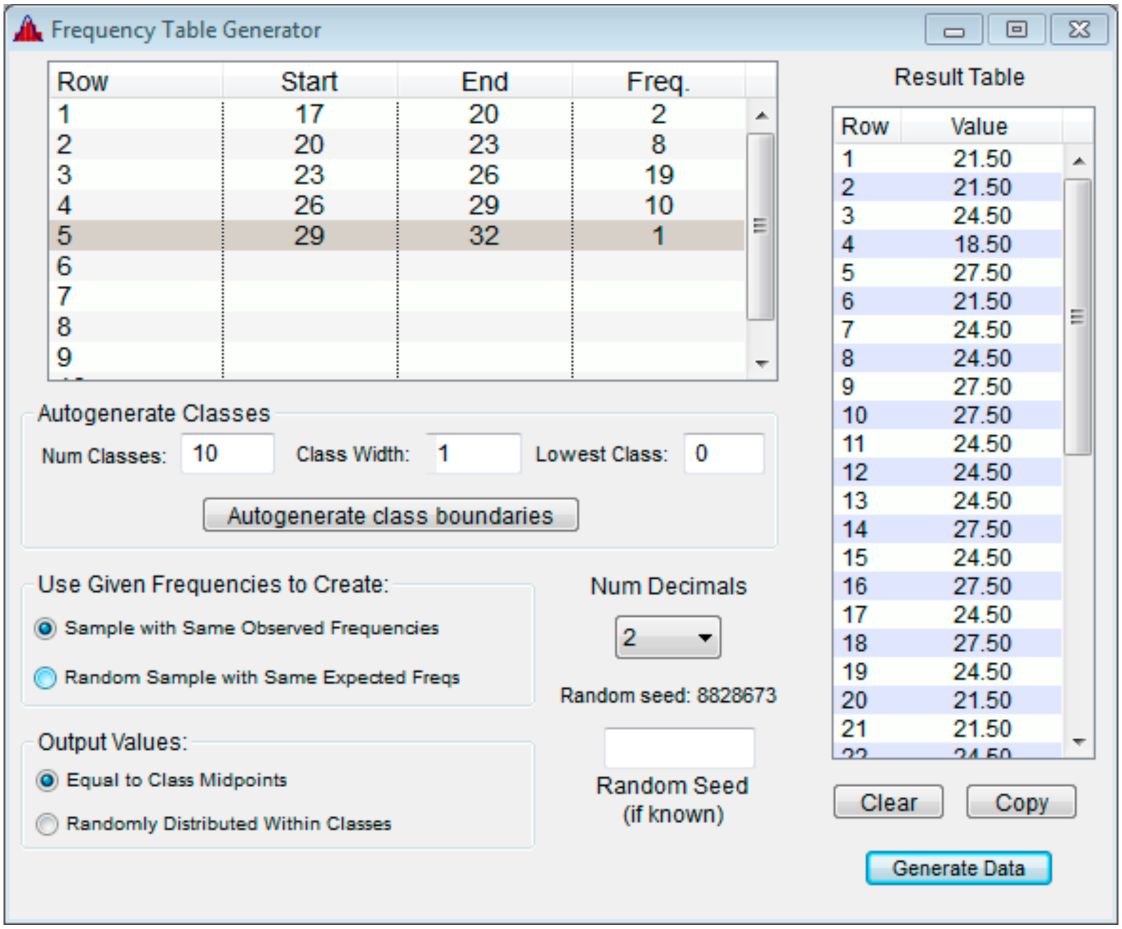
3. Click on **Frequency Table Generator**.

4. See the following Statdisk display showing the entries corresponding to the above frequency distribution.
   - Enter the lower class limits in the "Start" column as shown.
   - Enter the upper class limits in the "End" column as shown.
   - Enter the class frequencies in the "Freq." column as shown.
   - Be sure to select "Sample with Same Observed Frequencies."
   - For the output values, select "Equal to Class Midpoints."
   - For the number of decimals, select at least one more decimal place than is used for the class limits.

5. Click on the **Generate Data** button to get a list of sample values in the column at the right.

6. After the data have been generated and they appear in the column at the right, click on **Copy** and proceed to the main Sample Editor window, where the list of values can be pasted. Once the list of sample values is copied to the data window, you can use **Explore Data** to obtain the descriptive statistics.

Caution: When using the above procedure, realize that you are not generating a list of sample values with the *exact same* characteristics as the original list of sample data. If the original sample values are not known, there is no way to reconstruct the original list from a frequency distribution table. Statistics calculated from the generated data are likely to differ somewhat from the statistics that would be calculated using the original list of sample data. For example, using the original list of ages of chocolate chips, the mean is found to be 24.0 chocolate chips when rounded, but using the generated values from the frequency distribution results in a mean of 24.5 chocolate chips when rounded.


Frequency Table Generator
Row Start End Freq.
1 17 20 2
2 20 23 8
3 23 26 19
4 26 29 10
5 29 32 1
6
7
8
9
Autogenerate Classes
Num Classes: 10
Class Width: 1
Lowest Class: 0
Autogenerate class boundaries
Use Given Frequencies to Create:
Sample with Same Observed Frequencies
Random Sample with Same Expected Freqs
Output Values:
Equal to Class Midpoints
Randomly Distributed Within Classes
Num Decimals
2
Random seed: 8828673
Random Seed
(if known)
Result Table
Row Value
1 21.50
2 21.50
3 24.50
4 18.50
5 27.50
6 21.50
7 24.50
8 24.50
9 27.50
10 27.50
11 24.50
12 24.50
13 24.50
14 27.50
15 24.50
16 27.50
17 24.50
18 27.50
19 24.50
20 21.50
21 21.50
Clear
Copy
Generate Data

CHAPTER 3 EXPERIMENTS: Statistics for Describing, Exploring, and Comparing Data

3-1 **Comparing Heights of Men and Women** In this experiment we use two small data sets as a quick introduction to using some of the basic Statdisk features. (When beginning work with new software, it is wise to first work with small data sets so that they can be entered quickly if they are lost or damaged.) The data listed below are measured heights (cm) of random samples of men and women (taken from *Elementary Statistics 12th Edition* data set *1 - Body Measurements Male/Female* included in Statdisk and Appendix B of the textbook).

Men	178.8	177.5	187.8	172.4	181.7	169.0
Women	163.7	165.5	163.1	166.3	163.6	170.9

a. Find the indicated characteristics of the heights of *men* and enter the results below.

Center: Mean: _____ Median: _____

Variation: St. Dev.:_____ Range: _____

5-Number Summary: Min.:_____ Q_1:_____ Q_2:_____ Q_3:_____ Max.:_____

Outliers: ____________________

b. Find the characteristics of the heights of *women* and enter the results below.

Center: Mean: _____ Median: _____

Variation: St. Dev.:_____ Range: _____

5-Number Summary: Min.:_____ Q_1:_____ Q_2:_____ Q_3:_____ Max.:_____

Outliers: ____________________

c. Compare the results from parts a and b.

____

____

____

Statdisk

3-2 **Working with Larger Data Sets** Repeat Experiment 3-1, but use the sample data for all 40 males and 40 females included in *Elementary Statistics 12th Edition* data set *1 - Body Measurements Male/Female* included in Statdisk and in Appendix B of the textbook. Remember, instead of manually entering the 80 individual heights (which would be no fun at all), open the data sets in Statdisk by selecting **Data Sets** from the top menu bar.

a. Find the indicated characteristics of the heights of *men* and enter the results below.

Center: Mean: _____ Median: _____

Variation: St. Dev.:_____ Range: _____

5-Number Summary: Min.:_____ Q_1:_____ Q_2:_____ Q_3:_____ Max.:_____

Outliers: ___________________

b. Find the characteristics of the heights of *women* and enter the results below.

Center: Mean: _____ Median: _____

Variation: St. Dev.:_____ Range: _____

5-Number Summary: Min.:_____ Q_1:_____ Q_2:_____ Q_3:_____ Max.:_____

Outliers: ___________________

c. Compare the results from parts a and b.

____

____

____

Statdisk

3-3 **Boxplots** Use the same sets of data used in Experiment 3–2 and print boxplots for the heights of the 40 men and the heights of the 40 women. Include both boxplots in the same window so that they can be compared. (*Hint:* Because the two lists are in different data sets, use **Copy/Paste** so that they are in the same Sample Editor window.) Do the boxplots suggest any notable differences in the two sets of sample data?

____

____

Generate a modified boxplot for the heights of the men and Interpret the *asterisk* that appears in the Statdisk display.

____

____

3-4 **Effect of Outlier** In this experiment we will study the effect of an *outlier*. Use the same heights of *men* used in Experiment 3-1, but change the first entry from 178.8 cm. to 1788 cm. (This type of mistake often occurs when the key for the decimal point is not pressed with enough force.) The outlier of 1788 cm is clearly a mistake, because a male with of height of 1788 cm would be 59 feet tall, or about six stories tall. Although this outlier is a mistake, outliers are sometimes correct values that differ substantially from the other sample values.

Men	**1788**	177.5	187.8	172.4	181.7	169.0

Using this modified data set with the height of 178.8 cm changed to be the outlier of 1788 cm, find the following.

Center: Mean: _____ Median: _____

Variation: St. Dev.:_____ Range: _____

5-Number Summary: Min.:_____ Q_1:_____ Q_2:_____ Q_3:_____ Max.:_____

Outliers: ___________________

Based on a comparison of these results to those found in Experiment 3–1, how is the mean affected by the presence of an outlier?

____

How is the median affected by the presence of an outlier?

____

How is the standard deviation affected by the presence of an outlier?

____

3-5 **Describing Data** Use the 40 BMI (body mass index) indices of females from *Elementary Statistics 12th Edition* data set *1a - Body Measurements Female* included in Statdisk and Appendix B of the textbook, and enter the results indicated below.

Center: Mean: _____ Median: _____

Variation: St. Dev.:_____ Range: _____

5-Number Summary: Min.:_____ Q_1:_____ Q_2:_____ Q_3:_____ Max.:_____

Outliers: ___________________

3-6 **Describing Data** Use the 40 BMI (body mass index) indices of males from *Elementary Statistics 12th Edition* data set *1b - Body Measurements Male* included in Statdisk and Appendix B of the textbook, and enter the results indicated below.

Center: Mean: _____ Median: _____

Variation: St. Dev.:_____ Range: _____

5-Number Summary: Min.:_____ Q_1:_____ Q_2:_____ Q_3:_____ Max.:_____

Outliers: ___________________

Compare the BMI indices of females (from Experiment 3-5) and males.

____

____

3-7 **Boxplots** Use the BMI indices of females (see Experiment 3-5) and use the BMI indices of males (see Experiment 3-6). Print their two boxplots together in the same display. (*Hint:* Because the two lists are in different data sets, use **Copy/Paste** so that they are in the same Sample Editor window.) Do the boxplots suggest any notable differences in the two sets of sample data?

____

____

3-8 **Sorting Data** Use the earthquake magnitudes listed in *Elementary Statistics 12th Edition* data set *16 - Earthquake Measurements* in Statdisk and from Appendix B in the textbook. *Sort* that data set by arranging the magnitudes from lowest to highest (*A to Z*). List the first ten values in the sorted column.

____

3-9 **Combining Data** Open the Statdisk *Elementary Statistics 12th Edition* data set *20 - M and M Plain Candy Weights*, which consists of six different columns of data. Use **Copy/Paste** to make a copy of the data all stacked together in column 7 of the current Sample Editor.

a. Find the following results for the combined data in column C7.

Center: Mean: _____ Median: _____

Variation: St. Dev.:_____ Range: _____

5-Number Summary: Min.:_____ Q_1:_____ Q_2:_____ Q_3:_____ Max.:_____

Outliers: ___________________

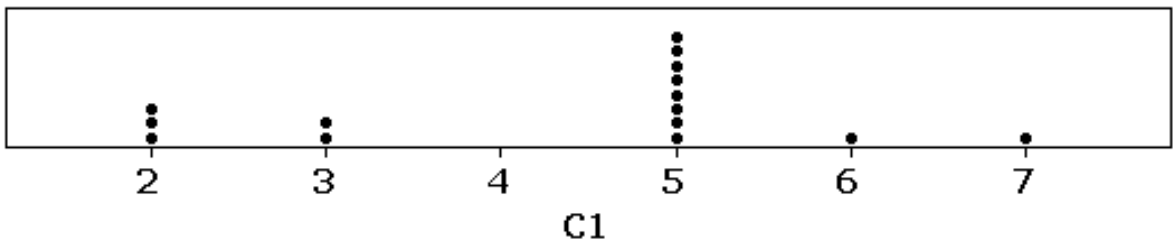
b. Print a boxplot of the data set.

c. Describe the important characteristics of the data set. Be sure to address the nature of the distribution, measures of center, measures of variation, and any other important and notable features.

____

____

3-10 **Statistics from a Dotplot** Shown below is a dotplot. Identify the values represented in this graph, enter them in Statdisk, then find the indicated results.



Center: Mean: _____ Median: _____

Variation: St. Dev.:_____ Range: _____

5-Number Summary: Min.:_____ Q_1:_____ Q_2:_____ Q_3:_____ Max.:_____

Outliers: ____________________

3-11 ***Frequency Distribution*** Use the frequency distribution below to first generate a list of sample data, then find the indicated statistics and enter them here.

Mean: __________ Standard deviation: __________

How does the mean compare to the value of 98.6°F, which is assumed to be the population mean by many people?

____

____

Temperature	Frequency
96.5–96.8	1
96.9–97.2	8
97.3–97.6	14
97.7–98.0	22
98.1–98.4	19
98.5–98.8	32
98.9–99.2	6
99.3–99.6	4

3–12 **Comparing Data** Open the Statdisk data set *Elementary Statistics 12th Edition* data set *19 - Cola Weights and Volumes* and use Statdisk to compare the weights of regular Coke and the weights of diet Coke. Obtain printouts of relevant results. What do you conclude? Can you explain any substantial difference?

____

____

3–13 **Comparing Data** In "Tobacco and Alcohol Use in G–Rated Children's Animated Films," by Goldstein, Sobel and Newman (*Journal of the American Medical Association,* Vol. 281, No. 12), the lengths (in seconds) of scenes showing tobacco use and alcohol use were recorded for animated children's movies. Refer to *Elementary Statistics 12th Edition* data set *8 - Alcohol and Tobacco Use in Animated Children's Movies* in Statdisk and Appendix B from the textbook and find the mean and median for the tobacco times, then find the mean and median for the alcohol times. Does there appear to be a difference between those times? Which appears to be the larger problem: scenes showing tobacco use or scenes showing alcohol use?

____

____

____

____

3-14 **Working with Your Own Data** Through observation or experimentation, collect your own set of sample values. Obtain at least 40 values and try to select data from an interesting population. List the values below.

____

____

____

____

____

____

____

____

Find the following results.

Center: Mean: _____ Median: _____

Variation: St. Dev.:_____ Range: _____

5-Number Summary: Min.:_____ Q_1:_____ Q_2:_____ Q_3:_____ Max.:_____

Outliers: ____________________

Use Statdisk to further explore the data. Obtain printouts of relevant results. Describe the nature of the data. That is, what do the values represent? Describe important characteristics of the data set, and include Statdisk displays to support your observations.

4 Probabilities Through Simulations

Statdisk

4-1 Simulation Methods

Chapter 4 in the Triola textbook focuses on principles of probability theory. That chapter presents a variety of rules and methods for finding probabilities of different events. The textbook focuses on traditional approaches to computing probability values, but this chapter in this manual/workbook focuses instead on an alternative approach based on *simulations*.

> A **simulation** of a procedure is a process that behaves the same way as the procedure, so that similar results are produced.

Mathematician Stanislaw Ulam once studied the problem of finding the probability of winning a game of solitaire, but the theoretical computations involved were too complicated. Instead, Ulam took the approach of programming a computer to simulate or "play" solitaire hundreds of times. The ratio of wins to total games played is the approximate probability he sought. This same type of reasoning was used to solve important problems that arose during World War II. There was a need to determine how far neutrons would penetrate different materials, and the method of solution required that the computer make various random selections in much the same way that it can randomly select the outcome of the rolling of a pair of dice. This neutron diffusion project was named the Monte Carlo Project and we now refer to general methods of simulating experiments as *Monte Carlo methods*. Such methods are the focus of this chapter. The concept of simulation is quite easy to understand with simple examples.

- We could simulate the rolling of a die by using Statdisk to randomly generate whole numbers between 1 and 6 inclusive, provided that the computer selects from the numbers 1, 2, 3, 4, 5, and 6 in such a way that those outcomes are equally likely.

- We could simulate births by flipping a coin, where "heads" represents a baby girl and "tails" represents a baby boy. We could also simulate births by using Statdisk to randomly generate 1s (for baby girls) and 0s (for baby boys).

It is extremely important to construct a simulation so that it behaves just like the real procedure. The following example illustrates a right way and a wrong way.

EXAMPLE Describe a procedure for simulating the rolling of a pair of dice.

In the procedure of rolling a pair of dice, each of the two dice yields a number between 1 and 6 (inclusive) and those two numbers are then added. Any simulation should do exactly the same thing.

Right way to simulate rolling two dice*:* Randomly generate one number between 1 and 6, then randomly generate another number between 1 and 6, then add the two results.

Statdisk

Wrong way to simulate rolling two dice: Randomly generate numbers between 2 and 12. This procedure is similar to rolling dice in the sense that the results are always between 2 and 12, but these outcomes between 2 and 12 are equally likely. With real dice, the values between 2 and 12 are *not* equally likely. This simulation would yield terrible results.

4-2 Statdisk Simulation Tools

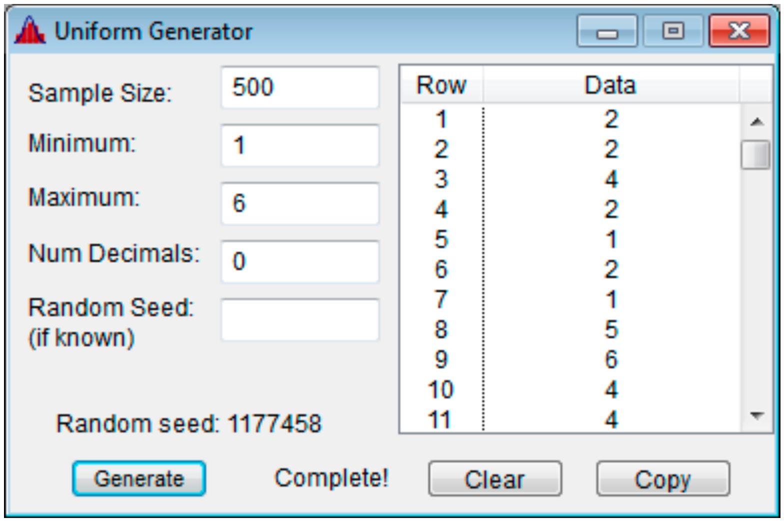
Statdisk includes several different tools that can be used for simulations. If you click on the main menu item of **Data**, you get a submenu that includes these items:

Normal Generator
Uniform Generator
Binomial Generator
Poisson Generator
Coins Toss Simulator
Dice Generator
Frequency Table Generator
Random Seed Generator

Here are descriptions of these menu items available by clicking on **Data** in the top menu:

- **Normal Generator**: Generates a sample of data randomly selected from a population having a normal distribution. You must enter the desired sample size, population mean and standard deviation. The number of decimal places can be specified, so enter 0 if you want only whole numbers. (Normal distributions are described in the textbook. For now, consider a normal distribution to be a distribution that is bell–shaped.)

- **Uniform Generator:** This tool is particularly good for the *random generation of integers*. It generates numbers between a desired minimum value and maximum value. You can specify the number of decimal places, so enter 0 if you want only whole numbers. The generated values are "uniform" in the sense that all possible values have the same chance of being selected. For example, if you select a sample size of 500, a minimum of 1, a maximum of 6, and 0 decimal places, the results simulate the rolling of a single die 500 times, as shown in the following Statdisk display. (See also the Dice Generator described below.)

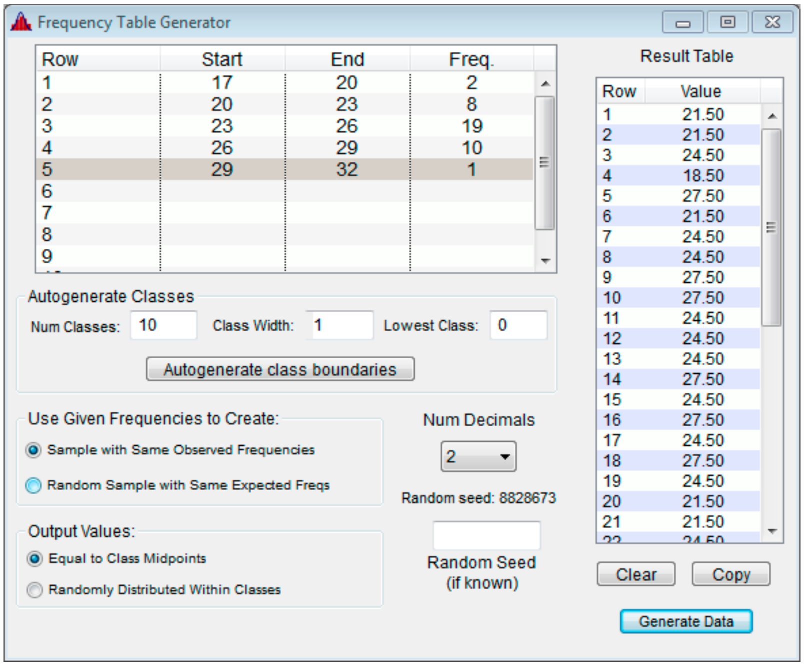


- **Binomial Generator**: Generates numbers of successes for a binomial probability distribution. You specify the number of values to be generated (sample size), the probability of success, and the number of trials in each case. Binomial probability distributions are discussed later in the textbook, so this item can be ignored for now.

- **Coins Toss Simulator**: This tool is particularly useful for those cases in which there are two possible outcomes (such as boy/girl) that are equally likely, as is the case with coin tosses. You select the number of generated values that you want (trials), and you also select the number of coins to be tossed in each trial. The generated values are the numbers of heads that turn up.

- **Dice Generator:** Select the number of generated values that you want (trials), and select the number of dice to be rolled in each trial. Also select the number of sides the dice have (use 6 for standard dice). The generated values are the totals of the numbers of dots that turn up on the dice.

- **Frequency Table Generator:** This feature can be used to generate sample data drawn from a population that can be described by a frequency distribution. See section 3-6 of this manual for more detail. Click on the main menu item of **Data** and then select **Frequency Table Generator** to get the window shown below. You can automatically generate the class boundaries. There are other options indicated in the window. Click **Generate Data** when you are ready. The generated data can be copied to the Sample Editor window where it can be used with other modules, such as Descriptive Statistics or Histogram.



Random Seed: The preceding Statdisk tools include an option for entering a "random seed" if it is known. This entry will be usually left blank, but if you record a seed that was used or if you enter a value for your own seed, you can duplicate results that were previously obtained. For example, an instructor might assign the generation of data with a particular random seed so that everyone in the class will get identical results. Most of the time, you will *not* enter a value for the random seed so that your results will be different each time. This makes life a bit more interesting.

The Random Seed menu item allows you generate a list of random seeds that could be used in the other options for randomly generating data.

4-3 Sorting Data

In some cases, it is very helpful to sort data. (Data are *sorted* when they are arranged in order.) In Section 2-5 of this manual/workbook we described the procedure for sorting data. This procedure is slightly modified for data that have been generated:

Procedure for Sorting Data

1. After generating sample data using the tools described in Section 4-2 of this manual/workbook, use Copy/Paste to copy the data to the Sample Editor window.

2. Click on **Data** in the top menu bar or **Data Tools** in the Sample Editor menu bar.

3. Click on **Sort Data** and the following toolbar will appear above the Sample Editor:

4. Selecting **Sort One column** and the select the column that contains the sample data.

5. Select the order in which you want to sort. **A to Z** will sort data in ascending order, **Z to A** will sort data in descending order.

6. Click **Sort**.

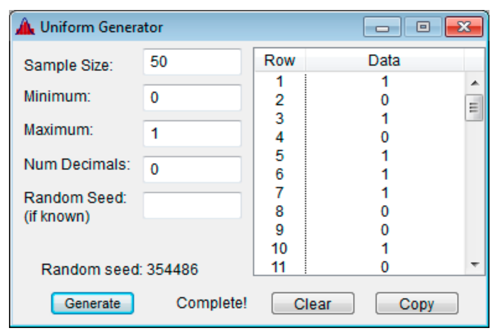
4-4 Simulation Examples

We now proceed to illustrate the preceding Statdisk features by describing specific simulations.

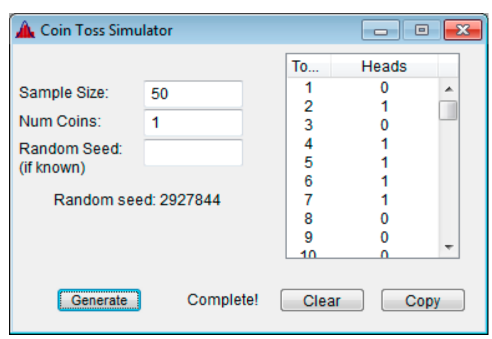
Simulation 1: Generating 50 births (boys/girls)

To simulate 50 births with the assumption that boys and girls are equally likely, use either of the following approaches:

- Use Statdisk's **Uniform Generator** (see Section 4-2) to generate 50 integers between 0 (minimum) and 1 (maximum). Be sure to enter 0 for the number of decimal places. If you arrange the results in order, it is very easy to count the number of 0s (or boys) and the number of 1s (or girls). See Section 4-3 of this manual/workbook for the procedure for sorting data.



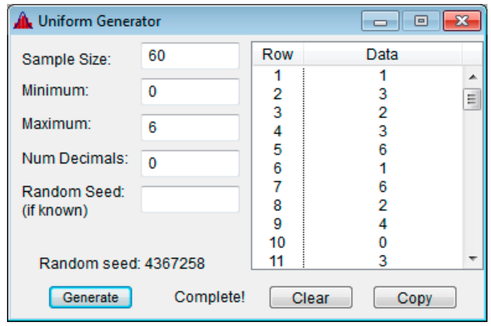
- Use Statdisk's **Coins Toss Simulator** (see Section 4-2). Enter 50 for the sample size and enter 1 for the number of coins. Again, it is very easy to count the number of 0s (or boys) and the number of 1s (or girls) if the data are sorted, as described in Section 4-3 of this manual/workbook.



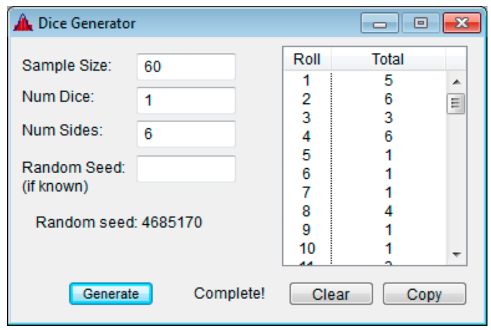
Simulation 2: Rolling a single die 60 times

To simulate 60 rolls of a single die, use either of these approaches:

- Use Statdisk's **Uniform Generator** (see Section 4-2) to generate 60 integers between 1 and 6. (Enter 60 for the sample size and be sure to enter 0 for the number of decimal places.) Again, arranging them in order makes it easy to count the number of 1s, 2s, and so on.

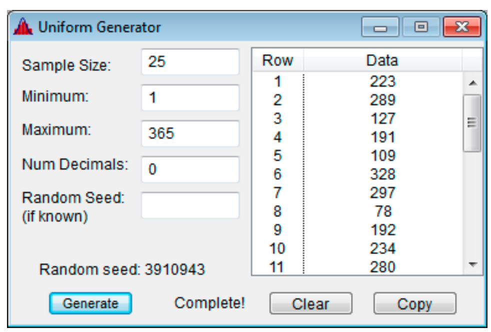


- Use Statdisk's **Dice Generator**. Enter 60 for the sample size, enter 1 for the number of dice, and enter 6 for the number of sides on the die.



Simulation 3: Generating 25 birth dates

Instead of generating 25 results such as "January 1," or "November 27," randomly generate 25 integers between 1 and 365 inclusive. (We are ignoring leap years). Use Statdisk's **Uniform Generator** and enter 25 for the sample size. Also enter 1 for the minimum, 365 for the maximum, and be sure to enter 0 for the number of decimal places (so that only integers are generated). See the sample Statdisk display shown here. The first generated value of 23 corresponds to January 23, because the 23rd day in a year is January 23.



Even though the display shows only the first 11 of the 25 birth dates, we can examine the complete list to determine whether two values occur twice. You can examine all 25 entries by scrolling, but if you sort the simulated birth dates by copying the list of data to the Sample Editor window and using the sort feature (by selecting **Data Tools – Sort Data** from the Sample Editor menu bar), it becomes much easier to scan the sorted list and determine whether there are two birth dates that are the same. If there are two birth dates that are the same, they will show up as *consecutive* equal values in the sorted list.

CHAPTER 4 EXPERIMENTS: Probabilities through Simulations

4-1 **Birth Simulation** Use Statdisk to simulate 500 births, where each birth results in a boy or girl. Sort the results, count the number of girls, and enter that value here:_________

Based on that result, estimate the probability of getting a girl when a baby is born. Enter the estimated probability here: __________

The preceding estimated probability is likely to be different from 0.5. Does this suggest that the computer's random number generator is defective? Why or why not?

____

____

4-2 **Dice Simulation** Use Statdisk to simulate 1000 rolls of a pair of dice. Sort the results, then find the number of times that the total was exactly 7. Enter that value here:_____

Based on that result, estimate the probability of getting a 7 when two dice are rolled. Enter the estimated probability here:__________

How does this estimated probability compare to the computed (theoretical) probability of 1/6 or 0.167? ____

____

4-3 **Probability of at Least 11 Girls**

a. Use Statdisk to simulate 20 births. Does the result consist of at least 11 girls? _____
b. Repeat part (a) nine more times and record the yes/no result from part (a) along with the other nine results here: ___ ___ ___ ___ ___ ___ ___ ___ ___ ___
c. Based on the results from part (b), what is the estimated probability of getting at least 11 girls in 20 births? __________

4-4 **Brand Recognition** The probability of randomly selecting an adult who recognizes the brand name of McDonald's is 0.95 (based on data from Franchise Advantage). Use Statdisk to develop a simulation so that each individual outcome should be an indication of one of two results: (1) The consumer recognizes the brand name of McDonald's; (2) the consumer does not recognize the brand name of McDonald's. (Hint: Randomly generate integers between 1 and 100, and consider results from 1 through 95 to be people who recognize the brand name of McDonald's.) Conduct the simulation 1000 times and record the number of consumers who recognize the brand name of McDonald's. If possible, obtain a printed copy of the results. Is the proportion of those who recognize McDonald's reasonably close to the value of 0.95?

____

____

4-5 **Simulating Hybridization** When Mendel conducted his famous hybridization experiments, he used peas with green pods and yellow pods. One experiment involved crossing peas in such a way that 75% of the offspring peas were expected to have green pods, and 25% of the offspring peas were expected to have yellow pods. Use Statdisk to simulate 1000 peas in such a hybridization experiment. Each of the 1000 individual outcomes should be an indication of one of two results: (1) The pod is green; (2) the pod is yellow. Is the percentage of yellow peas from the simulation reasonably close to the value of 25%?

____

____

4-6 **Probability of at Least 55 Girls** Use Statdisk to conduct a simulation for estimating the probability of getting at least 55 girls in 100 births. Enter the estimated probability here:___________ Describe the procedure used to obtain the estimated probability.

____

In testing a gender-selection method, assume that the Biogene Technology Corporation conducted an experiment with 100 couples who were treated, and that the 100 births included at least 55 girls. What should you conclude about the effectiveness of the treatment?

____

4-7 **Probability of at Least 65 Girls** Use Statdisk to conduct a simulation for estimating the probability of getting at least 65 girls in 100 births. Enter the estimated probability here:___________

Describe the procedure used to obtain the estimated probability.

____

In testing a gender-selection method, if the Biogene Technology Corporation conducted an experiment with 100 couples who were treated, and the 100 births included at least 65 girls, what should you conclude about the effectiveness of the treatment?

____

4-8 **Gender-Selection Method** As of this writing, the latest results available from the Microsort YSORT method of gender-selection consist of 127 boys in 152 births. That is, among 152 sets of parents using the YSORT method for increasing the likelihood of a boy, 127 actually had boys and the other 25 had girls. Assuming that the YSORT method has no effect and that boys and girls are equally likely, use Statdisk to simulate 152 births. Is it unusual to get 127 boys in 152 births? What does the result suggest about the YSORT method?

____

____

4-9 **Simulating Three Dice** Develop a simulation for rolling three dice. Simulate the rolling of the three dice 100 times. Describe the simulation, then use it to estimate the probability of getting a total of 10 when three dice are rolled.

____

____

4-10 **Simulating Left–Handedness** Ten percent of us are left-handed. In a study of dexterity, people are randomly selected in groups of five. Develop a simulation for finding the probability of getting at least one left-handed person in a group of five. Simulate 100 groups of five. How does the probability compare to the correct result of 0.410, which can be found by using the probability rules in the textbook?

____

____

4-11 **Nasonex Treatment** Nasonex is a nasal spray used to treat allergies. In clinical trials, 1671 subjects were given a placebo, and 2 of them developed upper respiratory tract infections. Another 2103 patients were treated with Nasonex and 6 of them developed upper respiratory tract infections. Assume that Nasonex has no effect on upper respiratory tract infections, so that the rate of those infections also applies to Nasonex users. Using the placebo rate of 2/1671, simulate groups of 2103 subjects given the Nasonex treatment, then determine whether a result of 6 upper respiratory tract infections could easily occur. Describe the results. What do the results suggest about Nasonex as a cause of upper respiratory tract infections?

____

____

4-12 **Birthdays** Simulate a class of 25 birth dates by randomly generating 25 integers between 1 and 365. (We will ignore leap years.) Arrange the birth dates in ascending order, then examine the list to determine whether at least two birth dates are the same. (This is easy to do, because any two equal integers must be next to each other.)

Generated "birth dates:" ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___
___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

Are at least two of the "birth dates" the same? __________

4-13 **Birthdays** Repeat the preceding experiment nine additional times and record all ten of the yes/no responses here:

____ ____ ____ ____ ____ ____ ____ ____ ____ ____

Based on these results, what is the probability of getting at least two birth dates that are the same (when a class of 25 students is randomly selected)? __________

4-14 **Birthdays** Repeat Experiments 4-12 and 4-13 for 50 people instead of 25. Based on the results, what is the estimated probability of getting at least two birth dates that are the same (when a class of 50 students is randomly selected)? __________

4-15 **Birthdays** Repeat Experiments 4-12 and 4-13 for 100 people instead of 25. Based on the results, what is the estimated probability of getting at least two birth dates that are the same (when a class of 100 students is randomly selected)? __________

4-16 **Normally Distributed IQ Scores** IQ scores are normally distributed with a mean of 100 and a standard deviation of 15. Generate a normally distributed sample of 800 IQ scores by using the given mean and standard deviation. Sort the results (arrange them in ascending order).

a. Examine the sorted results to estimate the probability of randomly selecting someone with an IQ score between 90 and 110 inclusive. Enter the result here.______
b. Examine the sorted results to estimate the probability of randomly selecting someone with an IQ score greater than 115. __________
c. Examine the sorted results to estimate the probability of randomly selecting someone with an IQ score less than 120. __________
d. Repeat part a of this experiment nine more times and list all ten probabilities here.

_____ _____ _____ _____ _____ _____ _____ _____ _____ _____

e. Examine the ten probabilities obtained above and comment on the *consistency* of the results.

____

f. How might we modify this experiment so that the results can become more consistent?

____

g. If the results appear to be very consistent, what does that imply about any individual sample result?

____

Statdisk

4-17 **Law of Large Numbers** In this experiment we test the Law of Large Numbers, which states that "as an experiment is repeated again and again, the empirical probability of success tends to approach the actual probability." We will use a simulation of a single die, and we will consider a success to be the outcome of 1. (Based on the classical definition of probability, we know that $P(1) = 1/6 = 0.167$.)

a. Simulate 5 trials by generating 5 integers between 1 and 6. Count the number of 6s that occurred and divide that number by 5 to get the empirical probability.
Based on 5 trials, $P(1)$ = _____.

b. Repeat part (a) for 25 trials. Based on 25 trials, $P(1)$= _____.

c. Repeat part (a) for 50 trials. Based on 50 trials, $P(1)$= _____.

d. Repeat part (a) for 500 trials. Based on 500 trials, $P(1)$= _____.

e. Repeat part (a) for 1000 trials. Based on 1000 trials, $P(1)$ = _____.

f. In your own words, generalize these results in a restatement of the Law of Large Numbers.

____

____

____

____

4-18 **Detecting Fabricated Results** Consider a class experiment in which some students actually flip a coin 200 times and record the results, while others make up their own results for 200 coin flips. It is easy to identify the fabricated results using this criterion: If there is a run of six heads or six tails, the results are real, but if there is no such run, the results are fabricated. This is based on the principle that when fabricating results, people almost never include a run of six or more heads or tails, but with 200 actual coin flips, there is a very high probability of getting such a run of at least six heads or tails. The calculation for the probability of getting a run of at least six heads or six tails is *extremely* difficult, so conduct simulations to estimate that probability. Describe the process and identify the estimated probability that when 200 coins are tossed, there is run of at least six heads or tails.

____

____

____

____

4-19 **Sticky Probability Problem** Consider the following exercise, which is extremely difficult to solve with the formal rules of probability.

> *Two points along a straight stick are randomly selected. The stick is then broken at these two points. Find the probability that the three pieces can be arranged to form a triangle.*

Instead of attempting a solution using formal rules of probability, we will use a simulation. The length of the stick is irrelevant, so assume it's one unit long and its length is measured from 0 at one end to 1 at the other end. Use Statdisk to randomly select the two break points with the random generation of two numbers from a uniform distribution with a minimum of 0, a maximum of 1, and 4 decimal places. Plot the break points on the "stick" below.

0__________________________________1

A triangle can be formed if the longest segment is less than 0.5, so enter the lengths of the three pieces here: ___________ ___________ ___________

Can a triangle be formed?

Now repeat this process nine more times and summarize all of the results below.

Trial	Break Points		Triangle formed?
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

Based on the ten trials, what is the estimated probability that a triangle can be formed? ______ This estimate gets better with more trials.

5 Probability Distributions

Statdisk

5-1 Exploring Probability Distributions

Chapter 5 in the Triola textbook discusses probability distributions, and its focus is *discrete* probability distributions only. These important definitions are introduced:

Definitions

A **random variable** is a variable (typically represented by *x*) that has a single numerical value, determined by chance, for each outcome of a procedure.

A **probability distribution** is a description that gives the probability for each value of the random variable. It is often expressed in the format of a table, formula, or graph.

When working with a probability distribution, we should consider the same important characteristics introduced in Chapter 2 of this workbook:

1. **Center:** Measure of center, which is a representative or average value that gives us an indication of where the middle of the data set is located.
2. **Variation:** A measure of the amount that the values vary among themselves.
3. **Distribution:** The nature or shape of the distribution of the data, such as bell-shaped, uniform, or skewed.
4. **Outliers:** Sample values that are very far away from the vast majority of the other sample values.
5. **Time:** Changing characteristics of the data over time.

Section 5-1 of this workbook addresses these important characteristics for probability distributions. The characteristics of center and variation are addressed with formulas for finding the mean, standard deviation, and variance of a probability distribution. The characteristic of distribution is addressed through the graph of a probability histogram.

Although Statdisk is not designed to deal directly with a probability distribution, it can often be used. Consider Table 5-1 from the textbook, reproduced here. This table lists the probabilities for the number girls in 2 births.

Table 5-1 Probability Distribution for the Number of Girls in Two Births

Number of Girls *x*	*P*(*x*)
0	0.25
1	0.50
2	0.25

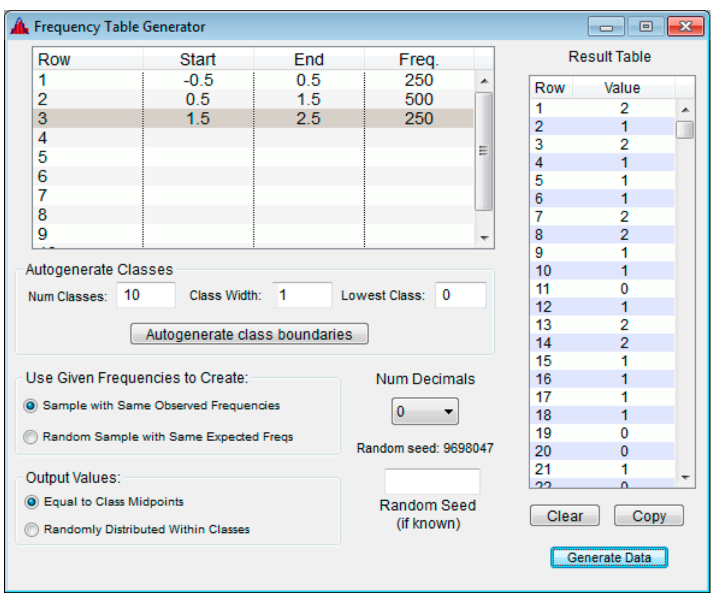
If you examine the data in the table, you can verify that a probability distribution is defined because the three key requirements are satisfied:

1. The variable *x* is a numerical random variable and its values are associated with probabilities, as in Table 5-1.
2. The sum of the probabilities is 1, as required. (0.25 + 0.50 + 0.25 = 1.)
3. Each value of *P*(*x*) is between 0 and 1. (Specifically, 0.25 and 0.50 and 0.25 are each between 0 and 1 inclusive.)

Having determined that Table 5-1 does define a probability distribution, let's now see how we can use Statdisk to find the mean μ and standard deviation σ. The basic approach is to use Statdisk's **Frequency Table Generator** to construct a table of actual values with the same distribution given in the table.

Statdisk Procedure for Working with a Probability Distribution

1. Click on **Data**.
2. Select the menu item of **Frequency Table Generator**.
3. Enter class limits and frequencies that correspond to the probability distribution. Shown below are the entries corresponding to the probability distribution given in Table 5-1. See the first class where the value of 0 is represented by the class limits of -0.5 to 0.5 and a frequency of 250 (based on a probability of 0.25).



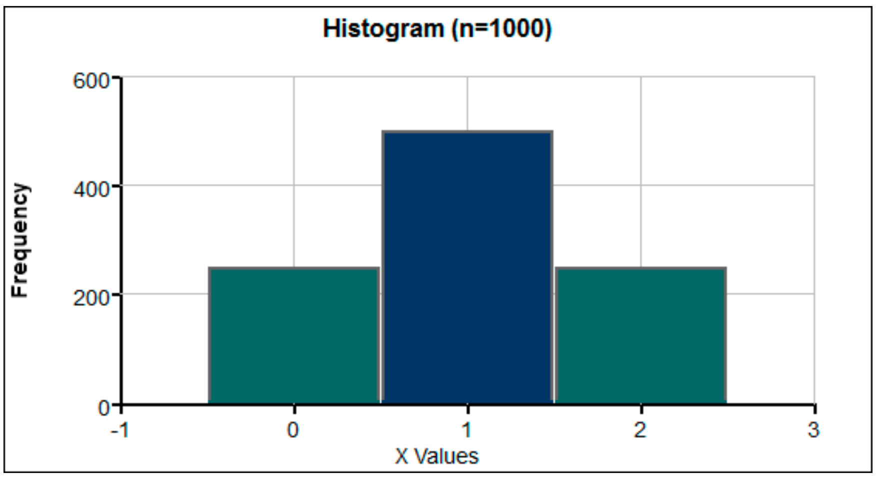
Note these settings in the preceding window:

- The number of decimals in the generated values is 0.
- Select "*Samples with Same Observed Frequencies*."
- Select output values "*Equal to Class Midpoints*."

4. Click on the **Generate Data** button.

5. Statdisk will proceed to generate a set of data corresponding to the probability distribution.

6. You can now use **Copy/Paste** to copy the data to the Sample Editor window where you can find the mean and standard deviation or you can construct a histogram. There's one correction needed: If you use the *Descriptive Statistics* function with the generated data, the computed standard deviation and variance could be off a little, because the calculation assumes the use of *sample* data, whereas we should consider the data to be a *population*. If the sample size is large, the discrepancy will be small. For the data from above table, the actual standard deviation is $\sigma = 0.7071068$, but Statdisk yields $\sigma = 0.7074606$. The value of the mean will be correct. Here is a Statdisk histogram that shows the shape of the probability distribution:



5-2 Binomial Distributions

A **binomial distribution** is defined in Section 5-3 of the Triola textbook to be to be a probability distribution that meets *all* of the following requirements:

1. The experiment must have a fixed number of trials.
2. The trials must be independent. (The outcome of any individual trial doesn't affect the probabilities in the other trials.)
3. Each trial must have all outcomes classified into two categories.
4. The probabilities must remain constant for each trial.

We also introduced notation with S and F denoting success and failure for the two possible categories of all outcomes. Also, p and q denote the probabilities of S and F, respectively, so that $P(S) = p$ and $P(F) = q$. We also use the following symbols.

Statdisk

n denotes the fixed number of trials

x denotes a specific number of successes in *n* trials so that *x* can be any whole number between 0 and *n*, inclusive.

p denotes the probability of success in *one* of the *n* trials.

q denotes the probability of failure in *one* of the *n* trials.

***P*(*x*)** denotes the probability of getting exactly *x* successes among the *n* trials.

Section 5-3 in the Triola textbook describes three methods for determining probabilities in binomial experiments. Method 1 uses the binomial probability formula:

$$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

Method 2 requires computer usage. We noted in the textbook that if a computer and software are available, this method of finding binomial probabilities is fast and easy, as shown in the Statdisk procedure that follows. We illustrate the Statdisk procedure with the following problem:

> **EXAMPLE** ***Devil of a Problem*** Based on a recent Harris poll, 60% of adults believe in the devil. Assuming that we randomly select 5 adults, find the probability that exactly 3 of the 5 adults believe in the devil.

For this example, $n = 5$, $p = 0.60$, and the possible values of x are 0, 1, 2, 3, 4, 5.

Statdisk Procedure for Finding Probabilities with a Binomial Distribution

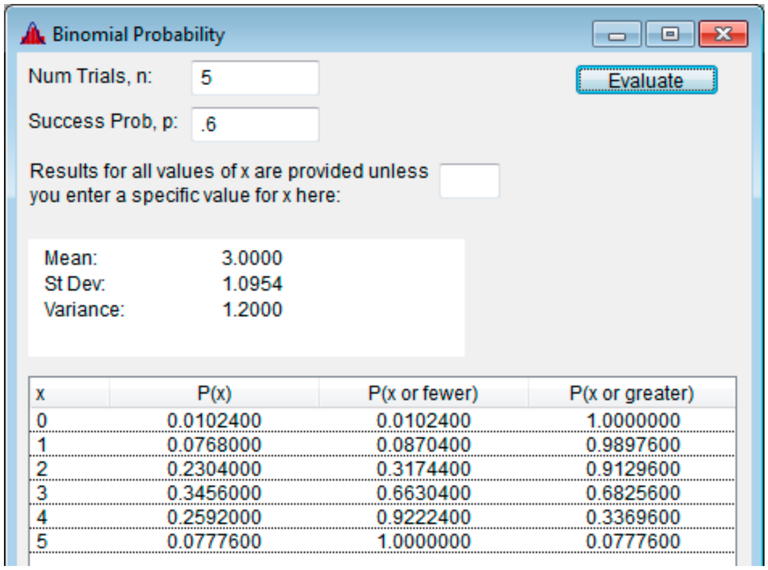
1. Click on **Analysis** from the top menu bar.

2. Select **Probability Distributions.**

3. Select **Binomial Distribution**.

4. You will now see a dialog box:
   - Enter the number of trials *n* (*n* = 5 in the example).
   - Enter the probability of success *p* (*p* = 0.60 in the example).

5. Click on **Evaluate**.



x	P(x)	P(x or fewer)	P(x or greater)
0	0.0102400	0.0102400	1.0000000
1	0.0768000	0.0870400	0.9897600
2	0.2304000	0.3174400	0.9129600
3	0.3456000	0.6630400	0.6825600
4	0.2592000	0.9222400	0.3369600
5	0.0777600	1.0000000	0.0777600

From this Statdisk display, we can see that *P*(3) = 0.34560. Note that the display includes values of the mean, standard deviation, and variance. Also, Statdisk includes cumulative probabilities along with probabilities for the individual values of *x*. From the above display we can find probabilities such as these:

- The probability of 2 or fewer correct responses is 0.31744.
- The probability of 3 or more correct responses is 0.68256.

If you only want the probabilities associated with a specific value displayed, you can enter a specific value for *x* in the dialog.

The table of binomial probabilities (Table A-1) in Appendix B of the textbook includes limited values of *n* and *p*, but Statdisk is so much more flexible in the values of *n* and *p* that can be used.

5-3 Poisson Distributions

Textbooks in the Triola Statistics Series (except for *Essentials of Statistics*) discuss the Poisson distribution. A Poisson distribution is a discrete probability distribution that applies to occurrences of some event *over a specified interval*. The random variable *x* is the number of occurrences of the event in an interval, such as time, distance, area, volume, or some similar unit. The probability of the event occurring *x* times over an interval is given by this formula:

$$P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!} \quad \text{where } e \approx 2.71828$$

The textbook also notes that the Poisson distribution is sometimes used to approximate the binomial distribution when $n \geq 100$ and $np \leq 10$; in such cases, we use $\mu = np$. If using Statdisk, the Poisson approximation to the binomial distribution isn't used much, because we can easily find binomial probabilities for a wide range of values for *n* and *p*, so an approximation is not necessary.

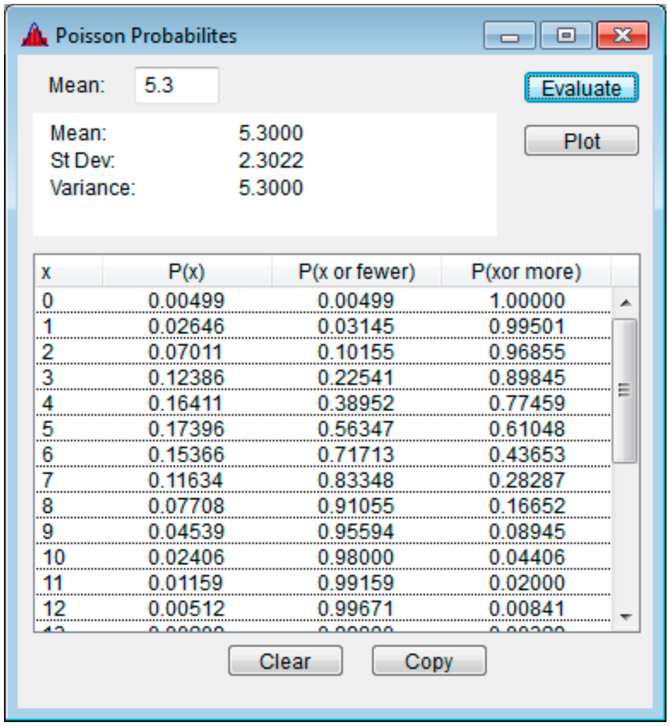
We illustrate the Statdisk procedure with the following problem:

> **EXAMPLE *Hurricanes*** For a recent period of 100 years, there were 530 Atlantic hurricanes (based on data from the University of Maryland Department of Geography and Environmental Systems), so that the mean number of hurricanes is 5.3 per year. Find the probabilities corresponding to *x* = 0, 1, 2, 3, 4, 5 hurricanes in a year.

For this example, $\mu = 5.3$ and the possible values of *x* are 0, 1, 2, 3, 4, 5.

Statdisk Procedure for Finding Probabilities for a Poisson Distribution

1. Determine the value of the mean μ.
2. Click on **Analysis** from the top menu bar.
3. Select **Probability Distributions.**
4. Select **Poisson Distribution**.
5. Enter the value of the mean μ, then click on the **Evaluate** button. In the example, $\mu = 5.3$.



Note that display includes values for the mean, standard deviation, and variance. The probabilities and cumulative probabilities are listed in lower portion of the window. For example, *P*(2) = 0.07011, which is the probability of exactly two Atlantic hurricanes in a given year. The probability of 0, 1, or 2 hurricanes is 0.10155. The probability of *x* having a value of 2 or more is 0.96855, which is the probability of at least two Atlantic hurricanes in a given year.

5-4 Cumulative Probabilities

The main objective of this section is to stress the point that *cumulative probabilities* are often critically important. By *cumulative* probability, we mean the probability that the random variable *x* has a range of values instead of a single value. Here are typical examples:

- Find the probability of getting *at least* 13 girls in 14 births.
- Find the probability of *more than* 5 wins when roulette is played 200 times.

A *cumulative* probability is often much more important than the probability of any individual event. Section 5-2 of the textbook includes criteria for determining when results are *unusual*, and the following criteria involve cumulative probabilities.

Using Probabilities to Determine When Results Are Unusual

- **Unusually *high* number of successes:** *x* successes among *n* trials is an *unusually high* number of successes if the probability of *x* or more successes is unlikely with a probability of 0.05 or less. This criterion can be expressed as follows:
  P(*x* or more) ≤ 0.05.*

- **Unusually *low* number of successes:** *x* successes among *n* trials is an *unusually low* number of successes if the probability of *x* or fewer successes is unlikely with a probability of 0.05 or less. This criterion can be expressed as follows:
  P(*x* or fewer) ≤ 0.05.*

  *The value 0.05 is not absolutely rigid. Other values, such as 0.01, could be used to distinguish between results that can easily occur by chance and events that are very unlikely to occur by chance.

We illustrate the importance of cumulative probabilities using Example 7 in Section 5-2 of the textbook:

> **EXAMPLE** The Chapter Problem includes results consisting of 879 girls in 945 births. Is 879 girls in 945 births an unusually high number of girls? What does it suggest about the effectiveness of the XSORT method of gender selection?

Here we want the cumulative probability of getting 879 or more girls, assuming that boys and girls are equally likely. Using Statdisk with the methods in Section 5-2 of this workbook, we find that the probability of 879 or more girls is 0.000 when rounded to three decimal places. This indicates that 879 girls in 945 is an unusually high number of girls, suggesting that the XSORT method of gender selection is effective. Here, cumulative probabilities play a critical role in identifying results that are considered to be unusual. Later chapters focus on this important concept. If you examine the dialog boxes used for finding binomial probabilities and Poisson probabilities, you can see that in addition to obtaining probabilities for specific numbers of successes, you can also obtain cumulative probabilities.

CHAPTER 5 EXPERIMENTS: Probability Distributions

5-1 **Genetic Disorder** Three males with an X–linked genetic disorder have one child each. The random variable *x* is the number of children among the three who inherit the X–linked genetic disorder, and the probability distribution for that number of children is given in the table below.

x	*P*(*x*)
0	0.4219
1	0.4219
2	0.1406
3	0.0156

Use Statdisk with the procedure described in Section 5-1 of this workbook for the following.

Find the mean and standard deviation:

Mean:______________ St. Dev.:______________

5-2 **Overbooked Flights** Air America has a policy of routinely overbooking flights. The random variable x represents the number of passengers who cannot be boarded because there are more passengers than seats (based on data from an IBM research paper by Lawrence, Hong, and Cherrier).

x	*P*(*x*)
0	0.051
1	0.141
2	0.274
3	0.331
4	0.187

Use Statdisk with the procedure described in Section 5-1 of this manual/workbook for the following.

Find the mean and standard deviation:

Mean:______________ St. Dev.:______________

5-3 **Genetics** Groups of five babies are randomly selected. In each group, the random variable *x* is the number of babies with green eyes (base on data from a study by Dr. Sorita Soni at Indiana University). (The symbol 0+ denotes a positive probability value that is very small.)

x	*P*(*x*)
0	0.528
1	0.360
2	0.098
3	0.013
4	0.001
5	0+

Use Statdisk with the procedure described in Section 5-1 of this manual/workbook for the following.

Find the mean and standard deviation:

Mean:_____________ St. Dev.:_____________

5-4 **Binomial Probabilities** Use Statdisk to find the binomial probabilities corresponding to n = 4 and p = 0.05. Enter the results below, along with the corresponding results found in Table A-1 of the textbook.

x	*P*(*x*) from Statdisk	*P*(*x*) from Table A-1
0		
1		
2		
3		
4		

By comparing the above results, what advantage does Statdisk have over Table A-1?

____

____

5-5 **Binomial Probabilities** Consider the binomial probabilities corresponding to n = 4 and p = 1/4 (or 0.25). If you attempt to use Table A-1 for finding the probabilities, you will find that the table does not apply to a probability of p = 1/4. Use Statdisk to find the probabilities and enter the results below.

$P(0)$ = _______________

$P(1)$ = _______________

$P(2)$ = _______________

$P(3)$ = _______________

$P(4)$ = _______________

5-6 **Binomial Probabilities** Assume that boys and girls are equally likely and 100 births are randomly selected. Use Statdisk with n = 100 and p = 0.5 to find $P(x)$, where x represents the number of girls among the 100 babies.

$P(35)$ = _______________

$P(45)$ = _______________

$P(50)$ = _______________

5-7 **Cumulative Probabilities** Assume that P(boy) = 0.5121, P(girl) = 0.4879, and that 100 births are randomly selected. Use Statdisk to find the probability that the number of girls among 100 babies is . . .

a. Fewer than 60 _______________

b. Fewer than 48 _______________

c. At most 30 _______________

d. At least 55 _______________

e. More than 40 _______________

5-8 **Identifying 0+** In Table A-1 from the textbook, the probability corresponding to n = 7, p = 0.95, and x = 2 is shown as 0+. Use Statdisk to find the corresponding probability and enter the result here. _______________

Statdisk

5-9 **Identifying 0+** In Table A-1 from the textbook, the probability corresponding to n =8, p = 0.01, and x = 6 is shown as 0+. Use Statdisk to find the corresponding probability and enter the result here. _______________

5-10. **Identifying a Probability Distribution** Use Statdisk to construct a table of x and $P(x)$ values corresponding to a binomial distribution in which n = 5 and p = 0.35. Enter the table in the space below.

Table:

Binomial *Exercises 5–11 through 5-15 involve binomial distributions. Use Statdisk for those exercises.*

5-11 **Brand Recognition** The brand name of Mrs. Fields (cookies) has a 90% recognition rate (based on data from Franchise Advantage). If Mrs. Fields herself wants to verify that rate by beginning with a small sample of ten randomly selected consumers, find the probability that exactly nine of the ten consumers recognize her brand name. Also find the probability that the number who recognize her brand name is not nine.

____

5-12 **Too Young to Tat** Based on a Harris poll, among adults who regret getting tattoos, 20% say that they were too young when they got their tattoos. Assume that 5 adults who regret getting tattoos are randomly selected, and find the indicated probability.

a. Find the probability that none of the selected adults say that they were too young to get tattoos. ___________
b. Find the probability that exactly one of the selected adults says that they were too young to get tattoos. ___________
c. Find the probability that the number of selected adults saying they were too young is 0 or 1. ___________

5-13 **Eye Color** In the United States, 40% of the population have brown eyes (based on data from Dr. P Sorita Soni at Indiana University). If 14 people are randomly selected, find the probability that at least 12 of them have brown eyes. Is it unusual to randomly select 14 people and find that at least 12 of them have brown eyes? Why or why not?

____

5-14 **Credit Rating** There is a 1% delinquency rate for consumers with FICO (Fair Isaac & Company) credit rating scores above 800. If the Jefferson Valley Bank provides large loans to 12 people with FICO scores above 800, what is the probability that at least one of them becomes delinquent? Based on that probability, should the bank plan on dealing with a delinquency?

____

5-15 **Genetics** Ten peas are generated from parents having the green/yellow pair of genes, so there is a 0.75 probability that an individual pea will have a green pod. Find the probability that among the 10 offspring peas, at least nine have green pods. Is it unusual to get at least nine peas with green pods when ten offspring peas are generated? Why or why not?

____

Poisson *Exercises 5-16 through 5-18 involve Poisson distributions. Use Statdisk for these exercises.*

5-16 **Radioactive Decay** Radioactive atoms are unstable because they have too much energy. When they release their extra energy, they are said to decay. When studying Cesium 137, it is found that during the course of decay over 365 days, 1,000,000 radioactive atoms are reduced to 977,287 radioactive atoms.

a. Find the mean number of radioactive atoms lost through decay in a day. _______
b. Find the probability that on a given day, 50 radioactive atoms decayed. _______

5-17 **Chocolate Chip Cookies** In the production of chocolate chip cookies, we can consider each cookie to be the specified interval unit required for a Poisson distribution, and we can consider the variable *x* to be the number of chocolate chips in a cookie. Table 3-1 is included with the Chapter Problem for Chapter 3 in the textbook, and it includes the numbers of chocolate chips in 34 different Keebler cookies. The Poisson distribution requires a value for μ, so use 30.4, which is the mean number of chocolate chips in the 34 Keebler cookies. Assume that the Poisson distribution applies.

a. Find the probability that a cookie will have 26 chocolate chips. ___________
b. Find the probability that a cookie will have 30 chocolate chips. ___________

5-18 **Deaths From Horse Kicks** A classic example of the Poisson distribution involves the number of deaths caused by horse kicks of men in the Prussian Army between 1875 and 1894. Data for 14 corps were combined for the 20-year period, and the 280 corps-years included a total of 196 deaths. After finding the mean number of deaths per corps-year, find the probability that a randomly selected corps-year has the following numbers of deaths.

a. 0 _____ b. 1 _____ c. 2 _____ d. 3 _____ e. 4 _____

The actual results consisted of these frequencies: 0 deaths (in 144 corps-years); 1 death (in 91 corps-years); 2 deaths (in 32 corps-years); 3 deaths (in 11 corps-years); 4 deaths (in 2 corps-years). Compare the actual results to those expected from the Poisson probabilities. Does the Poisson distribution serve as a good device for predicting the actual results?

____

____

6 Normal Distributions

Statdisk

6-1 Finding Areas and Values with a Normal Distribution

The Triola textbook describes methods for working with standard and nonstandard normal distributions. (A **standard normal distribution** has a mean of 0 and a standard deviation of 1.) Table A-2 in Appendix B of the textbook lists a wide variety of different *z* scores along with their corresponding areas. Statdisk's **Normal Probability** function can be used in place of Table A-2. Statdisk is much more flexible than the table, and it isn't limited to the values included in Table A-2.

Here is the procedure for using Statdisk's **Normal Probability** function.

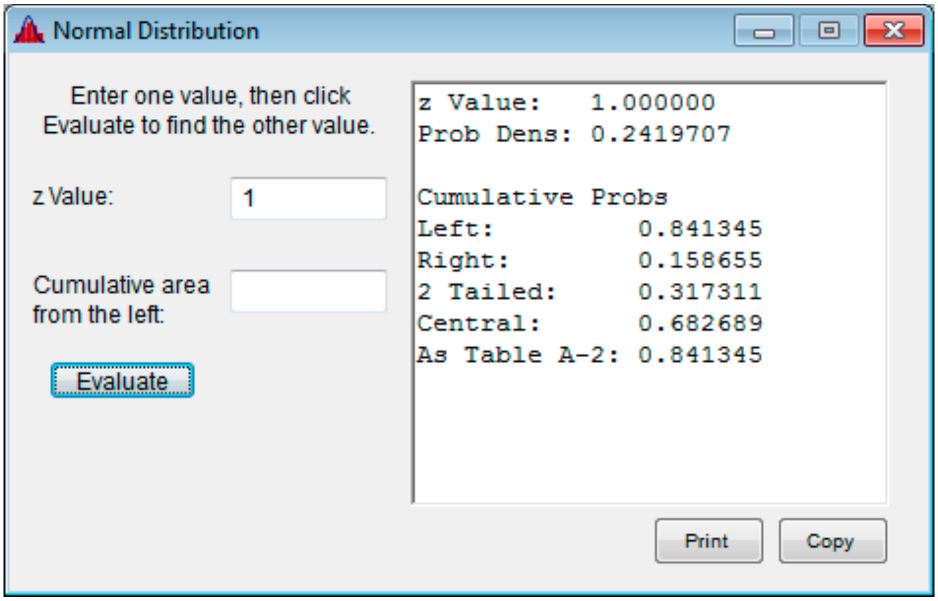
Statdisk Procedure for Finding Probabilities or *z* Scores with a Normal Distribution

1. Select **Analysis** from the menu at the top of the screen.
2. Select **Probability Distributions** from the subdirectory.
3. Select **Normal Distribution**.
4. Either enter the *z* score or enter the cumulative area from the left.
   - When working with a normal distribution that is nonstandard (with a mean different from 0 and/or a standard deviation different from 1) you must calculate the *z* score using the following formula:

$$z = \frac{x - \mu}{\sigma}$$

5. Click **Evaluate**.

If you enter a *z* score in Step 4, the display will include corresponding *areas*. If you enter the cumulative area from the left, the display will include the corresponding *z* score (along with other areas). For example, using the above procedure, enter a *z* score of 1 and click **Evaluate**. The screen display will be as shown below.



The following areas are included in the Statdisk display, and they correspond to the entry of z = 1.

Left	Area below the normal curve and to the left of z = 1:	0.841345
Right	Area below the normal curve and to the right of z = 1:	0.158655
2 Tailed	Twice the area in the tail bounded by z = 1:	0.317311
Central	Twice the area below the curve and bounded by the centerline and z = 1:	0.682689
As Table A-2	The area below the curve and to the *left* of z = 1: (The label "As Table A-2" indicates that the values are based on *cumulative areas from the left*, as in Table A-2.)	0.841345

The value shown for "*Prob Dens*" (probability density) is the height of the normal distribution curve for the value of z. The above display shows that when z = 1, the graph of the standard normal distribution has a height of 0.2419707. This particular value is not used in the textbook.

EXAMPLE Assume that IQ scores are normally distributed with a mean of 100 and a standard deviation of 15.

Finding Area: To find the area to the left of 115, first calculate the *z* score using the formula in step 4. Enter the z Value of 1.0 in the Normal Distribution Dialog and the **Left** result will be 0.843415, which is the area to the *left* of 115.

Finding a Value: To find the 90th percentile, enter 0.9 in the **Cumulative area from left** box. The z Value result will be 1.281553. The value (*x*) is calculated to be 119.223 using the formula below. This is the IQ score separating an area of 0.9 to its *left*.

$$x = z\sigma + \mu$$

6-2 Simulating and Generating Normal Data

We can learn much about the behavior of normal distributions by analyzing samples obtained from them. Sampling from real populations is often time consuming and expensive, but we can use the wonderful power of computers to obtain samples from theoretical normal distributions, and Minitab has such a capability, as described below.

Let's consider IQ scores. IQ tests are designed to produce a mean of 100 and a standard deviation of 15, and we expect that such scores are normally distributed. Suppose we want to learn about the variation of sample means for samples of IQ scores. Instead of going out into the world and randomly selecting groups of people and administering IQ tests, we can sample from theoretical populations. We can then learn much about the distribution of sample means. The following procedure allows you to obtain a random sample from a normally distributed population with a given mean and standard deviation.

Statdisk Procedure for Randomly Generating Sample Values from a Normally Distributed Population

1. Select **Data** from the menu at the top of the screen.

2. Select **Normal Generator** from the subdirectory.

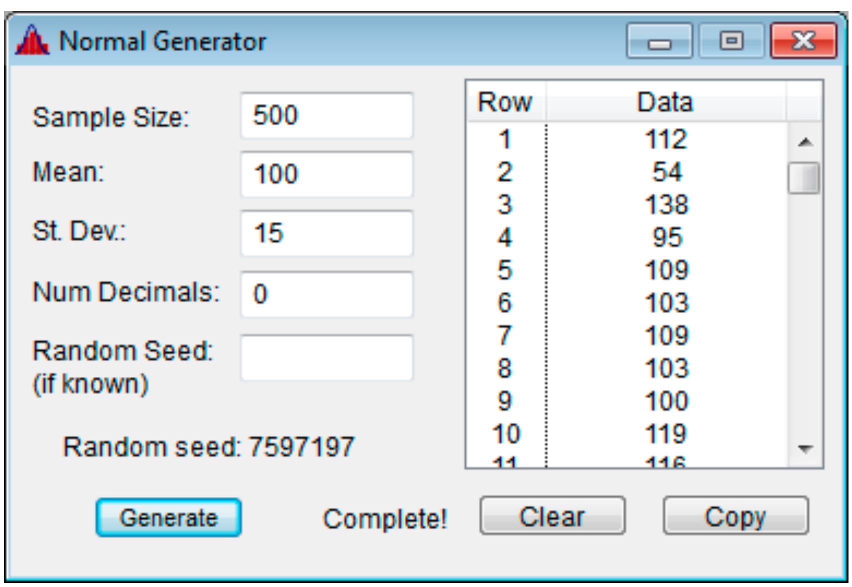
3. You will now get a dialog box, so enter the following.

- Enter the desired **sample size** (such as 500) for the number of values to be generated.
- Enter the desired **mean** (such as 100).
- Enter the desired **standard deviation** (such as 15).
- Enter the desired **number of decimal** places for the generated values.
- Enter a number for the Random Seed only if you want to *repeat* the generation of the specific data set. Otherwise, leave that box empty. (Leaving the Random Seed box empty causes a different data set to be randomly generated each time; using a specific seed number will generate the same data set each time.)

4. Click on the **Generate** button.

Shown below is the dialog box for generating 500 sample values (with 0 decimal places) from a normally distributed population with a mean of 100 and a standard deviation of 15. You can then click on **Copy** so that the 500 sample values can be copied to the Sample Editor window, where the values can be used for programs such as those for creating a histogram, boxplot, or calculating the descriptive statistics.

The result of this process is a collection of *sample* data randomly generated from a population with the specified mean and standard deviation, so the mean of the sample data might not be exactly the same as the value specified, and the standard deviation of the sample data might not be exactly the same as the value specified. The sample of IQ scores generated in this case has a mean of 99.51 and a standard deviation of 14.20.



Statdisk

6-3 The Central Limit Theorem

The textbook introduces the central limit theorem, which is then used in subsequent chapters when the important topics of estimating parameters and hypothesis testing are discussed. Here is the statement of the central limit theorem in the Triola textbook:

The Central Limit Theorem and the Sampling Distribution of $\bar{x}$

Given:

1. The random variable *x* has a distribution (which may or may not be normal) with mean μ and standard deviation σ.
2. Simple random samples all of the same size *n* are selected from the population. (The samples are selected so that all possible samples of size *n* have the same chance of being selected.)

Conclusions:

1. The distribution of sample means $\bar{x}$ will, as the sample size increases, approach a *normal* distribution.
2. The mean of all sample means is the population mean μ.
3. The standard deviation of all sample means is $\sigma/\sqrt{n}$.

Practical Rules Commonly Used

1. If the original population is *not normally distributed*, here is a common guideline: For *n* > 30, the distribution of the sample means can be approximated reasonably well by a normal distribution. (There are exceptions, such as populations with very nonnormal distributions requiring sample sizes larger than 30, but such exceptions are relatively rare.) The distribution of sample means gets closer to a normal distribution as the sample size *n* becomes larger.

2. If the original population is *normally distributed*, then for *any* sample size *n*, the sample means will be normally distributed.

In the study of methods of statistical analysis, it is extremely helpful to have a clear understanding of the statement of the central limit theorem. See Experiment 6-3 in this workbook. That experiment should show this: Even though the samples are obtained from a *uniform* distribution (outcomes of rolling a die), as the sample size increases from 1 to 2 to 10 to 20, the distribution of the sample means gets closer to a *normal* distribution.

6–4 Assessing Normality

Textbooks in the Triola statistics series discuss criteria for determining whether sample data appear to come from a population having a normal distribution. These criteria are listed:

1. **Histogram:** Construct a Histogram. Reject normality if the histogram departs dramatically from a bell shape. Statdisk can generate a histogram. Section 2-2 of this manual/workbook gives the Statdisk procedure for generating a histogram.

2. **Outliers:** Identify outliers. Reject normality if there is more than one outlier present. (Just one outlier could be an error or the result of chance variation, but be careful, because even a single outlier can have a dramatic effect on results.) Using Statdisk, we can sort the data and easily identify any values that are far away from the majority of all other values. Statdisk-generated boxplots can also be used to identify potential outliers. Section 3-3 of this manual/workbook provides the Statdisk procedure for generating a "modified boxplot." Modified boxplots are described in Part 2 of Section 3-4 in the textbook. In a Statdisk-generated boxplot, points identified with asterisks are potential outliers.

3. **Normal Quantile Plot:** If the histogram is basically symmetric and there is at most one outlier, construct a *normal quantile plot.* Examine the normal quantile plot and reject normality if the points do not lie close to a straight line, or if the points exhibit some systematic pattern that is not a straight-line pattern.

Statdisk includes the feature of **Normality Assessment** that allows you to address each of the above three criteria in the following simple procedure.

Statdisk Procedure for Normality Assessment

1. Put the list of sample data in a column of the Sample Editor window. Either manually enter the data or open a data set that had been previously saved, such as any of the data sets in Appendix B.

2. Click on the top menu item of **Data**.

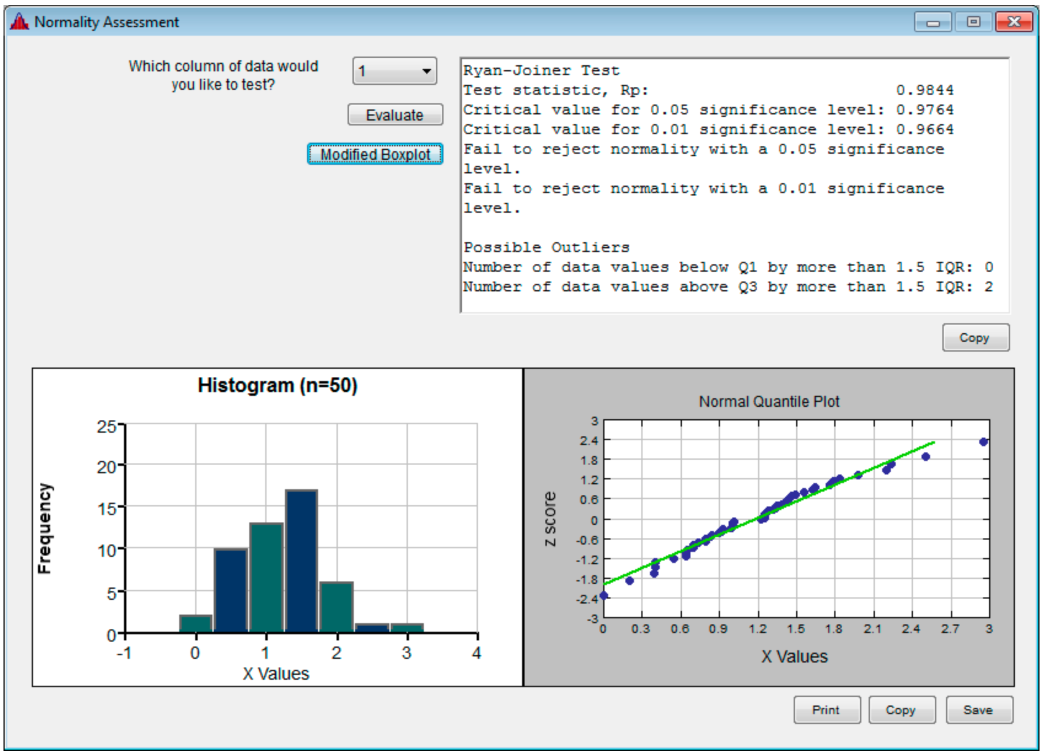
3. Click on the menu item of **Normality Assessment**.

4. Select the column containing the sample data that is to be analyzed.

5. Click on the **Evaluate** button.

EXAMPLE ***Earthquake Magnitudes*** As an example, consider the 50 earthquake magnitudes listed in the *Elementary Statistics 12th Edition* data set *16 – Earthquake Measurements*.

After opening this data set, the 50 earthquake magnitudes are listed in column 1 of the Sample Editor window. If you use the above Statdisk procedure for normality assessment, the Statdisk display will appear as shown below.



We now evaluate the results included in the above display.

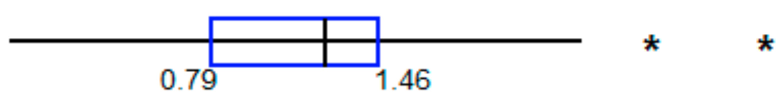
Histogram: The above Statdisk display includes a histogram showing that the shape of the distribution is approximately bell-shaped, and this suggests that the sample data are from a normally distributed population.

Outliers: The display includes the number of "Possible Outliers." In this case, we see that there are two potential outliers consisting of two values that are above the third quartile Q_3 by more than 1.5 times the IQR (interquartile range). If we examine the sorted list of sample values, we see that the highest values in the sorted list are 2.95, 2.50, 2.24, 1.98, 1.83, and so on. The potential outliers of 2.95 and 2.50 do not appear to be very far away from the other sample values, so these potential outliers do not appear to be actual outliers. Because there are only two *potential* outliers and they are not too far away from the other values, we should not reject normality because of outliers.

Normal Quantile Plot: The above display includes a normal quantile plot. (Normal quantile plots are discussed in Section6-6 of the textbook. Because the normal quantile plot in the above display shows a pattern of points that is reasonably close to a straight-line pattern and because there is no other systematic pattern that is not a straight-line pattern, the normal quantile plot suggests that the sample data are from a population with a normal distribution.

Ryan-Joiner Test: The Ryan-Joiner test is discussed briefly in Part 2 of Section 6-6 of the textbook in the Triola Statistics Series. It is one of several formal tests of normality, each having their own advantages and disadvantages. The above Statdisk display shows that the Ryan-Joiner test results in a conclusion of “fail to reject normality.” That is, it appears that the sample data are from a population having a normal distribution.

Modified Boxplot: The above Statdisk display shows that there is a button for generating a modified boxplot. (Modified boxplots are described in Part 2 of Section 3-4 in the textbook.) If we click on Statdisk’s **Modified Boxplot** button, we get the following graph. This modified boxplot includes the two asterisks, which are the potential outliers of 2.95 and 2.50. Refer to the discussion in the above in the paragraph with the heading of **Outliers**. Based on that discussion, we do not consider the potential outliers of 2.95 and 2.50 to be actual outliers. Also, the boxplot does not exhibit a lack of symmetry or lopsidedness that is too extreme.



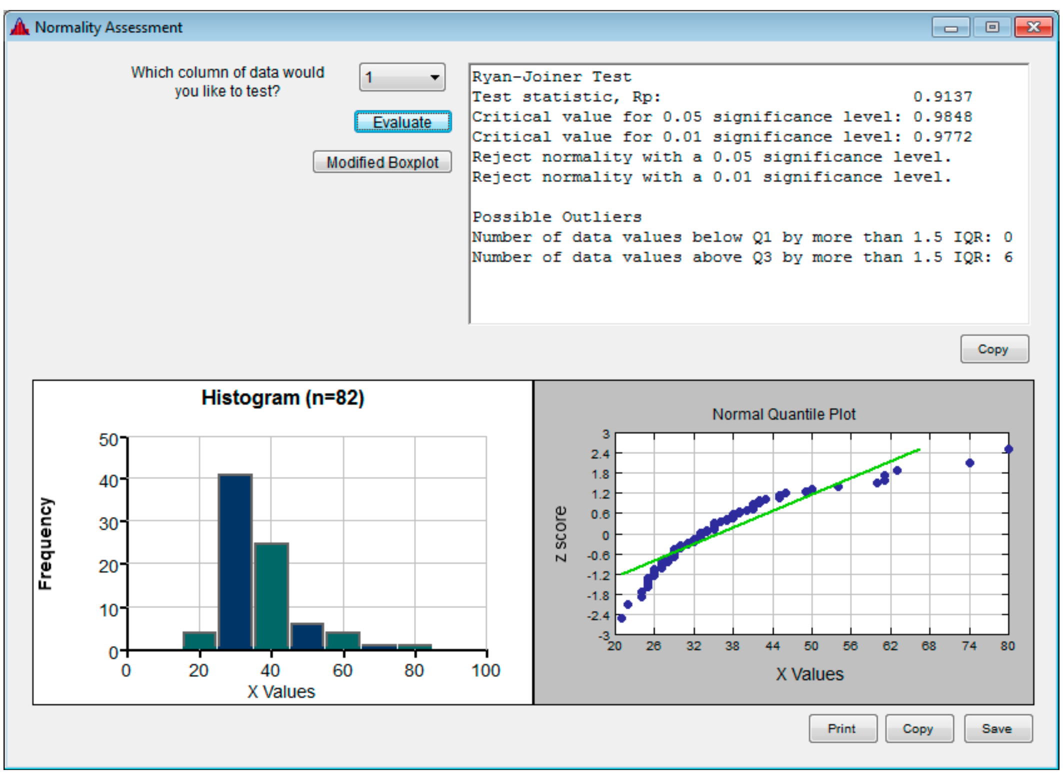
Conclusion: After considering all of the above elements from the single Statdisk display on the preceding page, we conclude that the sample appears to be from a population with a normal distribution.

EXAMPLE ***Age of Oscar Winning Actresses*** Consider the ages of 82 actresses at the time they won Oscars. These data are available in the Appendix B of the textbook (and in Statdisk) as data set *11 – Ages of Oscar Winners*.

Shown below is the Statdisk display of the *Normality Assessment* function.

- The histogram does not appear bell–shaped, which suggests that the actress ages do not have a normal distribution.
- The points in the normal quantile plot show a systematic pattern that is not a straight-line pattern, suggesting that the ages are not from a normally distributed population. There are 6 potential outliers, and examining the sorted list of 82 actress ages, we see that there appear to be five outliers. The presence of multiple outliers suggests that the data are not from a normally distributed population. The results from the Ryan-Joiner test suggest that we should reject normality.

Based on all of these factors, we conclude that the sample of actress ages appears to come from a population having a distribution that is *not* a normal distribution. If some particular method of statistical analysis requires that sample data are from a normally distributed population, the actress ages considered here do not satisfy that requirement.



6-5 Normal Approximation to Binomial

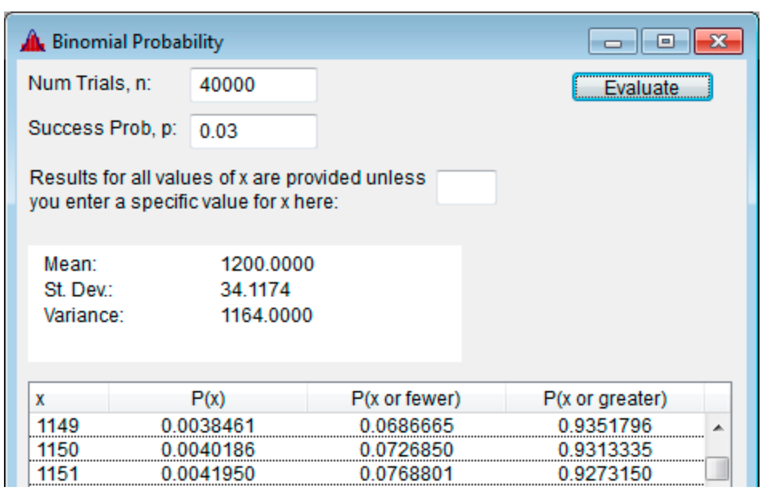
Section 6-6 in the Triola textbook discusses the topic of using a normal distribution as an approximation to a binomial distribution. That section includes this comment:

> *Note:* Instead of using a normal distribution as an approximation to a binomial probability distribution, most practical applications of the binomial distribution can be handled with computer software or a calculator.

When working with applications involving a binomial distribution, Statdisk can be used to find *exact* results, so there is no need to approximate the binomial distribution with a normal distribution. Consider Example 1 in Section 6-6 of the textbook:

> **EXAMPLE** The author was mailed a survey from Viking River Cruises, and the survey included a request for an e-mail address. Assume that the survey was sent to 40,000 people and that for such surveys, the percentage of responses with an e-mail address is 3%. If the true goal of the survey was to acquire a bank of at least 1150 e-mail addresses, find the probability of getting at least 1150 responses with e-mail addresses.

Based on the above statements from Example 1 in the textbook, we have n = 40,000 and p = 0.03, and we want to find $P(x \geq 1150)$, which is the probability of getting at least 1150 responses with e-mail addresses. Using Statdisk with the binomial distribution procedure described in Section 5-2 of this workbook, we can find that the *exact* probability of 1150 or greater responses with e-mail addresses is 0.9313335 as shown in the display below. This is a more accurate result than the result of 0.9306 found by using the normal distribution as an approximation to the binomial distribution. For Statdisk users, the method of approximating a binomial distribution with a normal distribution is generally obsolete.



CHAPTER 6 EXPERIMENTS: Normal Distributions

6-1 **Finding Probabilities for a Normal Distribution** Use Statdisk's *Normal Distribution* module to find the indicated probabilities. First select **Analysis** from the top menu bar, then select **Probability Distributions**, then **Normal Distribution**.

a. Given a population with a normal distribution, a mean of 0, and a standard deviation of 1, find the probability of a value less than 0.75.________________

b. Given a population with a normal distribution, a mean of 100, and a standard deviation of 15, find the probability of a value less than 83.________________

c. Given a population with a normal distribution, a mean of 120, and a standard deviation of 10, find the probability of a value greater than 135.______________

d. Given a population with a normal distribution, a mean of 98.6, and a standard deviation of 0.62, find the probability of a value between 97.0 and 99.0.__________

e. Given a population with a normal distribution, a mean of 150, and a standard deviation of 12, find the probability of a value between 147 and 160.__________

6-2. **Finding Values for a Normal Distribution** Use Statdisk's *Normal Distribution* module to find the indicated values. First select **Analysis** from the menu, then select **Probability Distributions**, then **Normal Distribution**.

a. Given a population with a normal distribution, a mean of 0, and a standard deviation of 1, what value has an area of 0.4 to its left?_____________________

b. Given a population with a normal distribution, a mean of 100, and a standard deviation of 15, what value has an area of 0.3 to its right?_________________

c. Given a population with a normal distribution, a mean of 85, and a standard deviation of 8, what value has an area of 0.95 to its left?____________________

d. Given a population with a normal distribution, a mean of 45.5, and a standard deviation of 3.5, what value has an area of 0.88 to its right?___________________

e. Given a population with a normal distribution, a mean of 94, and a standard deviation of 6, what value has an area of 0.01 to its left?____________________

6-3 **Central Limit Theorem** In this experiment, assume that all dice have 6 sides.

a. Use Statdisk to simulate the rolling of a single die 800 times. (Select **Data**, then **Dice Generator**.) Use Copy/Paste to copy the results to the Descriptive Statistics and Histogram modules, and enter the actual results below.

One Die: Mean: ___________
Standard Deviation: ___________
Distribution shape: ___________

b. Part (a) used a single die, but we will now use a pair of dice. Use Statdisk to "roll" two dice 800 times (again using Data/Dice Generator). The 800 values are *totals* for each pair of dice, so transform the totals to *means* by dividing each total by 2. (Use Copy/Paste to copy the results to the Sample Editor window, then use Statdisk's **Sample Transformations** feature to divide each value by 2. To divide each value by 2, select the operation of / and use a constant of 2.) Now use Copy/Paste to copy the 800 *means* to the Sample Editor window, then use the Descriptive Statistics and Histogram modules. Enter the results below.

Two Dice: Mean: ___________
Standard Deviation: ___________
Distribution shape: ___________

c. Repeat part (b) using 10 dice. When finding the mean of the 10 dice, divide each value by 10.

10 Dice: Mean: ___________
Standard Deviation: ___________
Distribution shape: ___________

d. Repeat part (b) using 20 dice.

20 Dice: Mean: ___________
Standard Deviation: ___________
Distribution shape: ___________

e. General conclusions:
What happens to the mean as the sample size increases from 1 to 2 to 10 to 20?

____

Statdisk

What happens to the standard deviation as the sample size increases?

____

What happens to the distribution shape as the sample size increases?

____

How do these results illustrate the central limit theorem?

____

____

6-4 **Identifying Significance** People generally believe that the mean body temperature is 98.6^{o}F. Appendix B from the textbook includes data set *3 – Body Temperatures of Healthy Adults,* which includes a sample of 106 body temperatures with these properties: The distribution is approximately normal, the sample mean is 98.20^{o}F, and the standard deviation is 0.62^{o}F. We want to determine whether these sample results differ from 98.6^{o}F by a *significant* amount. One way to make that determination is to study the behavior of samples drawn from a population with a mean of 98.6.

a. Use Statdisk to generate 106 values from a normally distributed population with a mean of 98.6 and a standard deviation of 0.62. Use Statdisk to find the mean of the generated sample. Record that mean here:__________

b. Repeat part (a) nine more times and record the 10 sample means here:

____

c. By examining the 10 sample means in part (b), we can get a sense for how much sample means vary for a normally distributed population with a mean of 98.6 and a standard deviation of 0.62. After examining those 10 sample means, what do you conclude about the likelihood of getting a sample mean of 98.20? Is 98.20 a sample mean that could easily occur by chance, or is it significantly different from the likely sample means that we expect from a population with a mean of 98.6?

____

d. Given that researchers did obtain a sample of 106 temperatures with a mean of 98.20^{o}F, what do their results suggest about the common belief that the population mean is 98.6^{o}F?

____

6-5. **Identifying Significance** This experiment involves the Statdisk data set *19 – Cola Weights and Volumes,* which can be opened by clicking on **Data Sets** on the top menu bar and then selecting *Elementary Statistics 12th Edition*.

a. Open the Cola data set and find the mean and standard deviation of the sample consisting of the volumes of cola in cans of regular Coke. Enter the results here. Sample mean:____________ Standard deviation:____________

b. Generate 10 different samples, where each sample has 36 values randomly selected from a normally distributed population with a mean of 12 oz and a standard deviation of 0.115 oz (based on the claimed volume printed on the cans and the data in the Cola data set). For each sample, record the sample mean and enter it here.

____

c. By examining the 10 sample means in part b, we can get a sense for how much sample means vary for a normally distributed population with a mean of 12 and a standard deviation of 0.115. After examining those 10 sample means, what do you conclude about the likelihood of getting a sample mean like the one found for the sample volumes in the Cola data set? Is the mean for the sample a value that could easily occur by chance, or is it significantly different from the likely sample means that we expect from a population with a mean of 12?

____

d. Consider the sample mean found from the volumes of regular Coke listed in the Cola data set. Does it suggest that the population mean of 12 oz (as printed on the label) is not correct?

____

Assesing Normality *In Experiments 6-6 through 6-10, refer to the indicated Statdisk data set. In each case, determine whether the sample data appear to come from a normally distributed population and give reasons explaining your conclusion. All data sets are from Appendix B in the Triola textbook (and are available in Statdisk).*

6-6 **Earthquake Depth**. The depth (km) measurements of earthquakes recorded in one year from a location in Southern California, as listed in the data set (12th Edition) *16 – Earthquake Measurements*.

6-7 **President Heights** The heights of U.S. Presidents, as listed in the data set (12th Edition) *18 – Voltage Measurements from a Home*.

6-8 **Oscar Winning Actor Ages** The ages of actors at the times that they won Oscars for Best Actor, as listed in the data set (12th Edition) *11 – Ages of Oscar Winners.*

6-9 **Generator Voltage** The measured voltage levels from a home generator, as listed in the data set (12th Edition) *18 – Voltage Measurements from a Home*.

7 Confidence Intervals and Sample Sizes

Statdisk

7-1 Confidence Intervals for Estimating p

Section 7-2 of the Triola textbook presents methods for using a sample proportion to estimate a population proportion, and a confidence interval is a common and important tool used for that purpose. A **confidence interval** (or **interval estimate**) is a range (or an interval) of values used to estimate the true value of a population parameter.

When finding a confidence interval estimate of a population proportion *p*, Statdisk requires the sample size *n* and the number of successes *x*. In some cases, the values of *x* and *n* are both known, but in other cases the given information may consist of the sample size *n* and a sample percentage. For example, suppose we know that among 1007 people surveyed, 85% said that they know what Twitter is. Based on that information, we know that *n* = 1007 and $\hat{p}$ = 0.85. Finding the value of *x* is quite simple: 85% of 1007 is 0.85 × 1007 = 856 (rounded to a whole number). (Because $\hat{p} = x/n$, it follows that $x = \hat{p}\,n$, so the number of successes can be found by multiplying the sample proportion $\hat{p}$ and the sample size *n*.)

To find the number of successes *x* from the sample proportion and sample size:
Calculate $x = \hat{p} \cdot n$ and round the result to the nearest whole number.

After having determined the value of the sample size *n* and the number of successes *x*, we can proceed to use Statdisk as follows.

Statdisk Procedure for Finding Confidence Intervals for *p*

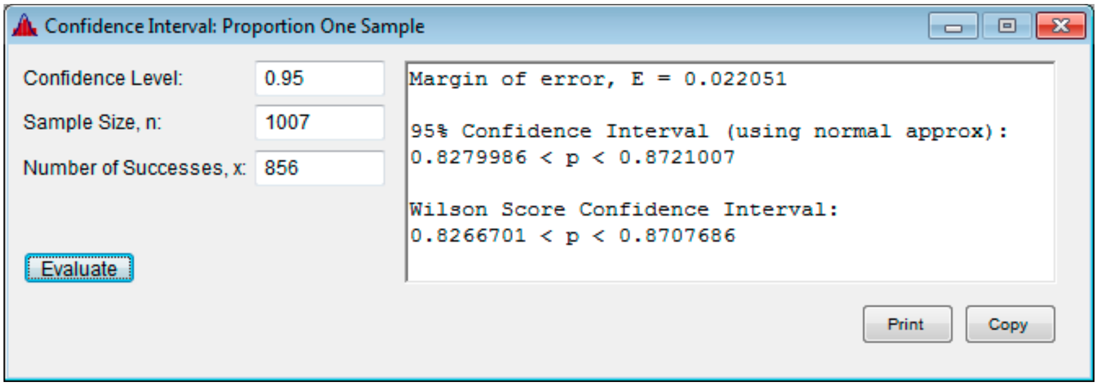
1. Select **Analysis** from the top menu bar.

2. Select **Confidence Intervals**.

3. Select **Proportion One Sample**.

4. Make these entries in the dialog box:
    - Enter a confidence level, such as 0.95 or 0.99.
    - Enter the value for the sample size *n.*
    - Enter the number of successes for *x*.

5. Click on the **Evaluate** button and the Statdisk results are as shown below.



Based on the above Statdisk display, we can express the 95% confidence interval estimate of p as $0.8279986 < p < 0.87210007$. After rounding, the confidence interval becomes $0.828 < p < 0.872$. This can also be expressed as $82.8\% < p < 87.2\%$ or as $85.0\% \pm 2.2\%$. (The Wilson Score confidence interval is discussed briefly in the textbook near the end of Section 7-2.)

7-2 Confidence Intervals for Estimating μ

Section 7-3 in the textbook introduces a method for using a sample mean $\bar{x}$ to estimate the value of a population mean μ.

Part 1 of Section 7-3 in the textbook focuses on the case in which the population standard deviation σ is not known, as is usually the case. Part 2 of Section 7-3 in the textbook discusses the case in which σ is known.

The textbook stresses the importance of selecting the correct distribution (normal or *t*), but Statdisk automatically chooses the correct distribution based on the information that is entered. Statdisk is very easy to use for constructing confidence interval estimates for a population mean μ. However, Statdisk requires that you first obtain the descriptive statistics of n, $\bar{x}$, and s, as indicated in Step 1 of the following procedure.

Statdisk Procedure for Finding Confidence Intervals for μ

1. If the sample data are known but n, $\bar{x}$, and s are not yet known, find the values of those sample statistics by using Statdisk's Explore Data or Descriptive Statistics features. (See Section 3-1 of this workbook.) You must know the values of n, $\bar{x}$, and s before proceeding to step 2.

2. Select **Analysis** from the menu at the top of the screen.

3. Select **Confidence Intervals** from the subdirectory.

4. Select **Mean – One Sample.**

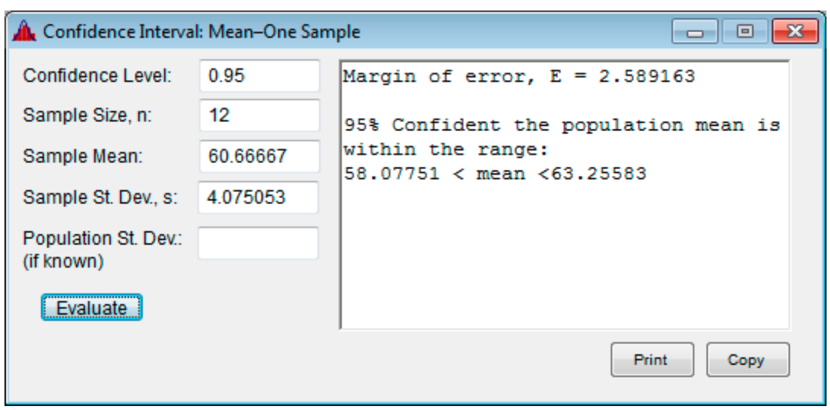
5. You will now see a dialog box allowing you to make the following entries.
   - Enter a Confidence Level, such as 0.95 or 0.99.
   - Enter the Sample Size *n*.
   - Enter the value of the sample mean $\overline{x}$.
   - Enter the value of the sample standard deviation *s*.
   - Enter the value of the *population* standard deviation σ if it is known. (The value of σ is usually unknown, so you will usually leave this entry blank.)

6. Click the **Evaluate** button.

Consider Example 2 in Section 7-3 of the textbook

> **EXAMPLE** The following speeds (mi/h) are measured from southbound traffic on I-280 near Cupertino, California (based on data from SigAlert).
> 62 61 61 57 61 54 59 58 59 69 60 67

The following screen shows the results obtained using the 12 speeds listed above. (The sample mean and sample standard deviation were obtained using Statdisk's Explore Data feature as described in Step 1.)



Based on the above Statdisk display, the 95% confidence interval is 58.1 < μ < 63.3 (rounded). The solution in the textbook includes this statement: We are 95% confident that the limits of 58.1 mi/h and 63.3 mi/h actually do contain the value of the population mean μ.

7-3 Confidence Intervals for Estimating σ

Section 7-4 of *Elementary Statistics*, 12th edition, and some other Triola textbooks describe the procedure for constructing a confidence interval for estimating a population standard deviation σ or variance σ^2. After selecting a confidence level and entering the sample size *n* and sample standard deviation *s*, Statdisk will automatically provide a confidence interval estimate of σ along with a confidence interval estimate of σ^2. You get both confidence intervals (for σ and for σ^2), whether you want them or not. Be careful to correctly identify the value of the sample standard deviation *s.* Be careful to enter the sample standard deviation where it is required; if only the sample variance is known, find its square root and enter that value for *s*. After obtaining the values of the sample size *n* and sample standard deviation *s*, proceed with the following Statdisk procedure.

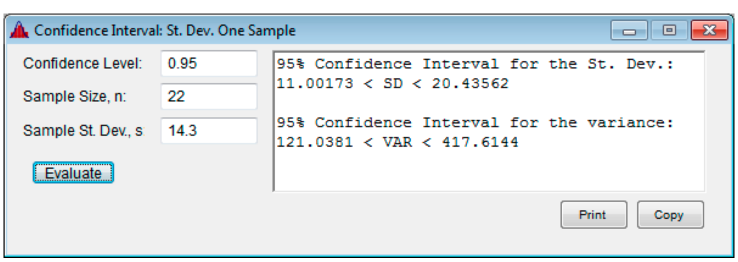
Statdisk Procedure for Finding Confidence Intervals for σ and σ^2

1. Select **Analysis** from the top menu bar.
2. Select **Confidence Intervals**.
3. Select **St Dev One Sample**
4. Make the following entries in the dialog box:
    - Enter a confidence level, such as 0.95 or 0.99.
    - Enter the value for the sample size *n.*
    - Enter the value of the sample *standard deviation s* (not the *variance* s^2)
5. Click on the **Evaluate** button.

Consider Example 2 from Section 7-4 in the textbook.

> **EXAMPLE** A sample of size *n* = 22 has standard deviation *s* = 14.3, and we want to construct a 95% confidence interval estimate of *σ.*

Because we know the values of *n* = 22 and *s* = 14.3, we can use the above procedure to obtain the 95% confidence interval estimate of the population standard deviation: 11.0 < σ < 20.4 (rounded). The Statdisk display is shown.



7-4 Sample Sizes for Estimating *p*

Section 7-2 of the Triola textbook describes methods for determining the *sample size* needed to estimate a population proportion *p*. Statdisk requires that you enter a confidence level (such as 0.95) and a margin of error *E* (such as 0.03). In addition to those two required entries, there are two optional entries. You can enter an estimate of *p* if one is known, based on such factors as prior knowledge or results from a previous study. You can also enter the population size *N* if it is known and if you are sampling without replacement. The textbook includes the following two cases.

When an estimate $\hat{p}$ is known: **Formula 7-2** $n = \dfrac{[z_{\alpha/2}]^2 \hat{p}\hat{q}}{E^2}$

When no estimate $\hat{p}$ is known: **Formula 7-3** $n = \dfrac{[z_{\alpha/2}]^2 0.25}{E^2}$

Statdisk Procedure for Finding Sample Sizes Required to Estimate *p*

1. Select **Analysis** from the top menu bar.

2. Select the subdirectory item of **Sample Size Determination**.

3. Select **Estimate Proportion**.

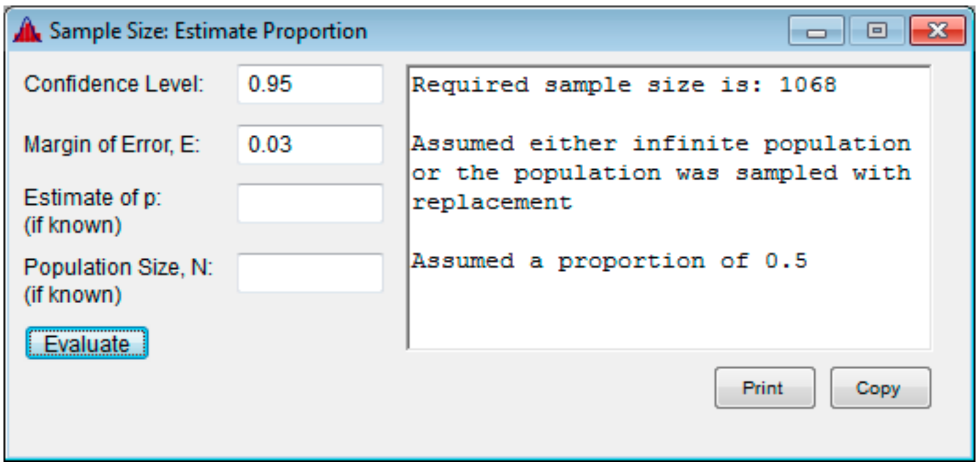
4. Make these entries in the dialog box:
   - Enter a confidence level, such as 0.95 or 0.99.
   - Enter a margin of error *E*. (*Hint:* The margin of error must be expressed in decimal form. For example, a margin of error of "three percentage points" should be entered as 0.03.)
   - Enter an estimated proportion if it is known. (This value might come from a previous study, or from knowledge about the value of the sample proportion. If such a value is not known, leave this box empty.)
   - Enter a value for the population size *N* if you will sample without replacement from a finite population of *N* subjects. (If the population is large or sampling is done with replacement, leave this box blank.)

5. Click the **Evaluate** button.

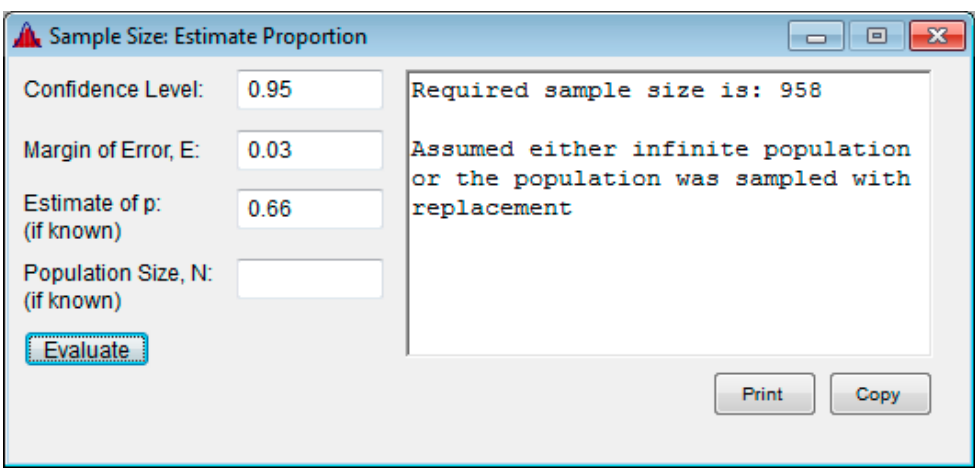
Consider Example 4 from Section 7-2 in the textbook

EXAMPLE: How many adults must be surveyed in order to be 95% confident that the sample percentage is in error by no more than three percentage points.

Using the above Statdisk procedure, we find that a simple random sample of 1068 adults is needed as shown in the display below.



Example 4 from Section 7-2 also asks how many adults must be surveyed in order to be 95% confident that the sample percentage is in error by no more than three percentage points when the estimate if estimated proportion is 66%. Using the above Statdisk procedure, we find that a simple random sample of 958 adults is needed as shown in the display below.



7-5 Sample Sizes for Estimating μ

Section 7-3 of the Triola textbook discusses the procedure for determining the sample size necessary to estimate a population mean μ. We use the formula

$$n = \left[\frac{z_{\alpha/2}\sigma}{E}\right]^2$$

Statdisk requires that we know the desired degree of confidence, the margin of error *E*, and the population standard deviation σ. The textbook notes that it is unusual to know σ without knowing μ, but σ might be known from a previous study or it might be estimated from a pilot study or the range rule of thumb. The entry of a finite population size *N* is optional, as described in the following steps.

Statdisk Procedure for Finding Sample Sizes Required to Estimate μ

1. Select **Analysis** from the top menu bar.

2. Select **Sample Size Determination** from the subdirectory.

3. Select the option of **Estimate Mean**.

4. In the dialog box, make these entries:

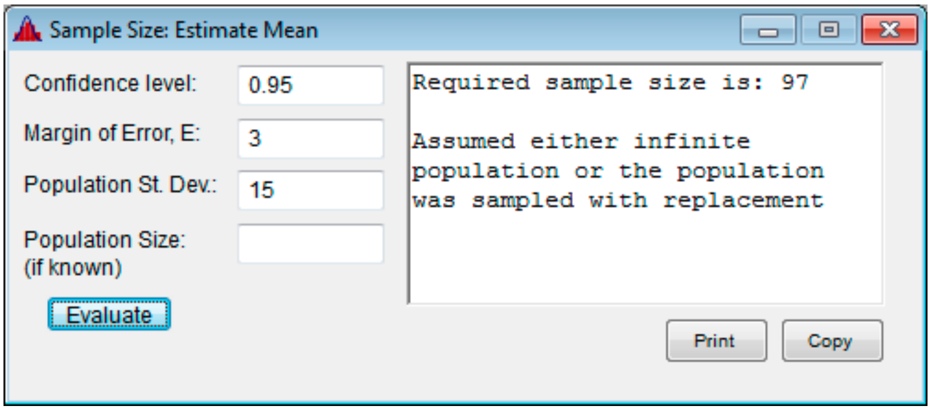
   - Enter a confidence level, such as 0.95 or 0.99.
   - Enter a margin of error *E*. (*Hint:* A margin of error of "three percentage points" should be entered as 3.0 in this dialog.)
   - Enter the value of the population standard deviation σ. (If σ is not known, consider estimating it from a previous study or pilot study or use the range rule of thumb.)
   - For the entry box labeled **Population Size**, leave it blank if you are sampling with replacement, or if you have a small sample drawn from a large population. (Consider a sample size *n* to be "small" if $n \leq 0.05N$.) Enter a value only if you are sampling without replacement from a finite population with known size *N*, and the sample is large so that $n > 0.05N$. This box is usually left blank.

5. Click on the **Evaluate** button.

Consider Example 6 in Section 7-3 of the textbook.

> **EXAMPLE** How many students must be sampled for IQ tests if we want 95% confidence that the sample mean is within 3 IQ points the population mean and the population standard deviation is assumed to be 15.

Using the above procedure, the Statdisk display below shows that the required sample size is 97.



7-6 Sample Sizes for Estimating σ

The textbook describes a procedure for determining the sample size required to estimate a population standard deviation σ or population variance σ^2. (See Section 7-4 in *Elementary Statistics*, 12th edition.) Table 7-2 in the textbook lists sample sizes for several different cases, but Statdisk is much more flexible and allows you to find sample sizes for many other cases.

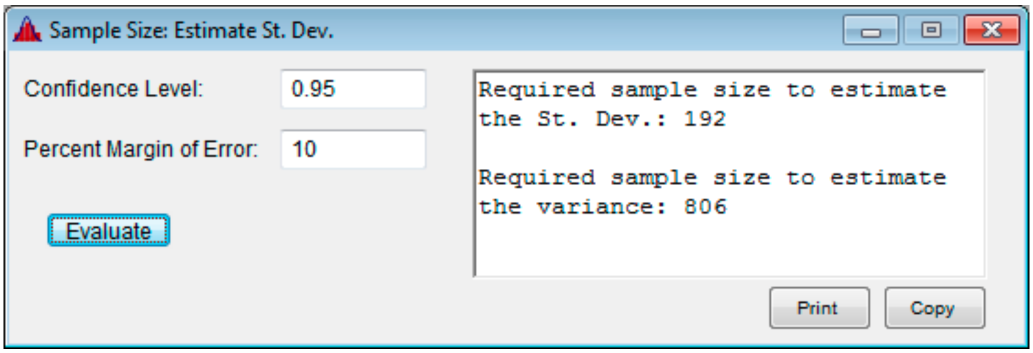
Statdisk Procedure for Finding Samples Sizes Required to Estimate σ or σ^2

1. Select **Analysis** from the top menu bar.
2. Select **Sample Size Determination**.
3. Select **Estimate St. Dev**.
4. Make these entries in the dialog box:
   - Enter a confidence level, such as 0.95 or 0.99.
   - Enter a "Percent Margin of Error." (For example, if you enter 20, Statdisk will provide the sample size required so that *s* is within 20% of σ; and it will also provide the sample

size required so that s^2 is within 20% of σ^2.)

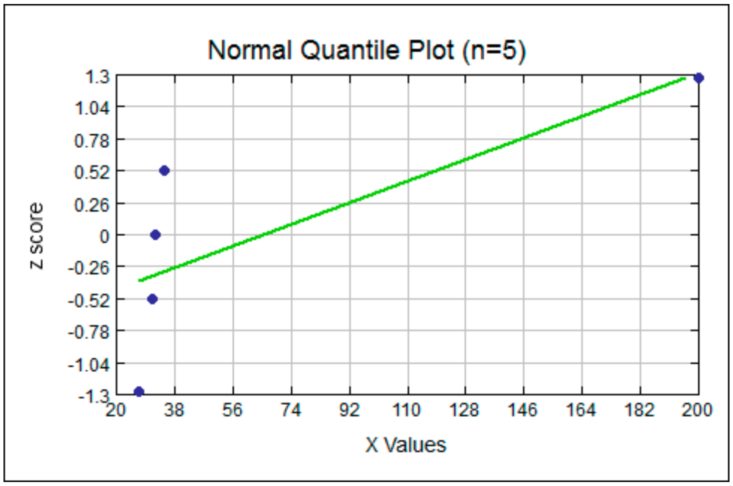
5. Click the **Evaluate** button.

Statistics textbooks tend to omit discussions about the issue of determining sample size for estimating a population standard deviation σ or variance σ^2, but Statdisk allows you to do these calculations with ease. For example, to be 95% confident that our estimate is within 10% of the true value of σ, a sample size of 192 is needed as shown in the Statdisk display below.



7-7 Bootstrap Resampling

Suppose that we have sample data consisting of the values 27, 31, 32, 35, and 200, and we want to use those values to construct a confidence interval estimate of the population mean. That small sample of five values includes the value of 200, which is an outlier. Shown below is the normal quantile plot of these values. Because the points do not fit a straight-line pattern reasonably well, we conclude that the data values are from a distribution that is not normal.



The sample is small, the σ is unknown, and the distribution is not normal, so the methods of Chapter 7 in the textbook cannot be used to construct a confidence interval estimate of the population mean. One alternative is to use the method of **bootstrap resampling**, which does not require normally distributed data. With bootstrap resampling of 5 sample values, we randomly select 5 of those values, but we sample with replacement. Then we repeat that selection process many times, thereby pulling the original sample up "by its own bootstraps."

Here is a bootstrap resampling procedure for using a small sample (from a non-normal population) to find a 95% confidence interval estimate of the population mean:

1. Generate a large number of new samples by randomly selecting values (with replacement) from the original sample.

2. Find the means of the generated samples.

3. Sort the sample means.

4. Construct a 95% confidence interval estimate of the population mean μ by finding the percentiles $P_{2.5}$ and $P_{97.5}$ for the list of sorted means. Those percentile values are the confidence interval limits.

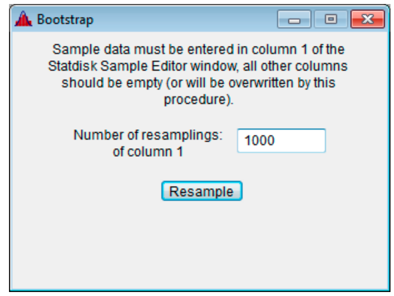
Statdisk

Statdisk Procedure for Bootstrap Resampling

1. Enter the sample data in column 1 of the Sample Editor window.

2. Select **Analysis**, then select the menu item of **Bootstrap Resampling**.

3. Enter the desired number of samples to be generated, such as 1000, then click **Resample**.
   - The sample means will be inserted in column 2 of the Sample Editor window, and the sample standard deviations will be inserted in column 3.



4. Click on **Data** and select **Sort Data**.
   - For Sort, select **One Column**.
   - Sort column 2 and then sort column 3. Sort columns using the **A to Z** selection.

5. Construct a 95% confidence interval estimate of the population mean μ by finding the percentiles $P_{2.5}$ and $P_{97.5}$ for the list of sorted means. (With 1000 sorted sample means, $P_{2.5}$ is the mean of the 25th and 26th sample means, and $P_{97.5}$ is the mean of the 975th and 976th sample means.) Those percentile values are the confidence interval limits. A confidence interval estimate of σ can also be found by using that same basic procedure of finding the percentiles $P_{2.5}$ and $P_{97.5}$ using the sorted liste of standard deviations in column 3.

Using this procedure with the sample values of 27, 31, 32, 35, and 200, the 95% confidence interval of 29.4 < μ < 132.8 was obtained. (Because the confidence interval obtained through bootstrap resampling is based on randomly generated values, different results will be obtained each time the bootstrap resampling method is used.)

If the methods of Section 7-4 in the textbook are used with the same sample data of 27, 31, 32, 35, and 200, the confidence interval of -28.7 < μ < 158.8 is obtained. Because the methods of Section 7-4 should not be used (because the requirement of a normal distribution is not met), this latter confidence interval should not be used as a reasonable estimate of the population mean. A confidence interval obtained through bootstrap resampling is likely to be a better estimate of the population mean.

CHAPTER 7 EXPERIMENTS: Confidence Intervals and Sample Sizes

In Experiments 7–1 through 7–4, use Statdisk with the sample data and confidence level to construct the confidence interval estimate of the population proportion p.

7-1 From a KRC Research poll in which respondents were asked if they felt vulnerable to identify theft: n = 1002, x = 531 who said "yes". Use a 95% confidence level. ____

7-2 From a 3M Privacy Filters poll in which respondents were asked to identify their favorite seat when they fly: n = 806, x = 492 who chose the window seat. Use a 99% confidence level. ____

7-3 From a Prince Market Research poll in which respondents were asked if the acted to annoy a bad driver: n = 2518, x = 1083 who said that they honked. Use a 90% confidence level. ____

7-4 From an Angus Reid Public Opinion poll in which respondents were asked if they felt that U. S. nuclear weapons made them feel safer: n = 1005, x = 543 who said "yes". Use a confidence level of 80%. ____

7–5 **Internet Shopping** In a Gallup poll, 1025 randomly selected adults were surveyed and 29% of them said that they used the Internet for shopping at least a few times a year.

a. Find the point estimate of the percentage of adults who use the Internet for shopping. _______________

b. Find a 99% confidence interval estimate of the percentage of adults who use the Internet for shopping. ____

c. Based on the result from part (b), if a traditional retail store wants to estimate the percentage of adult Internet shoppers in order to determine the maximum impact of Internet shoppers on its sales, what percentage of Internet shoppers should be used? _______________

7-6 **Death Penalty Survey** In a Gallup poll, 491 randomly selected adults were asked whether they are in favor of the death penalty for a person convicted of murder, and 65% of them said that they were in favor.

a. Find the point estimate of the percentage of adults who are in favor of this death penalty. _______________

b. Find a 95% confidence interval estimate of the percentage of adults who are in favor of this death penalty. ____

c. Can we safely conclude that the majority of adults are in favor of this death penalty? Explain.

____

____

____

7-7 **Mendelian Genetics** When Mendel conducted his famous genetics experiments with peas, one sample of offspring consisted of 428 green peas and 152 yellow peas.

a. Use Statdisk to find the following confidence interval estimates of the percentage of yellow peas.

99.5% confidence interval: ____
99% confidence interval: ____
98% confidence interval: ____
95% confidence interval: ____
90% confidence interval: ____

b. After examining the pattern of the above confidence intervals, complete the following statement. "As the degree of confidence decreases, the confidence interval limits

____."

c. In your own words, explain why the preceding completed statement makes sense. That is, why should the confidence intervals behave as you have described?

____

____

7-8 **Gender Selection** The Genetics and IVF Institute conducted a clinical trial of the XSORT method designed to increase the probability of conceiving a girl. As of this writing, 945 babies were born to parents using the XSORT method, and 879 of them were girls.

a. What is the best point estimate of the population proportion of girls born to parents using the XSORT method? __________________________________

b. Use the sample data to construct a 95% confidence interval estimate of the percentage of girls born to parents using the XSORT method.

____

c. Based on the results, does the XSORT method appear to be effective? Why or why not?

____

____

____

7-9 **Gender Selection** The Genetics and IVF Institute conducted a clinical trial of the YSORT method designed to increase the probability of conceiving a boy. As of this writing, 291 babies were born to parents using the YSORT method, and 239 of them were boys.

a. What is the best point estimate of the population proportion of boys born to parents using the YSORT method? ______________________________

b. Use the sample data to construct a 99% confidence interval estimate of the percentage of boys born to parents using the YSORT method.

____

c. Based on the results, does the YSORT method appear to be effective? Why or why not?

____

____

7-10 **Touch Therapy** When she was nine years of age, Emily Rosa did a science fair experiment in which she tested professional touch therapists to see if they could sense her energy field. She flipped a coin to select either her right hand or her left hand, then she asked the therapists to identify the selected hand by placing their hand just under Emily's hand without seeing it and without touching it. Among 280 trials, the touch therapists were correct 123 times (based on data in "A Close Look at Therapeutic Touch," *Journal of the American Medical Association*, Vol. 279, No. 13).

a. Given that Emily used a coin toss to select either her right hand or her left hand, what proportion of correct responses would be expected if the touch therapists made random guesses? __________

b. Using Emily's sample results, construct a 99% confidence interval estimate of the proportion of correct responses made by touch therapists.

____

c. What do the results suggest about the ability of touch therapists to select the correct hand by sensing an energy field?

____

In Exercises 7-11 through 7-14, use the Statdisk data sets from Appendix B in the Triola textbook (data sets are available in Statdisk by clicking on ***Data Sets*** *in top menu bar).*

7-11 **Nicotine in Cigarettes** Refer to the 12th Edition data set *10 - Cigarette Tar, Nicotine, and Carbon Monoxide* and assume that the samples are simple random samples obtained from normally distributed populations..

a. Construct a 95% confidence interval estimate of the mean amount of nicotine in cigarettes that are king size, non-filtered, non-menthol, and non-light.

____

b. Construct a 95% confidence interval estimate of the mean amount of nicotine in cigarettes that are 100 mm, filtered, non-menthol, and non-light.

____

c. Compare the results. Do filters on cigarettes appear to be effective?

____

____

7-12 **Pulse Rates** A physician wants to develop criteria for determining whether a patient's pulse rate is atypical, and she wants to determine whether there are significant differences between males and females. Use the sample pulse rates from data set *1 - Body Measurements Male/Female* found in Appendix B of the textbook.

a. Construct a 95% confidence interval estimate of the mean pulse rate for males.

____

b. Construct a 95% confidence interval estimate of the mean pulse rate for females.

____

c. Compare the preceding results. Can we conclude that the population means for males and females are different? Why or why not?

____

____

7-13 **Weights of Coins** Refer to the weights of pre-1964 quarters and the weights of post-1964 quarters listed in data set *21 – Coin Weights,* found in Appendix B from the textbook.

a. Construct a 95% confidence interval estimate of the mean weight of pre-1964 quarters.

____

b. Construct a 95% confidence interval estimate of the mean weight of post-1964 quarters.

____

c. Compare the preceding results. Can we conclude that the population means for pre-1964 quarters and post-1964 quarters are different? Why or why not?

____

____

7-14 **Second Hand Smoke** Refer to the measured cotinine levels of smokers, nonsmokers exposed to cigarette smoke, and nonsmokers not exposed to cigarette smoke. The cotinine measurements are listed in data set *9 – Passive and Active Smoke*, found in Appendix B from the textbook.

a. Construct a 95% confidence interval estimate of the mean cotinine level of smokers.

____

b. Construct a 95% confidence interval estimate of the mean cotinine level of nonsmokers exposed to cigarette smoke. (The column label is ETS.)

____

c. Construct a 95% confidence interval estimate of the mean cotinine level of nonsmokers not exposed to cigarette smoke. (The column label is No ETS.)

____

d. Compare the preceding results. What do you conclude?

____

____

7-15 **Simulated Data** In this experiment we will generate 500 IQ scores, then we will construct a confidence interval based on the sample results. IQ scores have a normal distribution with a mean of 100 and a standard deviation of 15. First generate the 500 sample values as follows.

1. Click on **Data,** then select **Normal Generator**.
2. In the dialog box, enter a sample size of 500, a mean of 100, and a standard deviation of 15. Click **Generate**.
3. Copy the generated data into the Sample Editor.
4. Use **Data/Descriptive Statistics** to find these statistics:

n = ___________ $\bar{x}$ = ___________ s = ___________

Using the generated values, construct a 95% confidence interval estimate of the population mean of all IQ scores. Enter the 95% confidence interval here.

____

Because of the way that the sample data were generated, we *know* that the population mean is 100. Do the confidence interval limits contain the true mean IQ score of 100?

____

If this experiment were to be repeated many times, how often would we expect the confidence interval limits to contain the true population mean value of 100? Explain how you arrived at your answer.

____

____

7-16 **Simulated Data** Follow the same steps listed in Experiment 7-15 to randomly generate 500 IQ scores from a population having a normal distribution, a mean of 100, and a standard deviation of 15. Record the sample statistics here.

n = ___________ $\bar{x}$ = ___________ *s* = ___________

Confidence intervals are typically constructed with confidence levels around 90%, 95%, or 99%. Instead of constructing such a typical confidence interval, use the generated values to construct a 50% confidence interval. Enter the result below.

____

Does the above confidence interval have limits that actually do contain the true population mean, which we know is 100? ___________

Repeat the above procedure 9 more times and list the resulting 50% confidence intervals here.

___________________________ ___________________________

___________________________ ___________________________

___________________________ ___________________________

___________________________ ___________________________

____

Among the total of the 10 confidence intervals constructed, how many of them actually do contain the true population mean of 100? Is this result consistent with the fact that the level of confidence used is 50%? Explain.

____

____

7-17 **M&M Weights** Refer to data set *20 - M and M Plain Candy Weights,* found in Appendix B from the textbook. Use the entire sample of 100 plain M&M candies to construct a 95% confidence interval for the mean weight of all M&Ms. (*Hint*: It is not necessary to manually enter the 100 weights, because they are already stored in separate columns in the data set. Use **Copy/Paste** to combine the different M&M columns into one big column.) Now construct a 95% confidence interval estimate of the population mean of all weights. Enter the 95% confidence interval here.

____

7-18 **Combining Data Sets** Refer to the word counts in data set *17 – Word Counts by Males and Females,* found in Appendix B in the textbook. Use Copy/Paste to combine all of the word counts from males in one column, and combine all of the word counts from females in another column.

a. Find the 95% confidence interval estimate of the mean word count for men.

____

b. Find the 95% confidence interval estimate of the mean word count for women.

____

c. Based on the results, does it appear the women talk more than men? Why or why not?

____

____

In Exercises 7-19 through 7-22, use the Statdisk data sets from Appendix B in the textbook to construct confidence interval estimates of the population standard deviations.

7-19 **Nicotine in Cigarettes** Refer to data set *10 – Cigarette Tar, Nicotine, and Carbon Monoxide* and assume that the samples are simple random samples obtained from normally distributed populations.

a. Construct a 95% confidence interval estimate of the standard deviation of amounts of nicotine in cigarettes that are king size, non-filtered, non-menthol, and non-light (column name KgNic). ____

b. Construct a 95% confidence interval estimate of the standard deviation of amounts of nicotine in cigarettes that are 100 mm, filtered, non-menthol, and non-light (column name FLNic). ____

c. Compare the results. Do the amounts of variation appear to be different?

____

____

Statdisk

7-20 **Pulse Rates** A physician wants to develop criteria for determining whether a patient's pulse rate is atypical, and she wants to determine whether there are significant differences between males and females. Use the sample pulse rates in data set *1 - Body Measurements Female/Male*.

a. Construct a 95% confidence interval estimate of the standard deviation of pulse rates (PULSE) for males.

____

b. Construct a 95% confidence interval estimate of the standard deviation of pulse rates (PULSE) for females.

____

c. Compare the preceding results. Do the amounts of variation appear to be different?

____

____

____

____

7-21 **Weights of Coins** Refer to the weights of pre-1964 quarters and the weights of post-1964 quarters listed in data set *21 – Coin Weights*.

a. Construct a 95% confidence interval estimate of the standard deviation of the weights of pre-1964 quarters.

____

b. Construct a 95% confidence interval estimate of the standard deviation of the weights of post-1964 quarters.

____

c. Compare the preceding results. Do the amounts of variation appear to be different?

____

____

7-22 **Second Hand Smoke** Refer to the measured cotinine levels of smokers, nonsmokers exposed to cigarette smoke, and nonsmokers not exposed to cigarette smoke. The cotinine measurements are listed in data set *9 – Passive and Active Smoke*.

a. Construct a 95% confidence interval estimate of the standard deviation of cotinine levels of smokers.

____

b. Construct a 95% confidence interval estimate of the standard deviation of cotinine levels of nonsmokers exposed to cigarette smoke. (The column label is ETS.)

____

c. Construct a 95% confidence interval estimate of the standard deviation of cotinine levels of nonsmokers not exposed to cigarette smoke. (The column label is NOETS.)

____

d. Compare the preceding results. What do you conclude?

____

____

____

7–23. **Sample Size for Proportion** Many states are carefully considering steps that would help them collect sales taxes on items purchased through the Internet. How many randomly selected sales transactions must be surveyed to determine the percentage that transpired over the Internet? Assume that we want to be 99% confident that the sample percentage is within two percentage points of the true population percentage for all sales transactions. _______________

7-24 **Sample Size for Proportion** As a manufacturer of golf equipment, the Spalding Corporation wants to estimate the proportion of golfers who are left-handed. (The company can use this information in planning for the number of right-handed and left-handed sets of golf clubs to make.) How many golfers must be surveyed if we want 99% confidence that the sample proportion has a margin of error of 0.025?

a. Assume that there is no available information that could be used as an estimate of $\hat{p}$.__________

b. Assume that we have an estimate of $\hat{p}$ found from a previous study that suggests that 15% of golfers are left-handed (based on a *USA Today* report).____________

c. Assume that instead of using randomly selected golfers, the sample data are obtained by asking TV viewers of the golfing channel to call an "800" phone number to report whether they are left–handed or right–handed. How are the results affected?

____

7-25 **Sample Size for Proportion** You have been hired by the Ford Motor Company to do market research, and you must estimate the percentage of households in which a vehicle is owned. How many households must you survey if you want to be 94% confident that your sample percentage has a margin of error of three percentage points?

a. Assume that a previous study suggested that vehicles are owned in 86% of households._______________
b. Assume that there is no available information that can be used to estimate the percentage of households in which a vehicle is owned._______________
c. Assume that instead of using randomly selected households, the sample data are obtained by asking readers of the *Washington Post* newspaper to mail in a survey form. How are the results affected?

____

7-26 **Sample Size for Mean** An economist wants to estimate the mean income for the first year of work for college graduates who have had the profound wisdom to take a statistics course. How many such incomes must be found if we want to be 95% confident that the sample mean is within \$500 of the true population mean? Assume that a previous study has revealed that for such incomes, σ = \$6250._______________

7-27 **Sample Size for Mean** Nielsen Media Research wants to estimate the mean amount of time (in minutes) that full-time college students spend watching television each weekday. Find the sample size necessary to estimate that mean with a 15 minute

Statdisk

margin of error. Assume that a 96% confidence level is desired. Also assume that a pilot study showed that the standard deviation is estimated to be 112.2 min.________________

7-28 **Sample Size for Variation** In each of the following, assume that each sample is a simple random sample obtained from a normally distributed population.

a. Find the minimum sample size needed to be 95% confident that the sample standard deviation *s* is within 10% of σ.________________

b. Find the minimum sample size needed to be 95% confident that the sample standard deviation *s* is within 30% of σ.________________

c. Find the minimum sample size needed to be 99% confident that the sample variance is within 1% of the population variance. Is such a sample size practical in most cases?________________

d. Find the minimum sample size needed to be 95% confident that the sample variance is within 20% of the population variance.________________

7-29 **Bootstrap Resampling** The sample values 2.9, 564.2, 1.4, 4.7, 67.6, 4.8, 51.3, 3.6, 18.0, and 3.6 are randomly selected from a population with a distribution that is far from normal. Use bootstrap resampling to construct a 95% confidence interval estimate of μ, and use bootstrap resampling to construct a 95% confidence interval estimate of σ. Enter the results here.

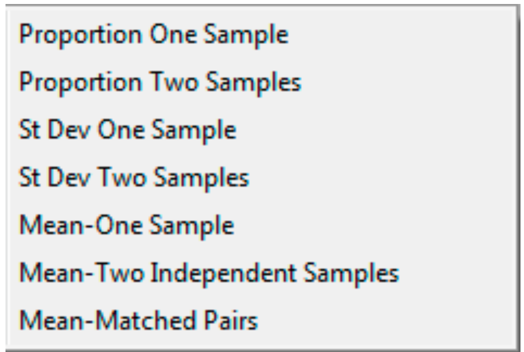
________________________________ ________________________________

Statdisk

8 Hypothesis Testing

Statdisk

Statdisk is designed for conducting a variety of hypothesis tests included in the textbook. If you click on the top menu item of **Analysis**, then select the subdirectory item of **Hypothesis Testing**, you will see the following menu of choices.



Among the items in the above list, three include a reference to *one sample*, and they involve hypothesis tests with claims made about a single population, as discussed in Chapter 8 of the Triola textbook. The other items involve *two* sets of sample data as described in Chapter 9 of the textbook. This chapter will consider hypothesis testing involving only one sample. The claim may be made about the proportion of a single population, or the mean of a single population, or the standard deviation or variance of a single population. In Chapter 9 we consider claims involving two samples.

We use hypothesis testing when we want to test some claim made about a particular characteristic of some population. In addition to the claim itself, we also need sample data and a significance level.

8-1 Testing Hypotheses About a Proportion *p*

Section 8-2 of the textbook introduces general concepts and terminology associated with hypothesis testing, and Section 8-3 introduces procedures for testing claims about the proportion of a single population. The Statdisk procedure for testing claims about a population proportion is quite easy. In addition to having a claim to be tested, Statdisk also requires a significance level, the sample size n, and the number of successes x.

Finding x: In Section 7-1 of this workbook, we briefly discussed one particular difficulty that arises when the available information provides the sample size n and the sample proportion $\hat{p}$ instead of the number of successes x. We provided this procedure for determining the number of successes x.

> **To find the number of successes x from the sample proportion and sample size:**
> **Calculate $x = \hat{p}\, n$, then round the result to the nearest whole number.**

Given a claim about a proportion and knowing *n* and *x*, we can use Statdisk as follows.

Statdisk Procedure for Testing Claims about *p*

1. Select **Analysis** from the top menu bar.
2. Select **Hypothesis Testing** from the subdirectory.
3. Select **Proportion One Sample**.
4. Make these entries in the dialog box.
   - Select the format of the claim that is being tested. The default will appear as

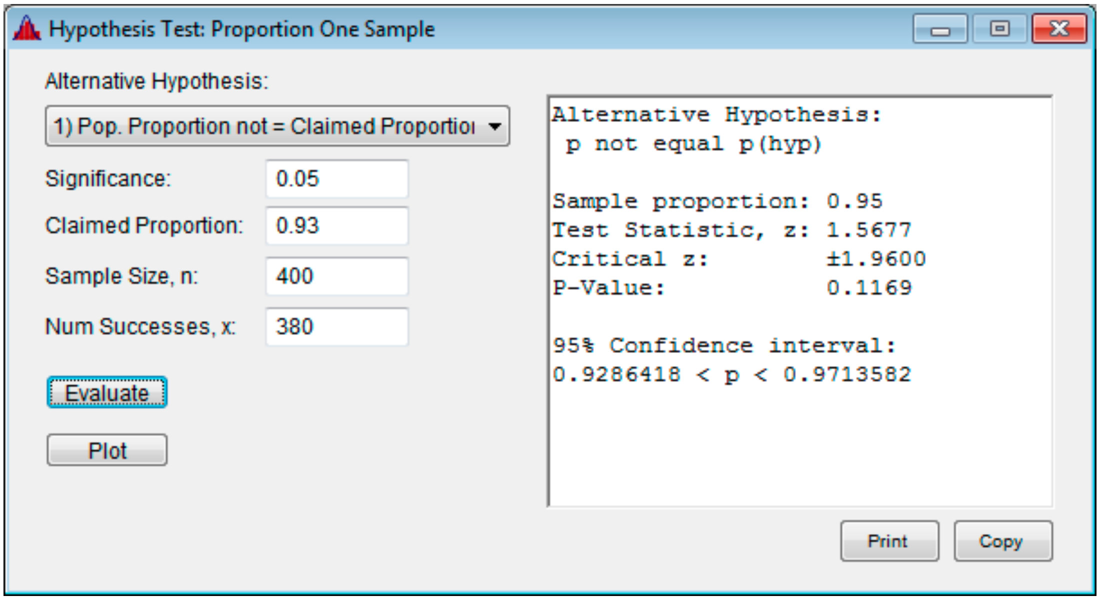
     1) Pop. Proportion = Claimed Proportion

     and this can be changed to any of the of 3 options. Click on the box to make the other options appear, then click on the format of the claim being tested.
   - Enter a significance level, such as 0.05 or 0.01.
   - Enter the *claimed* value of the population proportion.
   - Enter the sample size, *n*.
   - Enter the number of successes, *x*.
5. Click **Evaluate**.

Consider Example 1 in Section 8-3 of the textbook:

> **Example** 93% of computer owners believe that they have antivirus programs installed on their computers. In a random sample of 400 scanned computers, it is found that 380 of them (or 95%) actually have antivirus programs. Use the sample data from the scanned computers with a 0.05 significance level to test the claim that 93% of computers have antivirus programs.

Based on the information provided, the Statdisk dialog and results are as follows.



Important elements of the Statdisk display include the *P*-value of 0.1169, the test statistic of z = 1.5677, and critical values of ±1.9600. Having the *P*-value and critical values available, we can use either the critical value method of testing hypotheses or the *P*-value method. For this example, we know that the *P*-value of .1169 is greater than the significance level of 0.05, so we fail to reject the null hypothesis. If we were to use the critical value approach, we see that the test statistic of z = 1.5677 does not fall within the critical region, so we fail to reject the null hypothesis. Conclusion: There is not sufficient sample evidence to warrant rejection of the claim that 93% of computers have antivirus programs.

Note also that the Statdisk display includes this 95% confidence interval:

$$0.929 < p < 0.971 \text{ (rounded)}$$

The following points are important for interpreting this confidence interval.

1. Because the hypothesis test is two–tailed, the 0.05 significance level for the hypothesis test corresponds to a 95% confidence level for a confidence interval.

2. The textbook notes that both the critical value method and *P*–value method use the same standard deviation based on the *claimed proportion p*, but the confidence interval uses an estimated standard deviation based on the *sample proportion* $\hat{p}$. Consequently, it is possible that in some cases, the critical value and *P*–value methods of testing a claim about a proportion might yield a different conclusion than the confidence interval method.

8–2 Testing Hypotheses About a Mean μ

Section 8-4 in the textbook describes details for methods of testing claims about a population mean μ. Part 1 of Section 8-4 focuses on the realistic case in which the population standard deviation σ is not known, and Part 2 briefly considers the somewhat artificial case in which σ is somehow known.

The textbook explains that there are different procedures, depending on the size of the sample, the nature of the population distribution, and whether the population standard deviation σ is known. The textbook stresses the importance of selecting the correct distribution (normal or *t*). The criteria are summarized in the table below.

Choosing between *z* and *t*

Method	Conditions
Use normal (*z*) distribution.	σ known and normally distributed population *or* σ known and $n > 30$
Use *t* distribution.	σ not known and normally distributed population *or* σ not known and $n > 30$
Use a nonparametric method or bootstrapping.	Population is not normally distributed and $n \leq 30$

Statdisk greatly simplifies the process of choosing between the normal and *t* distributions because it is programmed to make the correct choice, depending on the information that is supplied. [One exception: Like other statistics software packages, Statdisk's hypothesis testing modules are not programmed to check for normality of the population, so you should not use *t*-test results if the sample size is small ($n \leq 30$) and the population has a distribution that is very non-normal.]

Statdisk Procedure for Hypothesis Tests about a Mean

1. Select the top menu item of **Analysis**.
2. Select **Hypothesis Testing** from the subdirectory.
3. Select **Mean - One Sample.**
4. You will now see a dialog box for entry of the claim, significance level, value of the claimed mean, and entry of the sample statistics.

Statdisk

- Select the format of the claim that is being tested. The default will appear as

  1) Pop. Mean = Claimed Mean

  This can be changed to any of the 3 possibilities. Click on the box to make the other options appear, then click on the format of the claim being tested.

- Enter a significance level, such as 0.05 or 0.01.

- Enter the *claimed* value of the population mean.

- Enter the *population* standard deviation σ if it is known. If σ is not known (as is usually the case), ignore that box and leave it empty. (*Caution*: Be careful to avoid the mistake of incorrectly entering the *sample* standard deviation in the box for the *population* standard deviation.)

- Enter the sample size *n*, sample mean $\bar{x}$, and the sample standard deviation *s*.

5. Click the **Evaluate** button to get the test results.

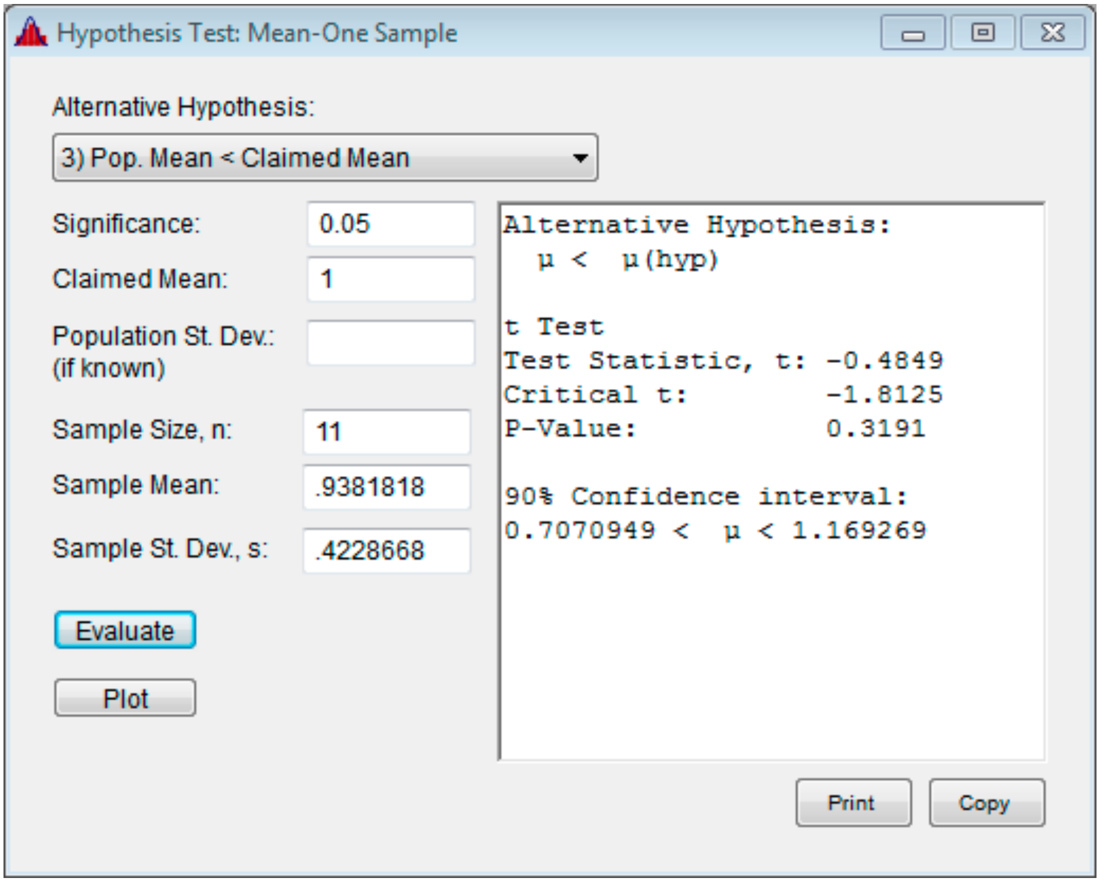
6. Click the **Plot** button to get a graph that shows the test statistic and critical values.

As an illustration, consider Example 1 in Section 8-4 of the textbook:

> **EXAMPLE *Cell Phone Radiation*** Listed below are the measured radiation emissions (in W/kg) corresponding to a sample of cell phones (based on data are from the Environmental Working Group). Use a 0.05 significance level to test the claim that cell phones have a mean radiation level that is less than 1.00 W/kg.
>
> 0.38 0.55 1.54 1.55 0.50 0.60 0.92 0.96 1.00 0.86 1.46

Because $n \leq 30$, we must verify that the sample data appear to be from a normally distributed population, and Example 1 in Section 8-4 of the textbook shows the normal quantile plot suggesting that this requirement is met.

Follow the procedure above and enter the sample data as shown in the Statdisk display. (Mean and standard deviation can be obtained by entering the data values in the example and using Statdisk's **Explore Data** function as described in Section 3-1 in this workbook.)



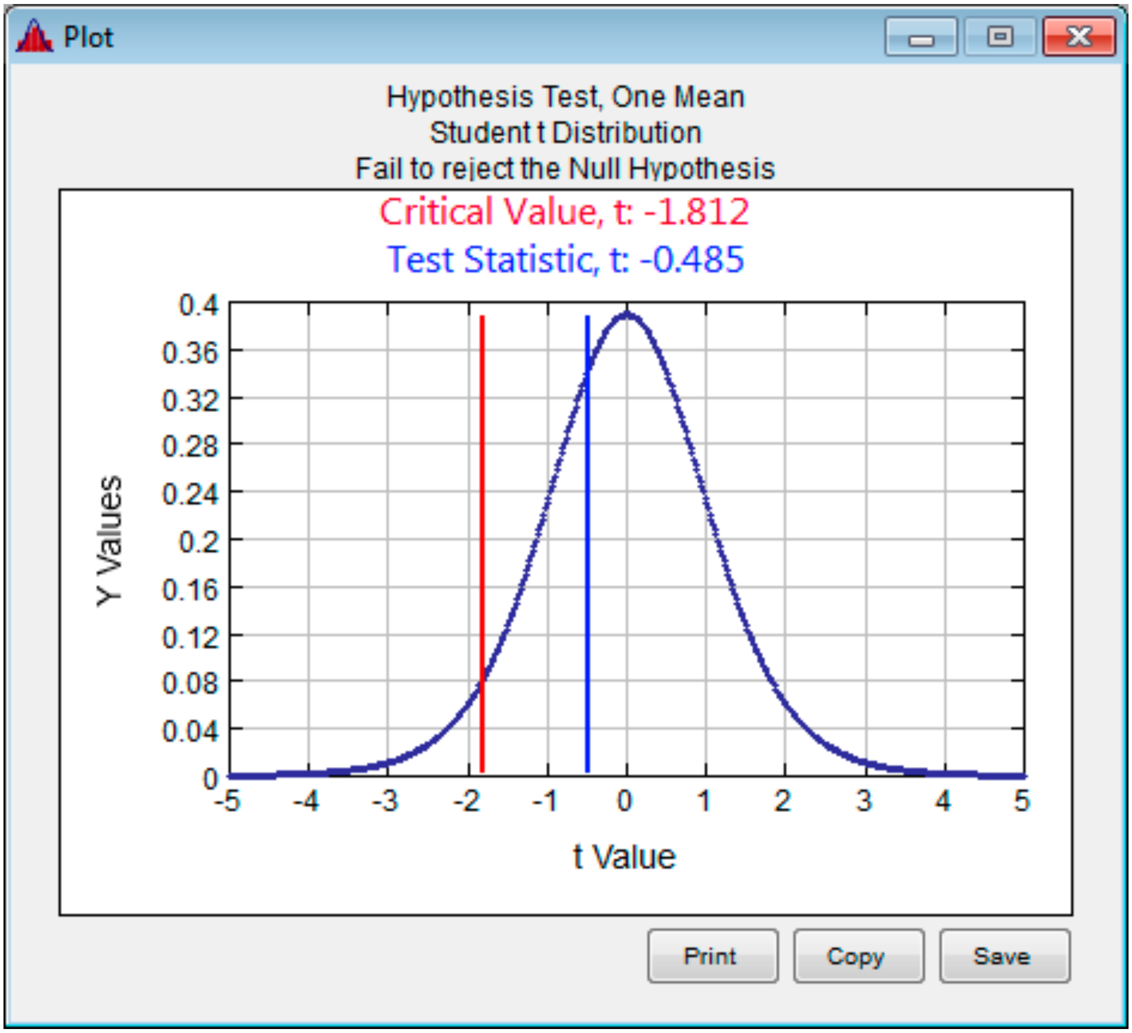
From the above Statdisk display, we see that the test statistic is t = -0.4849. The P-value is 0.3191. Because the P-value is greater than the significance level of 0.05, we fail to reject the null hypothesis and conclude that there is not sufficient evidence to support the claim that the population mean is less than 1.00 W/kg. (See -xample 1 in Section 8-4 in the textbook.)

Note that the display also includes the 90% confidence interval of $0.7071 < \mu < 1.1693$ (rounded). The following points are important for interpreting this confidence interval.

1. Because the hypothesis test is left–tailed, the 0.05 significance level for the hypothesis test corresponds to a 90% confidence level for a confidence interval.

2. The confidence interval of $0.707 < \mu < 1.169$ (rounded) suggests that the population mean μ can be any value between 0.707 and 1.169, and the assumed mean of 1.0 does fall within those limits, so there is not sufficient evidence to support the claim that $\mu <$ 1.0.

If you click on the **Plot** button, you will also obtain the graph shown below. The vertical line representing the test statistic will be in blue, and vertical lines representing critical values will be in red.



8-3 Testing Hypotheses About σ or σ^2

When testing claims about σ or σ^2, the textbook stresses that there are the following two important requirements:

1. The samples are simple random samples. (Remember the importance of good sampling methods.)

2. The sample values come from a population with a *normal distribution*.

The textbook makes the very important point that *for tests of claims about standard deviations or variances, the requirement of a normal distribution is very strict*. If the population does not have a normal distribution, then inferences about standard deviations or variances can be very misleading. *Suggestion:* Given sample data, use Statdisk's *Normality Assessment* tool to generate a display that includes a histogram, normal quantile plot, and information about potential outliers to determine whether the assumption of a normal distribution is reasonable. If the population distribution does not appear to have a normal distribution, do not use the methods described in Section 8-6 of the textbook or this section of this workbook. If the population distribution *does* appear to be normal and you want to test a claim about the population standard deviation or variance, use the Statdisk procedure given below.

Although Statdisk is designed to work only with standard deviations, claims about a population variance can be handled as well. For example, to test the claim that $\sigma^2 = 9$, take the square root of both sides of the equation, then restate the claim as $\sigma = 3$. Also, Statdisk requires entry of the sample *standard deviation s*, so if the sample variance is known, be sure to enter the value of *s*, which is the square root of the value of the sample variance.

Statdisk Procedure for Testing Hypotheses about σ or σ^2

1. Select **Analysis** from the top menu bar.

2. Select **Hypothesis Testing** from the subdirectory.

3. Select the option of **St. Dev. One Sample**. (Select this option for claims about standard deviations or variances.)

4. Make the following entries in the dialog box.

   - In the **Alternative Hypothesis** box, select the format of the claim being tested.
   - Enter a significance level, such as 0.05 or 0.01.
   - Enter the *claimed* value of the standard deviation. (This is the value used in the statement of the null hypothesis.)

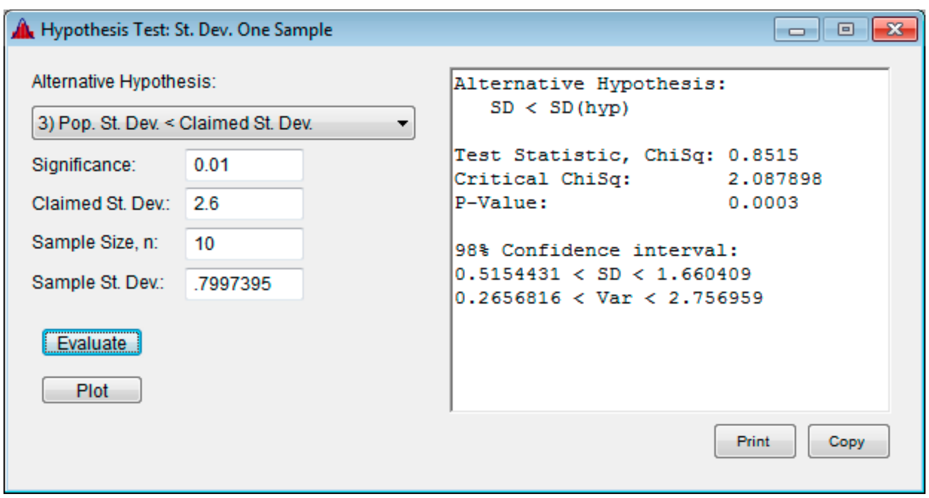
- Enter the sample size *n*.
- Enter the value of the sample standard deviation *s*.

5. Click **Evaluate**.

As an illustration, consider this from Example 1 in Section 8-5 of the textbook:

> **EXAMPLE** ***Supermodel Heights*** Listed below are the heights (inches) for the simple random sample of supermodels. Consider the claim that supermodels have heights that have much less variation than heights of women in the general population. We will use a 0.01 significance level to test the claim that supermodels have heights with a standard deviation that is less than 2.6 in. for the population of women.
>
> 70 71 69.25 68.5 69 70 71 70 70 69.5

From the above example, we see that we want to test the claim that σ < 2.6 in., and we want to use a 0.01 significance level. See Example 1 in Section 8-5 of the textbook, where the normality of the distribution is verified with a normal quantile plot. We can proceed with the hypothesis test, and the Statdisk dialog box with test results is shown below.



The Statdisk results include the test statistic of χ^2 = 0.8515 and a *P*-value of 0.0003. Because the *P*-value is low, we reject the null hypothesis and conclude that there is sufficient evidence to support the claim that the population standard deviation σ is less than 2.6 in.

Because critical values are included in the Statdisk display, the critical value approach to hypothesis testing can also be used. A confidence interval for σ (denoted by SD in the display) is also displayed, along with a confidence interval for σ^2 (denoted by Var, for variance).

Statdisk

8-4 Hypothesis Testing with Simulations

Sections 8-1, 8-2, and 8-3 of this manual/workbook have all described the use of Statdisk for hypothesis tests using the critical value method, *P*-value method, or confidence intervals. Another very different approach is to use *simulations*. Let's illustrate the simulation technique with an example.

Consider testing the claim that the population of body temperatures of healthy adults has a mean less than 98.6°F. That is, consider the claim that μ < 98.6°F. We will use sample data consisting of these 12 values:

98.0 97.5 98.6 98.8 98.0 98.5 98.6 99.4 98.4 98.7 98.6 97.6

For these 12 values, n = 12, $\bar{x}$ = 98.39167, and s = 0.53506. The key question is this:

> **If the population mean body temperature is really 98.6, then *how likely* is it that we would get a sample mean of 98.39167, given that the population has a normal distribution and the sample size is 12?**

If the probability of getting a sample mean such as 98.39167 is very small, then that suggests that the sample values are not the result of chance random fluctuation. If the probability is high, then we can accept random chance as an explanation for the discrepancy between the sample mean of 98.39167 and the assumed mean of 98.6. What we need is some way of determining the likelihood of getting a sample mean such as 98.39167. That is the precise role of *P*-values in the *P*-value approach to hypothesis testing. However, there is another approach. Statdisk and many other software packages are capable of generating random results from a variety of different populations.

Here is how Statdisk can be used: Determine the likelihood of getting a sample mean of 98.39167 by randomly generating several different samples from a population that is normally distributed with the claimed mean of 98.6. For the standard deviation, we will use the best available information: the value of s = 0.53506 obtained from the sample.

Statdisk Procedure for Testing Hypotheses with Simulations

1. Identify the values of the sample size *n*, the sample standard deviation *s*, and the claimed value of the population mean.

2. Click on **Data**.

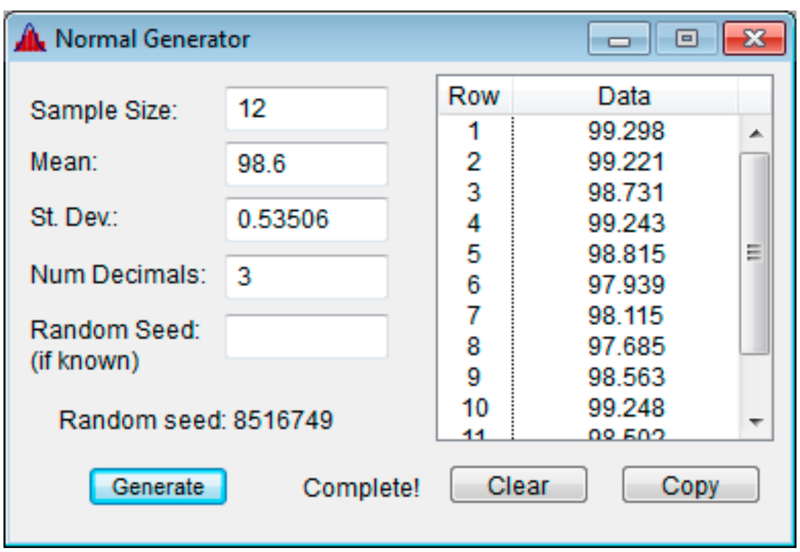
3. Click on **Normal Generator**.

4. Make these entries in the dialog box.
   - Enter sample size.
   - Enter the *claimed* mean (not the sample mean)
   - Enter the sample standard deviation *s*.
   - Specify the desired number of decimals. (Use at least as many decimal places as in the sample data values.)

5. Click **Generate** to generate the sample.

6. Calculate and record the mean of the sample generated in Step 5.

7. Continue to generate similar samples until it becomes clear that the given sample mean is or is not likely. (Here is one criterion: The given sample mean is *unlikely* if values at least as extreme occur 5% of the time or less.) If it is unlikely, reject the claimed mean. If it is likely, fail to reject the claimed mean.



For example, here are 20 results obtained from the random generation of samples of size 12 from a normally distributed population with mean 98.6 and standard deviation 0.53506:

Sample Means

98.754	**98.374**	**98.332**	98.638	98.513
98.551	98.566	98.760	**98.332**	98.603
98.407	98.640	98.655	98.408	98.802
98.505	98.699	98.609	98.206	98.582

Examining the 20 sample means, we see that three of them (displayed in bold) are 98.39167 or lower. Because 3 of the 20 results (or 15%) are at least as extreme as the sample mean of 98.39167, we see that a sample mean such as 98.39167 is *common* for these circumstances. This suggests that a sample mean of 98.39167 is not *significantly* different from the assumed mean of 98.6. We would feel more confident in this conclusion if we had more sample results, so we could continue to randomly generate simulated samples until we feel quite confident in our thinking that a sample mean such as 98.39167 is not an unusual result. It can easily occur as the result of chance random variation. We therefore fail to reject the null hypothesis that the mean equals 98.6. There is not sufficient evidence to support the claim that the mean is less than 98.6.

CHAPTER 8 EXPERIMENTS: Hypothesis Testing

Experiments 8–1 through 8–4 involve claims about proportions.

8-1 **Reporting Income** In a Pew Research Center poll of 745 randomly selected adults, 589 said that it is morally wrong to not report all income on tax returns. Use a 0.01 significance level to test the claim that 75% of adults say that it is morally wrong to not report all income on tax returns.

Test statistic: ___________ Critical value(s): ___________ *P*–value: _____________

Conclusion in your own words: ____

____

8-2 **Voting for the Winner** In a presidential election, 308 out of 611 voters surveyed said that they voted for the candidate who won (based on data from ICR Survey Research Group). Use a 0.01 significance level to test the claim that among all voters, the percentage believing that they voted for the winning candidate is equal to 43%, which is the actual percentage of votes for the winning candidate.

Test statistic: ___________ Critical value(s): ___________ *P*–value: _____________

Conclusion in your own words: ____

____

8-3 **Tennis Instant Replay** The Hawk-Eye electronic system is used in tennis for displaying an instant replay that shows whether a ball is in bounds or out of bounds, so players can challenge calls made by referees. In the most recent U.S. Open (as of this writing), singles players made 611 challenges and 172 of them were successful with the call overturned. Use a 0.01 significance level to test the claim that fewer than 1/3 of the challenges are successful

Test statistic: ___________ Critical value(s): ___________ *P*–value: _____________

Conclusion in your own words: ____

____

8-4 **Screening for Marijuana Usage** The company Drug Test Success provides a "1-Panel-THC" test for marijuana usage. Among 300 tested subjects, results from 27 subjects were wrong (either a false positive or a false negative). Use a 0.05 significance level to test the claim that less than 10% of the test results are wrong.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

Experiments 8–5 through 8–11 require tests of claims about means of populations. ***All Statdisk data sets are from Appendix B in the Triola textbook.***

8-5 **Earthquake Magnitudes** Use the earthquake magnitudes listed in the Statdisk data set *16 – Earthquake Measurements* and test the claim that the population of earthquakes has a mean magnitude greater than 1.00. Use a 0.05 significance level.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

8-6 **Blood Pressure** Use the systolic blood pressure measurements for females listed in data set *1a – Body Measurements Female* and test the claim that the female population has a mean sytolic blood pressure level less than 120.0 mm Hg. Use a 0.05 significance level.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

8-7 **College Weights** Use the September weights of males in data set *4 – Freshman 15 Data* and test the claim that male college students have a mean weight that is less than the 83 kg mean weight of males in the general population. Use a 0.01 significance level.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

8-8 **Olympic Winners** Listed below are the winning times (in seconds) of men in the 100-meter dash for consecutive summer Olympic games, listed in order by row. Assuming that these results are sample data randomly selected from the population of all past and future Olympic games, test the claim that the mean time is less than 11 sec. What do you observe about the precision of the numbers? What extremely important characteristic of the data set is not considered in this hypothesis test? Do the results from the hypothesis test suggest that future winning times should be around 10.5 sec, and is such a conclusion valid?

12.0	11.0	11.0	11.2	10.8	10.8	10.8	10.6	10.8	10.3	10.3	10.3
10.4	10.5	10.2	10.0	9.95	10.14	10.06	10.25	9.99	9.92	9.96	

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

8-9 **Is the Diet Practical?** When 40 people used the Weight Watchers diet for one year, their mean weight loss was 3.0 lb and the standard deviation was 4.9 lb (based on data from "Comparison of the Atkins, Ornish, Weight Watchers, and Zone Diets for Weight Loss and Heart Disease Reduction," by Dansinger, et al., Journal of the American Medical Association, Vol. 293, No. 1). Use a 0.01 significance level to test the claim that the mean weight loss is greater than 0 lb.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

8-10 **Weights of Pennies** The U. S. Mint has a specification that pennies have a mean weight of 2.5 g. data set *21 – Coin Weights* lists the weights (in grams) of 37 pennies manufactured after 1983. Those pennies have a mean weight of 2.49910 g and a standard deviation of 0.01648 g. Use a 0.05 significance level to test the claim that this sample is from a population with a mean weight equal to 2.5 g.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

8-11 **Analysis of Pennies** In an analysis investigating the usefulness of pennies, the cents portions of 100 randomly selected credit card charges are recorded. The sample has a mean of 47.6 cents and a standard deviation of 33.5 cents. If the amounts from 0 cents to 99 cents are all equally likely, the mean is expected to be 49.5 cents. Use a 0.01 significance level to test the claim that the sample is from a population with a mean equal to 49.5 cents.

Test statistic: ___________ Critical value(s): ___________ *P*–value: _____________

Conclusion in your own words: ____

____

Experiments 8–12 through 8–15 involve claims about a standard deviation or variance.

8-12 **Ages of Race Car Drivers** Listed below are the ages (years) of randomly selected race car drivers (based on data reported in USA Today). Most people in the general population have ages that vary between 0 and 90 years, so use of the range rule of thumb suggests that ages in the general population have a standard deviation of 22.5 years. Use a 0.01 significance level to test the claim that the standard deviation of ages of all race car drivers is less than 22.5 years.

32 32 33 33 41 29 38 32 33 23 27 45 52 29 25

Test statistic: ___________ Critical value(s): ___________ *P*–value: _____________

Conclusion in your own words: ____

____

8-13 **Highway Speeds** Listed below are speeds (mi/h) measured from southbound traffic on I-280 near Cupertino, California (based on data from SigAlert). This simple random sample was obtained at 3:30 PM on a weekday. Use a 0.05 significance level to test the claim of the highway engineer that the standard deviation of speeds is equal to 5.0 mi/h.

62 61 61 57 61 54 59 58 59 69 60 67

Test statistic: ___________ Critical value(s): ___________ *P*–value: _____________

Conclusion in your own words: ____

____

8-14 **Aircraft Altimeters** The Skytek Avionics company uses a new production method to manufacture aircraft altimeters. A simple random sample of new altimeters resulted in errors listed below. Use a 0.05 level of significance to test the claim that the new production method has errors with a standard deviation greater than 32.2 ft, which was the standard deviation for the old production method. If it appears that the standard deviation is greater, does the new production method appear to be better or worse than the old method? Should the company take any action?

−42 78 −22 −72 −45 15 17 51 −5 −53 −9 −109

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

8-15 **IQ of Professional Pilots** The Wechsler IQ test is designed so that the mean is 100 and the standard deviation is 15 for the population of normal adults. Listed below are IQ scores of randomly selected professional pilots. It is claimed that because professional pilots are a more homogeneous group than the general population, they have IQ scores with a standard deviation less than 15. Test that claim using a 0.05 significance level.

121 116 115 121 116 107 127 98 116 101 130 114

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

Experiments 8-16 through 8-18 involve the simulation approach to hypothesis testing.

8-16 **Hypothesis Testing with Simulations** Consider a test of the claim that μ = 75. The following sample results were obtained.

91 85 94 86 82

a. Generate and print a normal quantile plot. Based on the result, does this sample appear to be from a population with a normal distribution? Why or why not?

____

____

b. What does the value of the sample mean suggest about the claim that μ = 75?

____

c. Generate different samples of size *n* = 5 until getting a sense for the likelihood of getting a sample mean like the one obtained. (*Hint:* See Section 8-4 in this manual/workbook.) List the sample means below.

____

____

d. What do you conclude? Why?

____

____

8-17 **Hypothesis Testing with Simulations** Use a simulation approach for conducting the hypothesis test described in Experiment 8–8. Describe the procedure, results, and conclusions.

____

____

____

8–18 **Hypothesis Testing with Simulations** Use a simulation approach for conducting the hypothesis test described in Experiment 8–14. Describe the procedure, results, and conclusions.

____

____

____

9 Inferences from Two Samples

Statdisk

As the use of technology permeates introductory statistics courses, instructors are expanding the scope of topics covered. They often accomplish this with assignments from chapters not formally covered in class. Once students understand the basic concepts of confidence interval construction and hypothesis testing, it becomes relatively easy to use a technology such as Statdisk to apply that understanding to other circumstances, such as those involving two samples instead of only one.

This chapter deals with inferences based on two sets of sample data. Some of the most important applications of statistics require the methods of this chapter, such as determining whether the proportion of adverse reactions in a sample of people using a new drug is the same as the proportion of adverse reactions in a sample of people using a placebo.

9-1 Two Proportions

The textbook makes the point that the section discussing inferences involving two proportions is one of the most important sections in the book because the main objective is to provide methods for dealing with two sample proportions – a situation that is very common in real applications.

When working with two proportions, Statdisk requires that we identify the number of successes x_1 and the sample size n_1 for the first sample, and identify x_2 and n_2 for the second sample. Sample data often consist of sample proportions or percentages instead of the actual numbers of successes, so we must know how to determine the number of successes. From $\hat{p}_1 = x_1/n_1$, we know that $x_1 = n_1 \cdot \hat{p}_1$ so that x_1 can be found by multiplying the sample size for the first sample by the sample proportion expressed in decimal form. For example, if 35.2% of 1200 interviewed subjects answer "yes" to a question, we have $n = 1200$ and $x = 1200 \cdot 0.352 = 422.4$, which we round to 422.

Statdisk Procedure for Testing Hypotheses About Two Proportions

To conduct hypothesis tests about two population proportions use the following procedure.

1. For both samples, find the sample size *n* and the number of successes *x*.
2. Select **Analysis** from the top menu bar.
3. Select **Hypothesis Testing** from the subdirectory.
4. Select **Proportion Two Samples**.

5. Make the following entries in the dialog box:
   - In the "Alternative Hypothesis" box, select the format corresponding to the claim.
   - Enter a significance level, such as 0.05 or 0.01.
   - For each sample, enter the sample size *n* and the number of successes *x*.

6. Click on the **Evaluate** button to obtain the test results.

7. Click on the **Plot** button to obtain a graph that includes the test statistic and critical value(s).

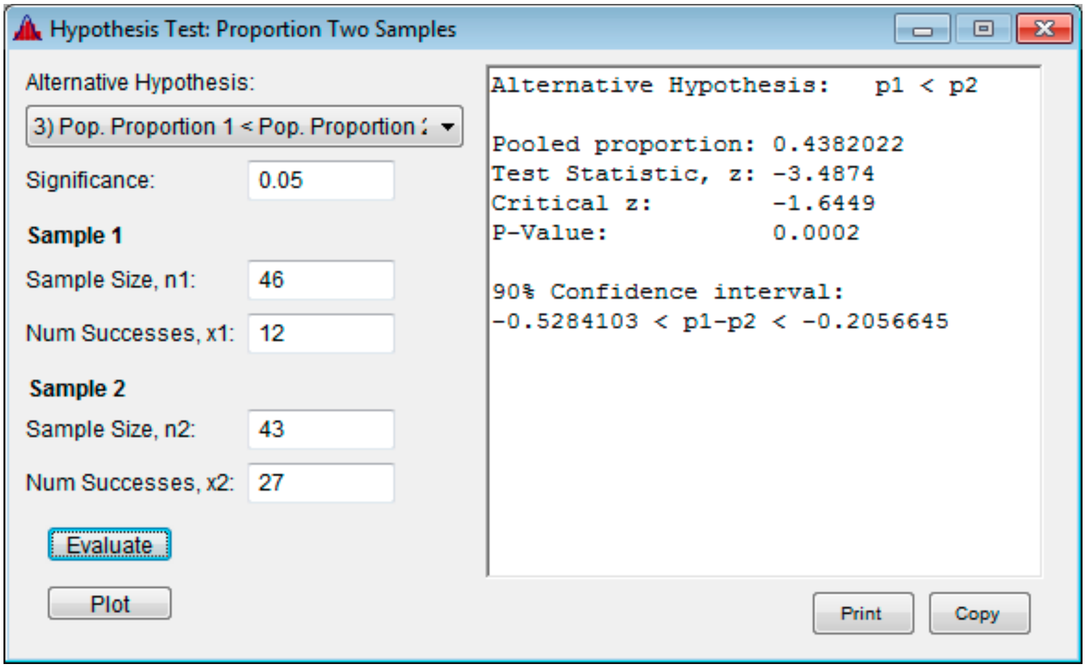
As an illustration, consider the following sample data from Example 1 in Section 9-2 of the textbook:

EXAMPLE ***Large Denominations Less Likely to be Spent?*** The table below lists sample results from a study conducted to determine whether people are less likely to spend money when it is in the form of larger denominations. Use a 0.05 significance level to test the claim that "money in a large denomination is less likely to be spent relative to an equivalent amount in many smaller denominations." That is, test the claim that the proportion of people who spend a $1 bill is less than the proportion who spend 4 quarters.

	Group 1	Group 2
	Subjects Given $1 Bill	**Subjects Given 4 Quarters**
Spent the money	$x_1 = 12$	$x_2 = 27$
Subjects in group	$n_1 = 46$	$n_2 = 43$

We can now proceed to use the above Statdisk procedure. After selecting **Analysis**, **Hypothesis Testing**, and **Proportion Two Samples**, we make the required entries in the dialog box as shown below. Note that the claim of a lower spending rate for the population given the $1 bill is the claim that $p_1 < p_2$, so we select the option of "less than" for the alternative hypothesis, as shown in the Statdisk display that follows.

The Statdisk results include the pooled proportion, test statistic, critical value, and *P*–value. Because the *P*-value of 0.0002 is less than the significance level of 0.05, we reject the null hypothesis and support the claim that those given $1 spend at a rate less than those given 4 quarters.



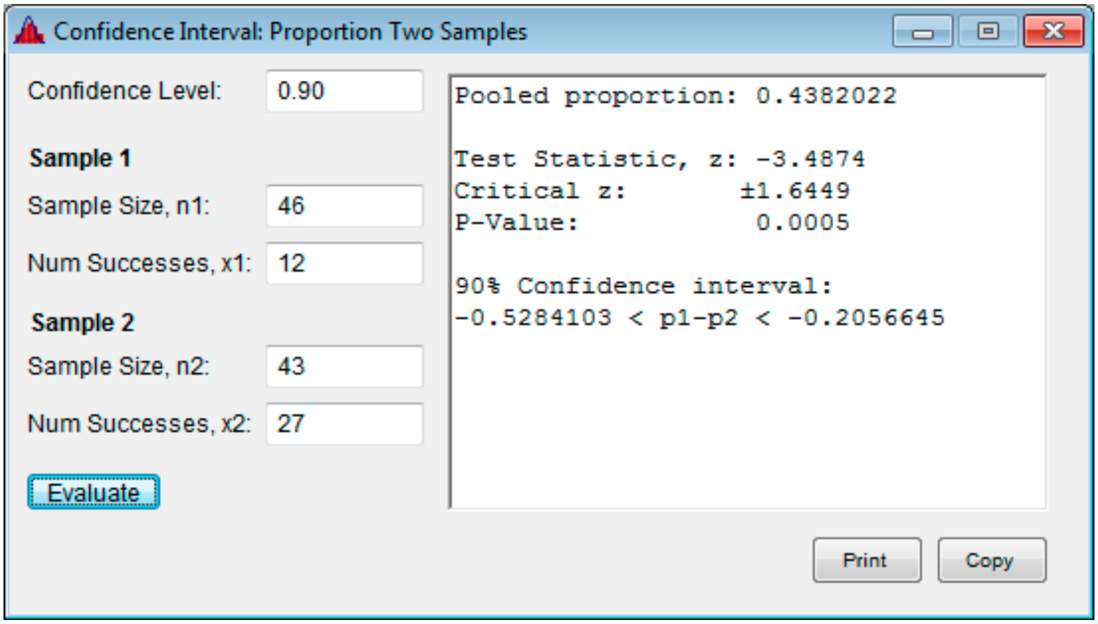
Note that the above results also provide the 90% confidence interval estimate of the difference between population proportions ($p_1 - p_2$). This confidence interval estimate does not include 0, so we have evidence suggesting that p_1 and p_2 have different values.

Statdisk can also determine the confidence interval estimate of the difference between population proportions using the following procedure. This procedure enables you to specify the confidence level.

Statdisk Procedure for a Confidence Interval Estimate of the Difference Between Two Proportions

1. For each of the two samples, find the sample size *n* and the number of successes *x*.
2. Select **Analysis** from the top menu bar.
3. Select **Confidence Intervals** from the subdirectory.
4. Select **Proportion Two Samples**.
5. Enter the confidence level.
6. Enter the sample size and number of successes for each of the two samples.
7. Click on the **Evaluate** button.

Shown below is a Statdisk display corresponding to the denomination/spending sample data in the preceding table. The confidence level of 0.90 was chosen so that the confidence interval corresponds to the one-sided hypothesis test conducted with a 0.05 significance level. Note that the confidence interval limits do not include zero, suggesting that there is a significant difference between the two proportions.



9-2 Two Means: Independent Samples

The textbook notes that two samples are **independent** if the sample values selected from one population are not related to or somehow paired or matched with the sample values selected from the other population. If there is some relationship so that each value in one sample is paired with a corresponding value in the other sample, the samples are **dependent.** Dependent samples are often referred to as **matched pairs,** or **paired samples.**

When testing a claim about the means of two independent samples, or when constructing a confidence interval estimate of the difference between the means of two independent samples, the textbook describes procedures based on the requirement that the two population standard deviations σ_1 and σ_2 are *not known*, and there is no assumption that $\sigma_1 = \sigma_2$. We focus on the first of the following three cases, and the other two cases are discussed briefly:

1. σ_1 and σ_2 are not known and are not assumed to be equal.
2. σ_1 and σ_2 are known.
3. It is assumed that $\sigma_1 = \sigma_2$.

Although the first of these three cases is the main focus of Section 9–3 in the textbook, Statdisk allows you to work with all three cases.

Statdisk Procedure for Tests of Hypotheses about Two Means: Independent Samples

1. For each of the two samples, identify the sample size, sample mean, and sample standard deviation. That is, find the values of n_1, $\bar{x}_1$, s_1, n_2, $\bar{x}_2$, and s_2. (If necessary, use Statdisk's *Descriptive Statistics* function to find the required sample statistics.)

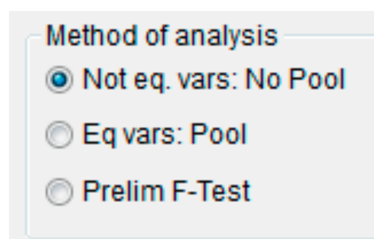
2. Select **Analysis** from the top menu bar.

3. Select **Hypothesis Testing** from the submenu.

4. Choose the option of **Mean - Two Independent Samples**.

5. Enter the required values in the dialog box. *Caution:* Be particularly careful with these items:
   - Alternative Hypothesis: Be sure to select the form of the alternative hypothesis to be tested.
   - Avoid confusion between the *sample* standard deviation and the *population* standard deviation. Values of the population standard deviation are rarely known, so the boxes for population standard deviation are usually left blank.

6. Select the **Method of Analysis**



   - If σ_1 and σ_2 are not known and there is no sound reason to assume that $\sigma_1 = \sigma_2$, select the first option "Not eq. vars: No Pool (which means that the sample variances will not be pooled as described in Section 9–3 of the textbook) as shown below.
   - The option of "Eq. vars: Pool" is used when there is a sound reason to assume that $\sigma_1 = \sigma_2$, so that the sample variances will be pooled to form an estimate of the population variance.
   - The option of "Prelim F–test" is not recommended, but it conducts a preliminary *F* test of the null hypothesis that $\sigma_1 = \sigma_2$ and, based on the results, proceeds with one of these two cases: (1) Do not assume that $\sigma_1 = \sigma_2$ and do not pool the sample variances; (2) Assume that $\sigma_1 = \sigma_2$ and pool the sample variances.

7. Click **Evaluate.**

Statdisk

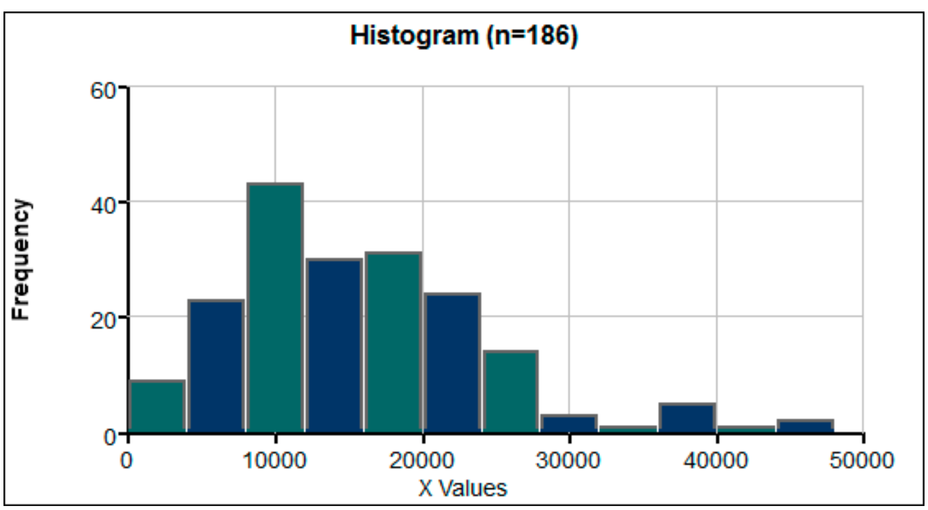
Consider the following exercise from the textbook.

EXAMPLE ***Are Men and Women Equal Talkers?*** A headline in *USA Today* proclaimed that "Men, women are equal talkers." That headline referred to a study of the numbers of words that samples of men and women spoke in a day. Given below are the results from the study, which are included in Appendix B of the Triola textbook as data set *17 – Word Counts by Males and Females* (based on "Are Women Really More Talkative Than Men?" by Mehl, et al, *Science*, Vol. 317, No. 5834). Use a 0.05 significance level to test the claim that men and women speak the same mean number of words in a day. Does there appear to be a difference?

Number of Words Spoken in a Day

Men	Women
$n_1 = 186$	$n_2 = 210$
$\bar{x}_1 = 15{,}668.5$	$\bar{x}_2 = 16{,}215.0$
$s_1 = 8632.5$	$s_2 = 7301.2$

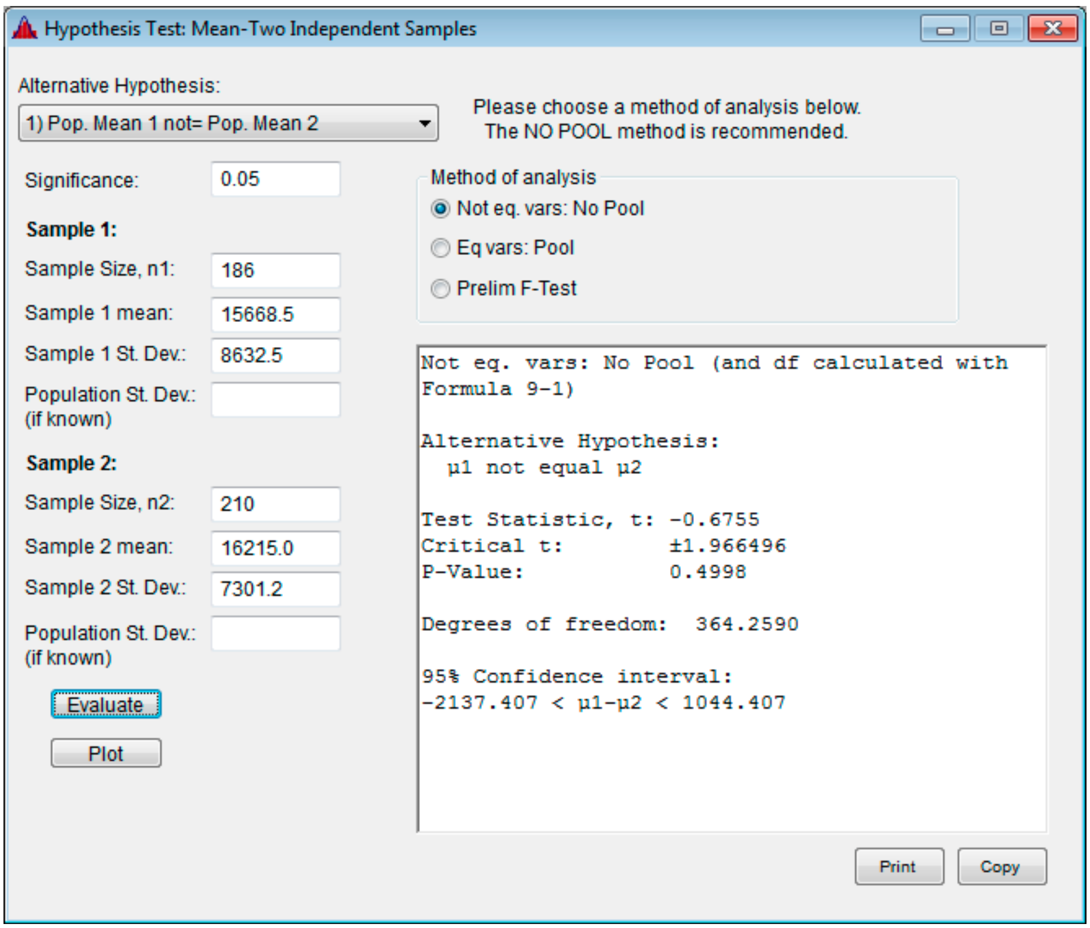
Both samples are large, so it is not necessary to verify that each sample appears to come from a population with a normal distribution, but the accompanying Statdisk display of the histogram for the word counts (in thousands) for men shows that the distribution is not substantially far from being a normal distribution. The histogram for the word counts for the women is very similar.



Statdisk

The Statdisk results are shown below. Because the *P*-value of 0.4998 is greater than the significance level of 0.05, we fail to reject $\mu_1 = \mu_2$. We conclude that there is not sufficient evidence to reject the claim that men and women speak the same mean number of words in a day.

Note that Statdisk uses the complicated Formula 9-1 for computing the number of degrees of freedom, whereas the textbook uses "the smaller of $n_1 - 1$ and $n_2 - 1$." Consequently, Statdisk results will be somewhat different than those obtained by letting the number of degrees of freedom equal to the smaller of $n_1 - 1$ and $n_2 - 1$. By using Formula 9-1, Statdisk provides better results.



Note that the above results also provide the 95% confidence interval estimate of the difference between means (μ_1 - μ_2). This confidence interval includes 0, so there does not appear to be significant difference between the two means.

Statdisk can also determine the confidence interval estimate of the difference between two means using the following procedure. This procedure enables you to specify the confidence level.

Statdisk Procedure for Confidence Interval Estimates of the Difference Between Two Means: Independent Samples

1. For each of the two samples, identify the sample size, sample mean, and sample standard deviation. That is, find the values of n_1, $\bar{x}_1$, s_1, n_2, $\bar{x}_2$, and s_2. (If necessary, use Statdisk's Descriptive Statistics function to find the required sample statistics.)

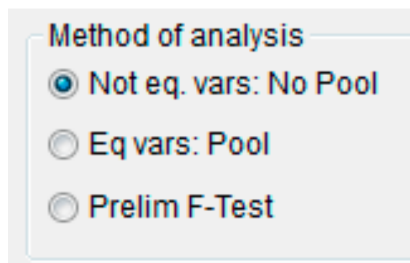
2. Select **Analysis** from the top menu bar.

3. Select **Confidence Intervals** from the submenu.

4. Choose the option of **Mean - Two Independent Samples**.

5. Enter the required values in the dialog box.
   - Enter the confidence level.
   - Enter the statistics for each of the two samples.

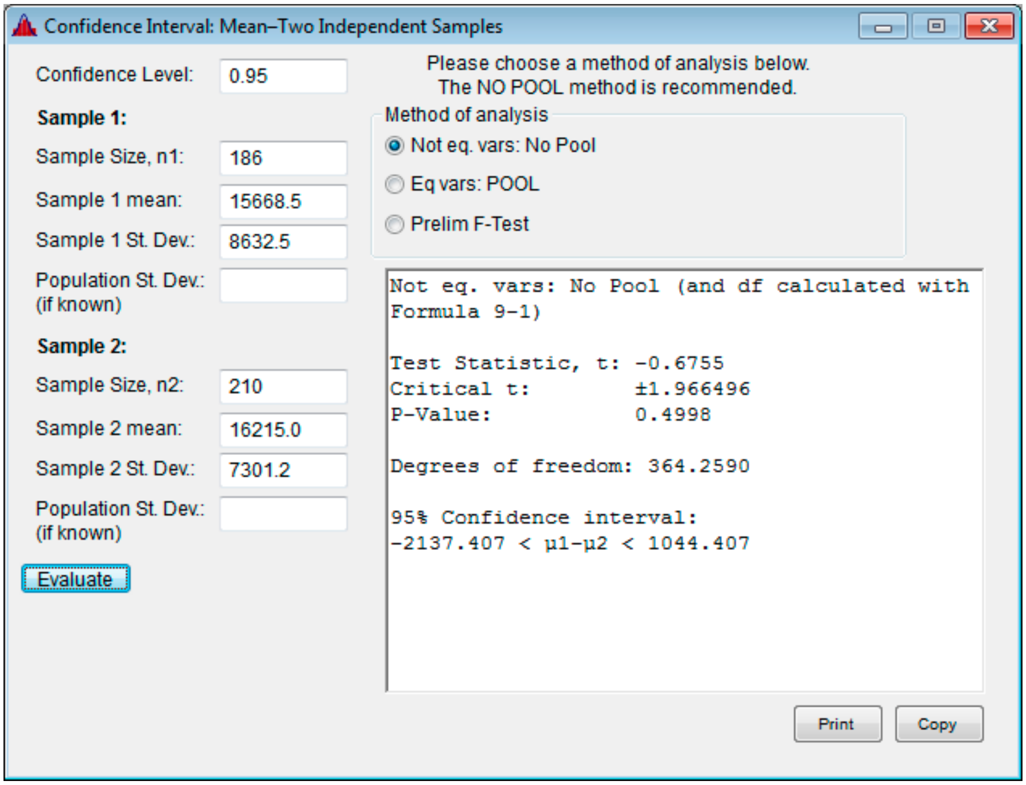
6. Select the **Method of Analysis**
   - If σ_1 and σ_2 are not known and there is no sound reason to assume that $\sigma_1 = \sigma_2$, select the first option "Not eq. vars: No Pool (which means that the sample variances will not be pooled as described in Section 9–3 of the textbook) as shown below.



   - The option of "Eq. vars: Pool" is used when there is a sound reason to assume that $\sigma_1 = \sigma_2$, so that the sample variances will be pooled to form an estimate of the population variance.
   - The option of "Prelim F–test" is not recommended, but it conducts a preliminary *F* test of the null hypothesis that $\sigma_1 = \sigma_2$ and, based on the results, proceeds with one of these two cases: (1) Do not assume that $\sigma_1 = \sigma_2$ and do not pool the sample variances; (2) Assume that $\sigma_1 = \sigma_2$ and pool the sample variances.

7. Click **Evaluate.**

If this procedure is used with the sample data from the preceding example, the 95% confidence interval is included in the following Statdisk display. Because the confidence interval limits do include zero, there does not appear to be significant difference between the two means.



9-3 Two Dependent Samples (Matched Pairs)

Section 9-4 of the Triola textbook describes methods for testing hypotheses and constructing confidence interval estimates of the differences between samples consisting of *matched pairs*. Note that there is a requirement that the number of matched pairs of sample data must be large ($n > 30$) or the pairs of values have differences that are from a population having a distribution that is approximately normal.

Statdisk Procedure for Testing Hypothesis About the Mean of the Differences from Matched Pairs

1. Enter the paired data into two columns of the Statdisk Sample Editor window. Either manually enter the data (if the lists are not very long) or open existing Statdisk data sets.
   - If the number of matched pairs is small ($n \leq 30$), use a histogram or normal quantile plot to verify that the differences appear to come from a population with a normal distribution.

2. Select **Analysis** from the top menu bar.

3. Select **Hypothesis Testing** from the subdirectory.

4. Select **Mean - Matched Pairs**.

5. Make these entries and selections in the dialog box:
   - Select the appropriate format in the "Alternative Hypothesis" box. The default is Mean of Differences not = 0. Scroll through the other 2 possibilities and select the format corresponding to the claim being tested.
   - Enter a significance level, such as 0.05 or 0.01.
   - Select the columns of the data window to be used for the calculations.

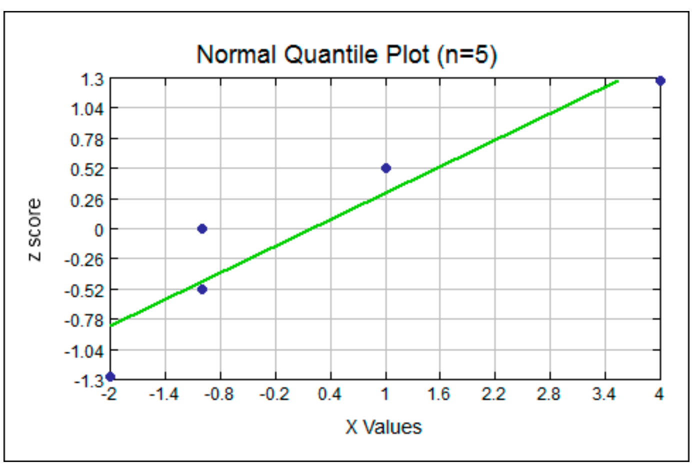
6. Click the **Evaluate** button.

EXAMPLE ***Hypothesis Test of Claimed Freshman Weight Gain*** The table below lists weights (in kg) of a random sample of college Freshman. Use the sample data with a 0.05 significance level to test the claim that for the population of college students, the mean change in weight from September to April is equal to 0 kg.

Weight (kg) Measurements of Students in Their Freshman Year

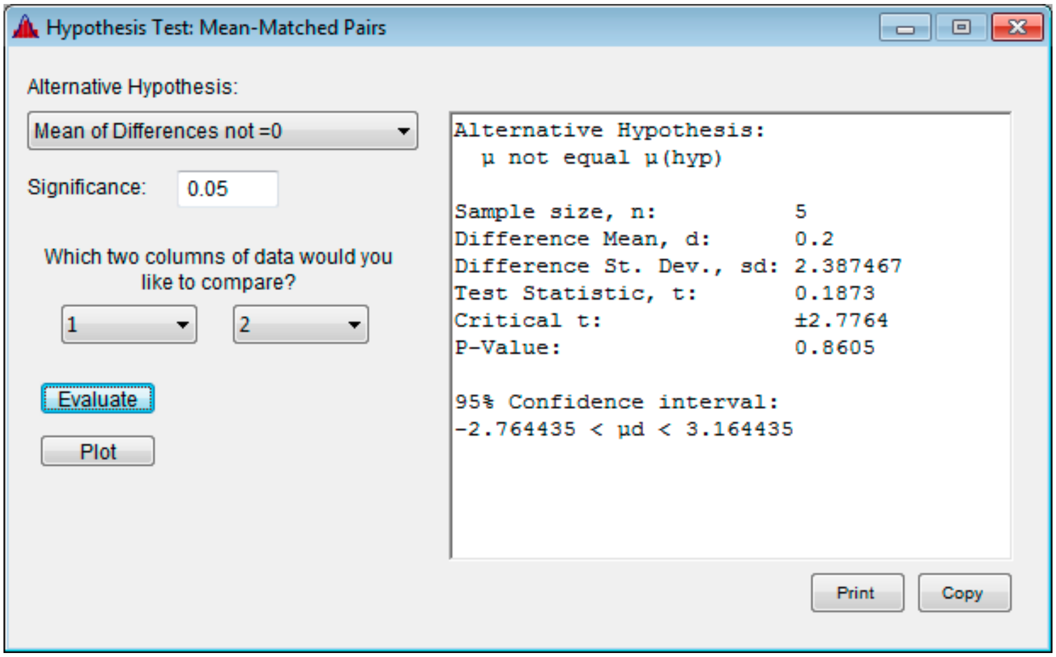
April weight	66	52	68	69	71
September weight	67	53	64	71	70

Because the sample is small, we should verify that the sample differences appear to come from a population with a normal distribution. A normal quantile plot does suggest that the differences appear to be from a normally distributed population.



Statdisk

Using the Statdisk procedure for matched pairs, we get the following display when testing the claim that the mean change in weight from September to April is equal to 0 kg. The *P*–value of 0.8605 is greater than the significance level of 0.05, so we fail to reject the null hypothesis that the mean of the differences between the April weights and the September weights is equal to 0 kg. There does not appear to be a significant difference.



Note that the above results also provide the 95% confidence interval estimate of the mean of the difference from matched pairs (μ_d). The confidence interval limits do contain 0, indicating that the true value of μ_d is not significantly different from 0. We cannot conclude that there is a significant difference between the April weights and September weights.

Statdisk can also determine the confidence interval estimate of the mean of the difference from matched pairs using the following procedure. This procedure enables you to specify the confidence level.

Statdisk Procedure for a Confidence Interval Estimate of the Mean of the Differences from Matched Pairs

1. Enter the paired data into two columns of the Statdisk Sample Editor window. Either manually enter the data (if the lists are not very long) or open existing data sets.
   - If the number of matched pairs is small ($n \leq 30$), use a histogram or normal quantile plot to verify that the differences appear to come from a population with a normal distribution.

2. Select **Analysis** from the top menu bar.

3. Select **Confidence Intervals** from the subdirectory.

4. Select **Mean - Matched Pairs**.

5. Enter a confidence level and select the columns containing the sample data.

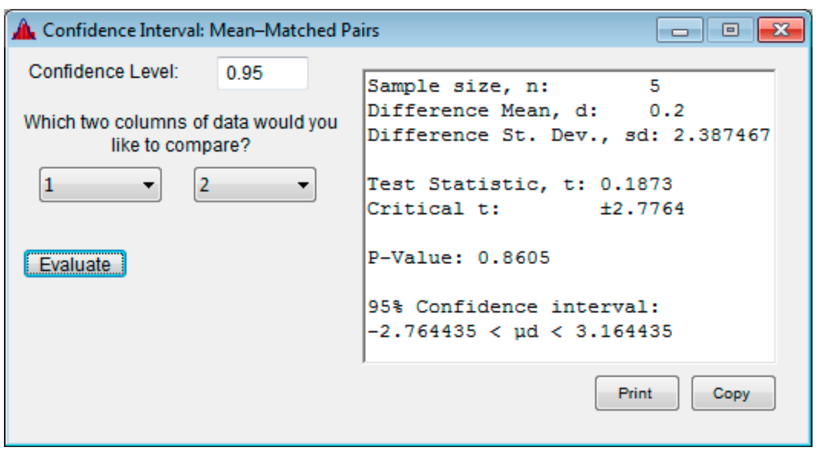
6. Click the **Evaluate** button.

Shown below is the Statdisk display that includes a 95% confidence interval based on the data in the following table.

Weight (kg) Measurements of Students in Their Freshman Year

April weight	66	52	68	69	71
September weight	67	53	64	71	70

Note that the confidence interval limits do contain 0, indicating that the true value of μ_d is not significantly different from 0. We cannot conclude that there is a significant difference between the April weights and September weights.



9-4 Two Variances

Section 9-5 in the textbook describes the use of the *F* distribution in testing a claim that two populations have the same variance (or standard deviation). Statdisk can conduct such tests. Section 9-5 in the textbook focuses on hypothesis tests, but Statdisk can be used to find confidence intervals for the ratio σ_1^2/σ_2^2 and also for the ratio σ_1/σ_2.

Before using the Statdisk procedures for testing hypotheses or constructing confidence intervals, the requirements should be verified. When making inferences about two standard deviations or variances, we require that the two samples are independent. Also, the two populations must be *normally distributed*. Verification of normality is important because the methods of this section are not robust, meaning that they are extremely sensitive to departures from normality. Use Statdisk's *Normality Assessment* feature to determine whether the two samples are from normally distributed populations. The *Normality Assessment* feature is described in Section 6-4 of this workbook.

Statdisk Procedure for Hypothesis Tests About

Two Standard Deviations or Two Variances

1. For each of the two samples, obtain the sample size *n* and the sample standard deviation *s*. (If you have two samples of raw data, you can find these statistics using Statdisk's Descriptive Statistics function. Also, if the available information includes sample sizes and sample *variances*, be sure to take the square root of the sample variances to obtain the sample *standard deviations*.)

2. Select **Analysis** from the top menu bar.

3. Select **Hypothesis Testing** from the subdirectory.

4. Select **St. Dev. Two Samples**.

5. Make the following entries in the dialog box:
   - In the "Alternative Hypothesis" box, select the case corresponding to the claim being tested.
   - Enter a significance level, such as 0.05 or 0.01.
   - In the appropriate boxes, enter the sample size and sample standard deviation for the first sample, then do the same for the second sample.

6. Click on the **Evaluate** button to obtain the test results.

Statdisk

Consider Example 1 from Section 9-5 of the textbook.

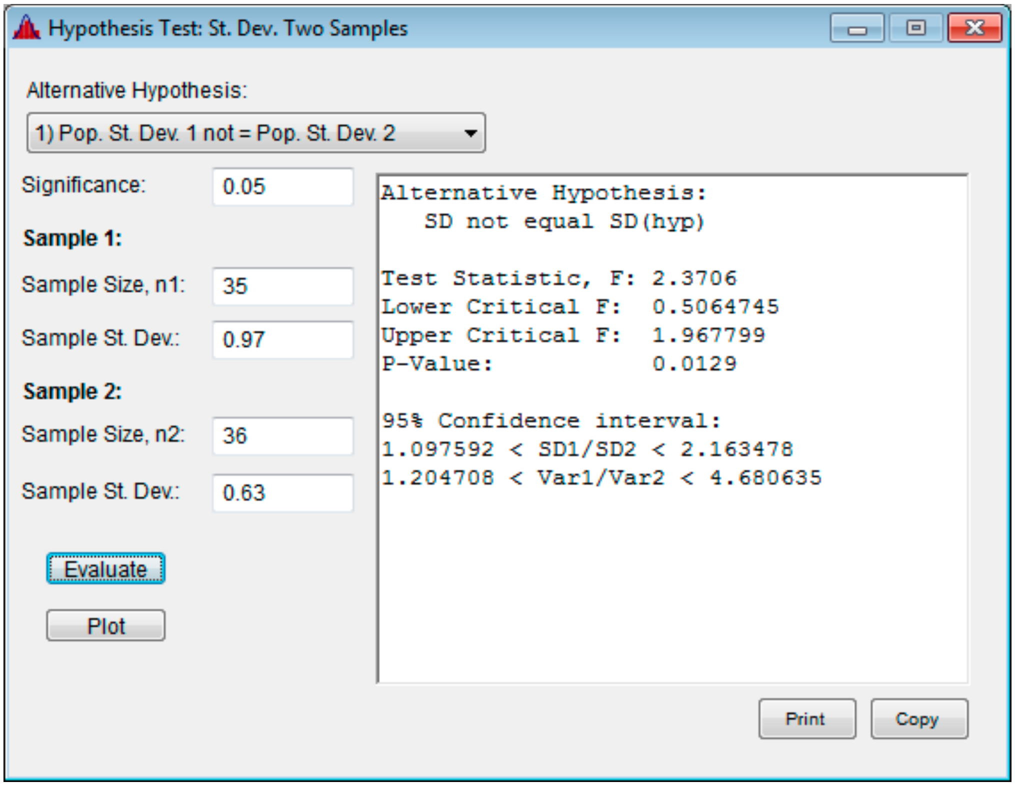
EXAMPLE Use the following sample data to test the claim that subjects tested with a red background have creativity scores with a standard deviation equal to the standard deviation for those tested with a blue background. Use a 0.05 significance level.

Creativity Scores	
Red Background:	$n = 35, \bar{x} = 3.39, s = 0.97$
Blue Background:	$n = 36, \bar{x} = 3.97, s = 0.63$

Using the above Statdisk procedure, we will use the sample data with a 0.05 significance level to test the claim that both samples of creativity scores are from populations with the same standard deviation. The procedure described in the Triola textbook requires that the sample with the larger variance be designated as Sample 1, but this is not necessary with Statdisk. Statdisk automatically does the required calculations and it correctly handles cases in which the first sample has a variance smaller than the second sample.

The following Statdisk display shows a *P*–value of 0.0129. Because that *P*–value is less than the significance level of 0.05, we reject the null hypothesis. There is sufficient evidence to warrant rejection of the claim that the two standard deviations are equal. The variation among creativity scores for those with a red background appears to be different from the variation among creativity scores for those with a blue background.

The results also provide the test statistic F = 2.3706 and the upper and lower critical F values. Figure 9-6 in the textbook shows that the test statistic F = 2.3706 does fall within the critical region, so we reject the null hypothesis of equal standard deviations.



Because a confidence interval based on two standard deviations or variations consists of confidence interval limits for the *ratio* of the population standard deviations or variances, such confidence intervals are not used as often as those involving two means.

Statdisk Procedure for Confidence Interval Estimates of the Ratio of Two Standard Deviations or Variances

1. For each of the two samples, obtain the sample size *n*, and the sample standard deviation *s*. (If you have two samples of raw data, you can find these statistics using the Descriptive Statistics module. Also, if the available information includes sample sizes and sample *variances*, be sure to take the square root of the sample variances to obtain the sample *standard deviations*.)

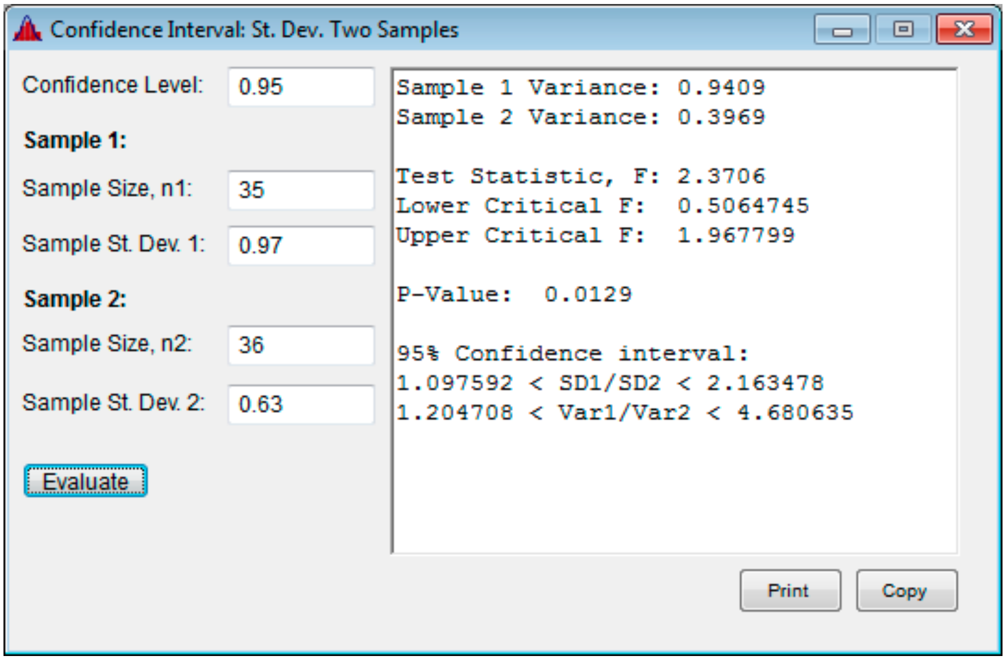
2. Select **Analysis** from the top menu bar.

3. Select **Confidence Intervals** from the subdirectory.

4. Select **St. Dev. Two Samples.**

5. Enter a confidence level and the sample statistics.

6. Click on the **Evaluate** button to obtain the test results.



CHAPTER 9 EXPERIMENTS: Inferences from Two Samples

9.1 **Drug Use in College** In a 1993 survey of 560 college students, 171 said that they used illegal drugs during the previous year. In a recent survey of 720 college students, 263 said that they used illegal drugs during the previous year (based on data from the National Center for Addiction and Substance Abuse at Columbia University). Use a 0.05 significance level to test the claim that the proportion of college students using illegal drugs in 1993 was less than it is now.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-2 **Drug Use in College** Using the sample data from Experiment 9-1, construct the confidence interval corresponding to the hypothesis test conducted with a 0.05 significance level. What conclusion does the confidence interval suggest?

____

____

9-3 **Are Seat Belts Effective?** A simple random sample of front-seat occupants involved in car crashes is obtained. Among 2823 occupants not wearing seatbelts, 31 were killed. Among 7765 occupants wearing seatbelts, 16 were killed (based on data from "Who Wants Airbags?" by Meyer and Finney, Chance, Vol. 18, No. 2). Construct a 90% confidence interval estimate of the difference between the fatality rates for those not wearing seat belts and those wearing seat belts. What does the result suggest about the effectiveness of seat belts?

____

____

Statdisk

9-4 **Are Seat Belts Effective?** Use the sample data in Experiment 9-3 with a 0.05 significance level to test the claim that the fatality rate is higher for those not wearing seat belts.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-5 **Morality and Marriage** A Pew Research Center poll asked randomly selected subjects if they agreed with the statement that "It is morally wrong for married people to have an affair." Among the 386 women surveyed, 347 agreed with the statement. Among the 359 men surveyed, 305 agreed with the statement. Use a 0.05 significance level to test the claim that the percentage of women who agree is different from the percentage of men who agree. Does there appear to be a difference in the way women and men feel about this issue?

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-6 **Morality and Marriage** Using the sample data from Experiment 9-5, construct the confidence interval corresponding to the hypothesis test conducted with a 0.05 significance level. What conclusion does the confidence interval suggest?

____

____

9-7 **Cardiac Arrest at Day and Night** A study investigated survival rates for in-hospital patients who suffered cardiac arrest. Among 58,593 patients who had cardiac arrest during the day, 11,604 survived and were discharged. Among 28,155 patients who suffered cardiac arrest at night, 4139 survived and were discharged (based on data from "Survival from In-Hospital Cardiac Arrest During Nights and Weekends," by Peberdy et al., *Journal of the American Medical Association,* Vol. 299, No. 7). Use a 0.01 significance level to test the claim that the survival rates are the same for day and night.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-8 **Cardiac Arrest at Day and Night** Using the sample data from Experiment 9-7, construct the confidence interval corresponding to the hypothesis test conducted with a 0.01 significance level. What conclusion does the confidence interval suggest?

____

____

In Experiments 9–9 through 9–20, assume that the two samples are independent simple random samples selected from normally distributed populations. Do not assume that the population standard deviations are equal.

9-9 **Hypothesis Test of Effectiveness of Humidity in Treating Croup** Treatment In a randomized controlled trial conducted with children suffering from viral croup, 46 children were treated with low humidity while 46 other children were treated with high humidity. Researchers used the Westley Croup Score to assess the results after one hour. The low humidity group had a mean score of 0.98 with a standard deviation of 1.22 while the high humidity group had a mean score of 1.09 with a standard deviation of 1.11 (based on data from "Controlled Delivery of High vs Low Humidity vs Mist Therapy for Croup Emergency Departments," by Scolnik et al, *Journal of the American Medical Association,* Vol. 295, No. 11). Use a 0.05 significance level to test the claim that the two groups are from populations with the same mean. What does the result suggest about the common treatment of humidity?

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-10 **Confidence Interval for Effectiveness of Humidity in Treating Croup** Use the sample data given in Experiment 9-9 and construct a 95% confidence interval estimate of the difference between the mean Westley Croup Score of children treated with low humidity and the mean score of children treated with high humidity. What does the confidence interval suggest about humidity as a treatment for croup?

____

____

9-11 **Confidence Interval for Cigarette Tar** The mean tar content of a simple random sample of 25 unfiltered king-size cigarettes is 21.1 mg, with a standard deviation of 3.2 mg. The mean tar content of a simple random sample of 25 filtered 100 mm cigarettes is 13.2 mg with a standard deviation of 3.7 mg (based on data from 12th Edition data set *10 – Cigarette Tar, Nicotine, and Carbon Monoxide*). Construct a 90% confidence interval estimate of the difference between the mean tar content of unfiltered king size cigarettes and the mean tar content of filtered 10mm cigarettes. Does the result suggest that 100 mm filtered cigarettes have less tar than unfiltered king size cigarettes?

____

____

9-12 **Hypothesis Test for Cigarette Tar** Refer to the sample data in Experiment 9-11 and use a 0.05 significance level to test the claim that unfiltered king size cigarettes have a mean tar content greater than that of filtered 100 mm cigarettes. What does the result suggest about the effectiveness of cigarette filters?

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-13 **BMI for Miss America** The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) for Miss America winners from two different time periods. Consider the listed values to be simple random samples selected from larger populations.

BMI (from recent winners): 19.5 20.3 19.6 20.2 17.8 17.9 19.1
18.8 17.6 16.8

BMI (from the 1920’s and 1930’s): 20.4 21.9 22.1 22.3 20.3 18.8 18.9
19.4 18.4 19.1

a. Use a 0.05 significance level to test the claim that recent winners have a lower mean BMI than winners from the 1920’s and 1930’s.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Statdisk

Conclusion in your own words: ____

____

b. Construct a 90% confidence interval for the difference between the mean BMI of recent winners and the mean BMI of winners from the 1920’s and 1930’s.

____

____

9-14 **Radiation in Baby Teeth** Listed below are amounts of Strontium-90 (in millibecquerels or mBq per gram of calcium) in a simple random sample of baby teeth obtained from Pennsylvania residents and New York residents born after 1979 (based on data from “An Unexpected Rise in Strontium-90 in U.S. Deciduous Teeth in the 1990s,” by Mangano, et. al., *Science of the Total Environment*).

Pennsylvania:	155	142	149	130	151	163	151	142	156	133	138	161
New York:	133	140	142	131	134	129	128	140	140	140	137	143

a. Use a 0.05 significance level to test the claim that the mean amount of Strontium-90 from Pennsylvania residents is greater than the mean amount from New York residents.

Test statistic: ___________ Critical value(s): ___________ *P*–value: _____________

Conclusion in your own words: ____

____

b. Construct a 90% confidence interval of the difference between the mean amount of Strontium-90 from Pennsylvania residents and the mean amount from New York residents.

____

____

9-15 **Longevity** Listed below are the numbers of years that popes and British monarchs (since 1690) lived after their election or coronation (based on data from *Computer-Interactive Data Analysis*, by Lunn and McNeil, John Wiley & Sons). Treat the values as simple random samples from a larger population.

Popes: 2 9 21 3 6 10 18 11 6 25 23 6 2 15 32 25
11 8 17 19 5 15 0 26

Kings and Queens: 17 6 13 12 13 33 59 10 7 63 9 25 36 15

a. Use a 0.01 significance level to test the claim that the mean longevity for popes is less the mean for British monarchs after coronation.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

b. Construct a 98% confidence interval of the difference between the mean longevity of popes and the mean longevity for kings and queens. What does the result suggest about those two means?

____

____

9-16 **Sex and Blood Cell Counts** White blood cell counts are helpful for assessing liver disease, radiation, bone marrow failure, and infectious diseases. Listed below are white blood cell counts found in simple random samples of males and females (based on data from the Third National Health and Nutrition Examination Survey).

Female: 8.90 6.50 9.45 7.65 6.40 5.15 16.60 5.75 11.60 5.90 9.30 8.55
10.80 4.85 4.90 8.75 6.90 9.75 4.05 9.05 5.05 6.40 4.05 7.60
4.95 3.00 9.10

Male: 5.25 5.95 10.05 5.45 5.30 5.55 6.85 6.65 6.30 6.40 7.85 7.70 5.30
6.50 4.55 7.10 8.00 4.70 4.40 4.90 10.75 11.00 9.60

a. Use a 0.01 significance level to test the claim that females and males have different mean white blood cell counts.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

b. Construct a 98% confidence interval of the difference between the mean white blood cell count of females and males. Based on the result, does there appear to be a difference?

____

____

9-17 **Baseline Characteristics** Reports of results from clinical trials often include statistics about "baseline characteristics," so that we can see that different groups have the same basic characteristics. Refer to Data Set 1 in Appendix B and construct a 95% confidence interval estimate of the difference between the mean age of men and the mean age of women. Based on the result, does it appear that the sample of men and the sample of women are from populations with the same mean?

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

b. Construct a 95% confidence interval estimate of the difference between the mean age of men and the mean age of women. What does the confidence interval suggest?

____

____

9-18 **Statdisk data set: Word Counts** Refer to Appendix B in the Triola textbook for data set *17 – Word Counts by Males and Females*. Use the word counts for male and female psychology students recruited in Mexico (see the columns labeled M3 and F3).

a. Use a 0.05 significance level to test the claim that male and female psychology students speak the same mean number of words in a day.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

b. Construct a 95% confidence interval estimate of the difference between the mean number of words spoken in a day by male and female psychology students in Mexico. Do the confidence interval limits include 0, and what does that suggest about the two means?

____

____

9-19 **Statdisk data set: Voltage** Refer to Appendix B in the Triola textbook for data set *18 – Voltage Measurements from a Home*. Use a 0.05 significance level to test the claim that the sample of home voltages and the sample of generator voltages are from populations with the same mean. If there is a statistically significant difference, does that difference have practical significance?

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-20 **Statdisk data set: Weights of Coke** Refer to Appendix B in the Triola textbook for data set*19 – Cola Weights and Volumes* and test the claim that because they contain the same amount of cola, the mean weight of cola in cans of regular Coke is the same as the mean weight of cola in cans of Diet Coke. If there is a difference in the mean weights, identify the most likely explanation for that difference.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

Statdisk

9-21 **Does BMI Change During Freshman Year?** Listed below are body mass indices (BMI) of students. The BMI of each student was measured in September and April of the freshman year (based on data from "Changes in Body Weight and Fat Mass of Men and Women in the First Year of College: A Study of the 'Freshman 15'," by Hoffman, Policastro, Quick, and Lee, *Journal of American College Health*, Vol. 55, No. 1). Use a 0.05 significance level to test the claim that the mean change in BMI for all students is equal to 0. Does BMI appear to change during freshman year?

April BMI	20.15	19.24	20.77	23.85	21.32
September BMI	20.68	19.48	19.59	24.57	20.96

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-22 **Confidence Interval for BMI Changes** Use the same paired data from Experiment 9-21 to construct a 95% confidence interval estimate of the change in BMI during freshman year. Does the confidence interval include 0, and what does that suggest about BMI during freshman year?

____

____

9-23 **Are Best Actresses Younger than Best Actors?** Listed below are ages of actresses and actors at the times that they won Oscars. The data are paired according to the years that they won. Use a 0.05 significance level to test the common belief that best actresses are younger than best actors. Does the result suggest a problem in our culture?

Best Actresses	28	32	27	27	26	24	25	29	41	40	27	42	33	21	35
Best Actors	62	41	52	41	34	40	56	41	39	49	48	56	42	62	29

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

Statdisk

9-24 **Are Flights Cheaper When Scheduled Earlier?** Listed below are the costs (in dollars) of flights from New York (JFK) to San Francisco for US Air, Continental, Delta, United, American, Alaska, and Northwest. Use a 0.01 significance level to test the claim that flights scheduled one day in advance cost more than flights scheduled 30 days in advance. What strategy appears to be effective in saving money when flying?

Flight scheduled one day in advance	456	614	628	1088	943	567	536
Flight scheduled 30 days in advance	244	260	264	264	278	318	280

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-25 **Does Your Body Temperature Change During the Day?** Listed below are body temperatures (in ^oF) of subjects measured at 8 AM and at 12 AM (from University of Maryland physicians listed in 12^{th} Edition data set *3 – Body Temperatures of Healthy Adults*). Construct a 95% confidence interval estimate of the difference between the 8 AM temperatures and the 12 AM temperatures. Is body temperature basically the same at both times?

8 AM	97.0	96.2	97.6	96.4	97.8	99.2
12 AM	98.0	98.6	98.8	98.0	98.6	97.6

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-26 **Is Blood Pressure the Same for Both Arms?** Listed below are systolic blood pressure measurements (mm Hg) taken from the right and left arms of the same woman (based on data from "Consistency of Blood Pressure Differences Between the Left and Right Arms," by Eguchi et al, *Archives of Internal Medicine*, Vol. 167). Use a 0.05 significance level to test for a difference between the measurements from the two arms. What do you conclude?

Right arm	102	101	94	79	79
Left arm	175	169	182	146	144

Statdisk

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-27 **Statdisk data set: Paper or Plastic?** Refer to Appendix B in the Triola textbook for data set *23 –Weights of Discarded Garbage for one Week*. Construct a 95% confidence interval estimate of the mean of the differences between weights of discarded paper and weights of discarded plastic. Which seems to weigh more: discarded paper or discarded plastic?

____

____

9-28 **Statdisk data set: Glass and Food** Refer Appendix B in the Triola textbook for data set *23 –Weights of Discarded Garbage for one Week*. Construct a 95% confidence interval estimate of the mean of the differences between weights of discarded glass and weights of discarded food. Which seems to weigh more: discarded glass or discarded food?

____

____

9-29 **Testing Effects of Alcohol** Researchers conducted an experiment to test the effects of alcohol. The errors were recorded in a test of visual and motor skills for a treatment group of 22 people who drank ethanol and another group of 22 people given a placebo. The errors for the treatment group have a standard deviation of 2.20, and the errors for the placebo group have a standard deviation of 0.72 (based on data from "Effects of Alcohol Intoxication on Risk Taking, Strategy, and Error Rate in Visuomotor Performance," by Streufert, et al., *Journal of Applied Psychology,* Vol. 77, No. 4). Use a 0.05 significance level to test the claim that the treatment group has errors that vary more than the errors of the placebo group

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-30 **Heights** Listed below are heights (cm) of randomly selected females and males taken from Appendix B in the Triola textbook for data set *1 – Body Measurements Female/Male* in Appendix B. Use a 0.05 significance level to test the claim that females and males have heights with the same amount of variation.

Female	163.7	165.5	163.1	166.3	163.6	170.9	153.5	155.7	153.0	157.0
Male	178.8	177.5	187.8	172.4	181.7	169.0	186.9	183.1	176.4	183.4

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-31 **Freshman 15 Study**Use the sample weights (kg) of male and female college students measured in April of their freshman year, as listed in Appendix B in the Triola textbook for data set *4 – Freshman 15 Data*. Use a 0.05 significance level to test the claim that near the end of the freshman year, weights of male college students vary more than weights of female college students.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

9-32 **M&Ms** Refer to Appendix B in the Triola textbook for data set *20 – M and M Plain Candy Weights* and use the weights (g) of the red M&Ms and the orange M&Ms. Use a 0.05 significance level to test the claim that the two samples are from populations with the same amount of variation.

Test statistic: __________ Critical value(s): __________ *P*–value: ____________

Conclusion in your own words: ____

____

10 Correlation and Regression

Statdisk

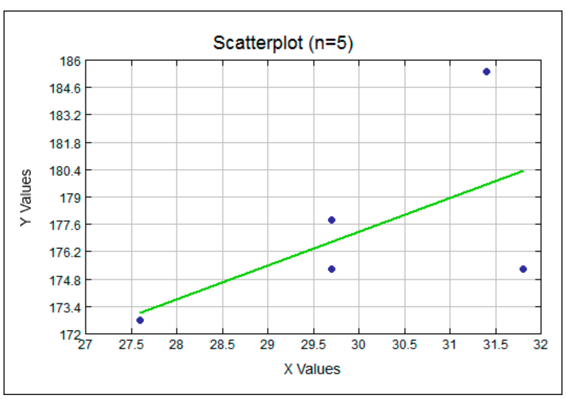
10-1 Scatterplot

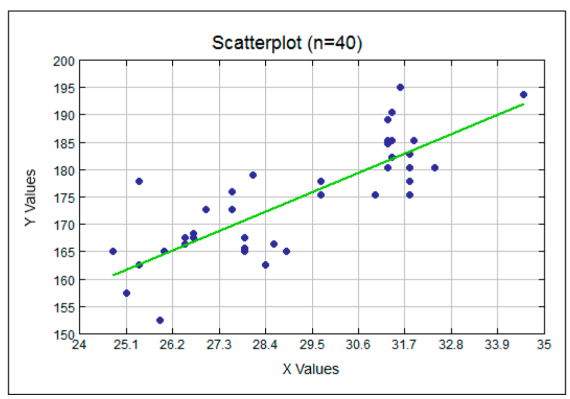
Chapter 10 in the Triola textbook introduces the basic concepts of linear correlation and regression. The basic objective is to use paired sample data to determine whether there is a relationship between two variables and, if so, identify what the relationship is. Consider the paired sample data in the table below. The data consist of measured shoe print lengths and heights of five males. We want to determine whether there is a relationship between shoe print length and height. If such a relationship exists, we want to identify it with an equation so that we can predict the height of a male based on the length of his shoe print.

Table 10-1 Shoe Print Lengths and Heights of Males

Shoe Print (cm)	29.7	29.7	31.4	31.8	27.6
Height (cm)	175.3	177.8	185.4	175.3	172.7

The data from the table is shown on the scatterplot on the left, which suggests there might be a pattern, but it isn't very strong. The scatterplot on the right shows the data of all 40 pairs of shoe print and height measurements from the Statdisk data set (*Elementary Statistics 12th Edition*) *2 - Foot and Height Measurements*. This scatterplot provides better evidence that a relationship may exist between the two variables. (The Statdisk procedure for generating scatterplots was introduced in Section 2-4 of this manual/workbook.)





10-2 Correlation and Regression

Sections 10-2, 10-3, and 10-4 in the textbook introduce the basic concepts of linear correlation and regression. The basic objective is to use sample paired data to determine whether there is a relationship between two variables and, if so, identify what the relationship is.

The *linear correlation coefficient r* is a measure of the strength of the linear relationship between two variables. Statdisk can compute the value of *r* for paired data using the procedure that follows.

Given a collection of paired sample data, the *regression equation*

$$\hat{y} = b_0 + b_1x$$

describes the relationship between the two variables. The graph of the regression equation is called the *regression line* (or *line of best fit*, or *least-squares line*). The regression equation expresses a relationship between *x* (called the *independent variable* or *predictor variable*) and *y* (called the *dependent variable* or *response variable*). The *y*-intercept of the line is b_0 and the slope is b_1. Given a collection of paired data, Statdisk can find the regression equation as follows.

Statdisk Procedure for Correlation and Regression

1. Enter the paired data into two columns of the Statdisk Sample Editor. Either manually enter the data (if the lists are not very long) or open existing s.
   - In this example we will open the *Elementary Statistics 12th Edition 2 - Foot and Height Measurements.*

2. Select **Analysis** from the top menu bar.

3. Select the menu item of **Correlation and Regression**.

4. Select a significance level, such as 0.05 or 0.01.

5. Proceed to select the columns to be used for the *x* variable and the *y* variable.

6. Click the **Evaluate** button to get the correlation/regression results.

7. Click the **Scatterplot** button to get a graph of the scatterplot.

8. Click the Residual Plot button to get a graph of residuals versus x.

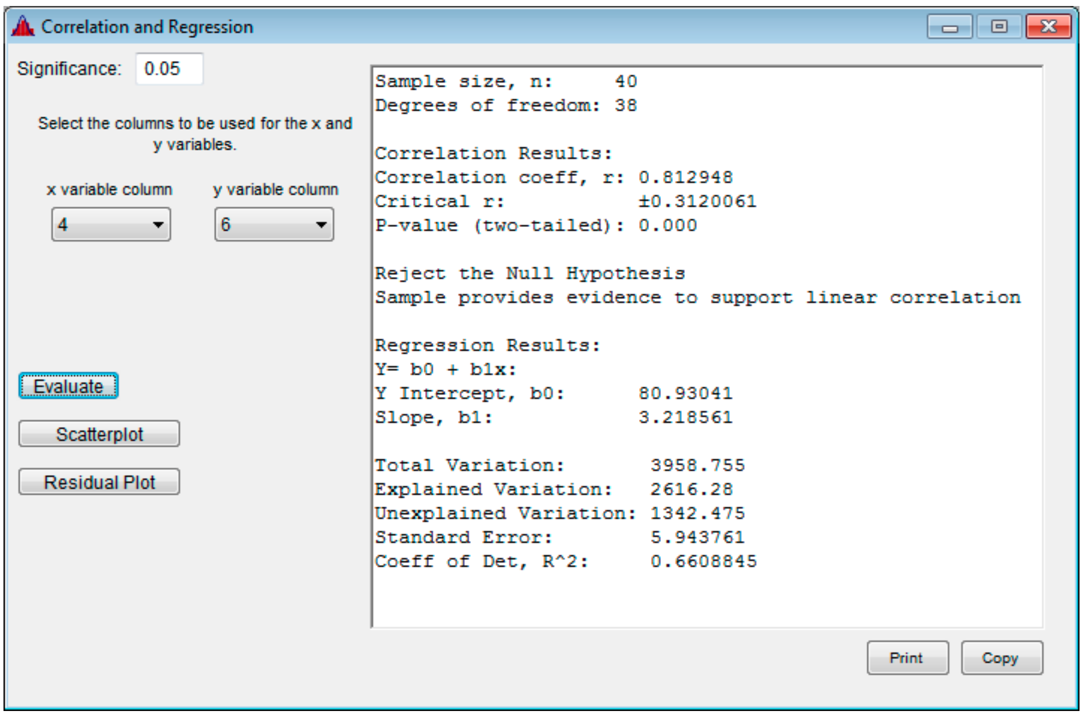
Correlation
If you follow the above steps using *2 - Foot and Height Measurements*, the Statdisk results will be as shown on the following page. Also shown is the scatterplot that is obtained by clicking on the **Plot** button in the display window. The results include the linear correlation coefficient of r = 0.812948, the critical values of $r = \pm 0.3120061$, the P-value of 0.000, and the conclusion that there is sufficient evidence to reject the null hypothesis (of no correlation) and support a claim of a linear correlation between the two variables. There does appear to be a linear correlation between shoe print length and height.

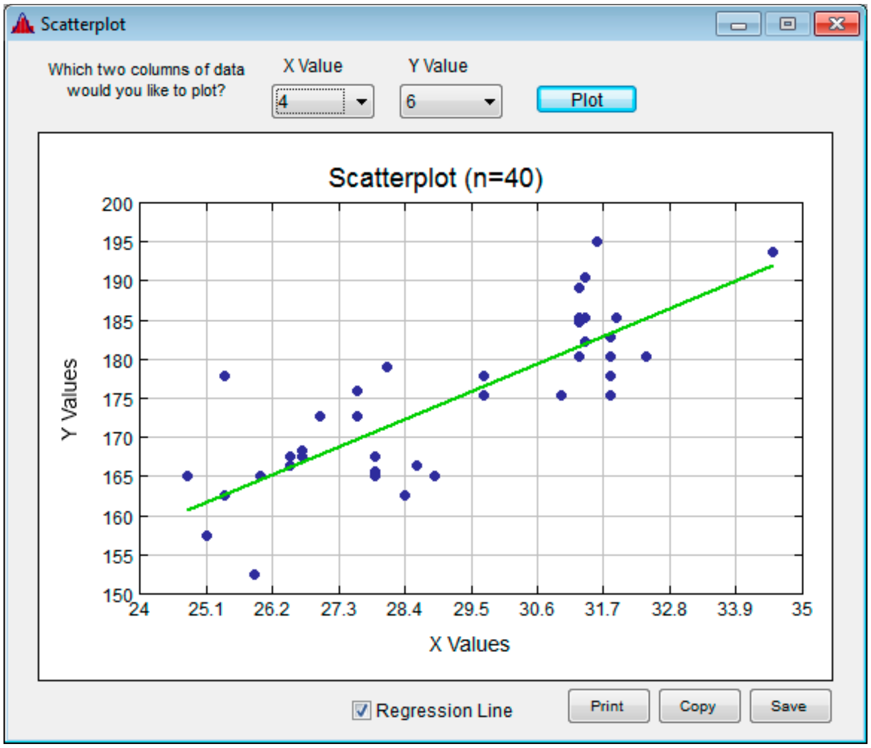
Regression
Also included in the display shown below are the *y*-intercept b_0 and slope b_1 of the estimated regression line. Using the Statdisk results, the estimated regression equation is

$$\hat{y} = 80.9 + 3.219x$$

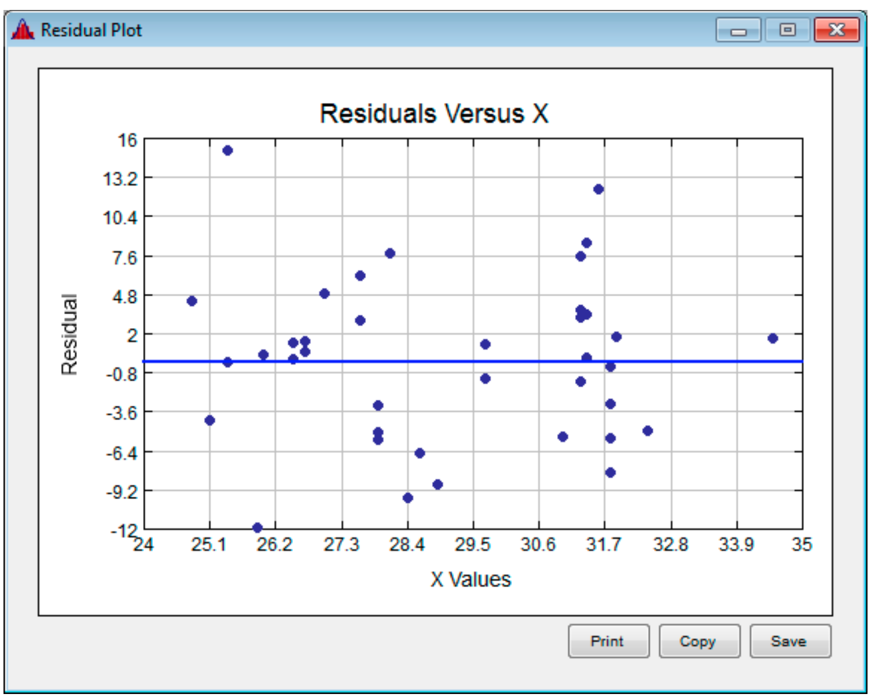
The graph of the scatterplot (shown on the next page) includes the regression line.

The residual plot provides a scatterplot of the (*x*, *y*) values after each of the *y*-coordinate values has been replaced by the residual value $y - \hat{y}$ (where $\hat{y}$ denotes the predicted value of *y*). There is no noticeable pattern in the residual plot, suggesting that the regression equation is a good model.





In the display of the scatterplot, note that the **Regression Line** box at the bottom of the screen is checked. The Regression Line button is checked by default. If you click on that box to remove the check mark, the graph of the regression line will not be included in the scatterplot.



Statdisk

10-3 Multiple Regression

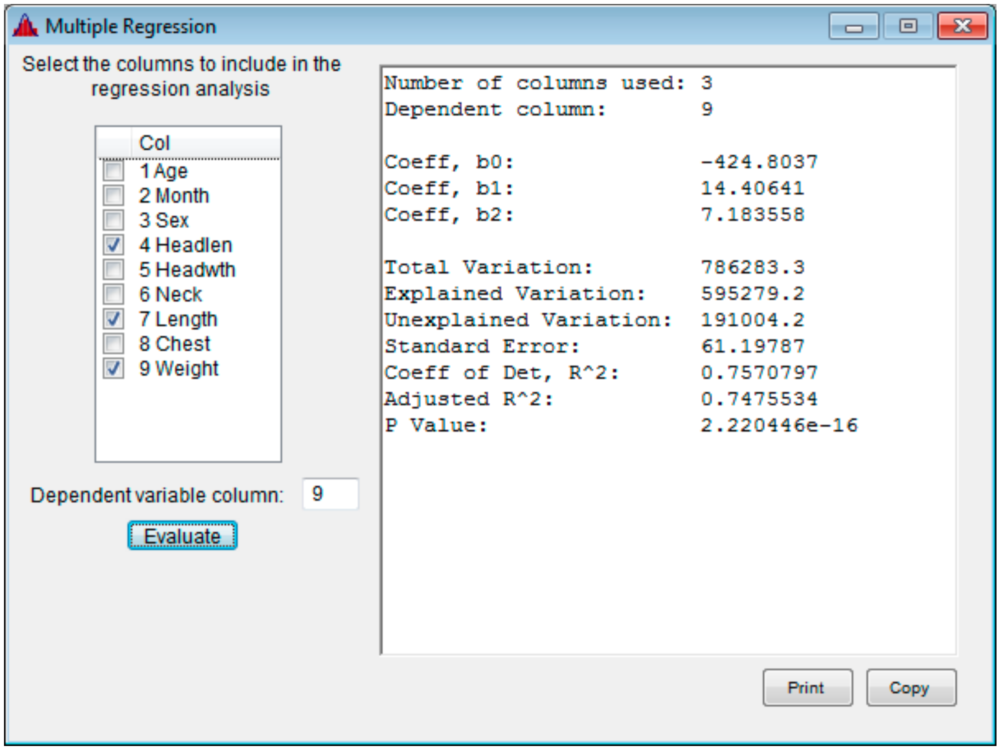
Section 10-5 of the textbook discusses multiple regression, and Statdisk does allow you to obtain multiple regression results. Once a collection of sample data has been entered, you can easily experiment with different combinations of columns to find the combination that is best.

Statdisk Procedure for Multiple Regression

1. Either enter the data in columns of the Statdisk Data Window, or open an existing .
2. Select **Analysis** from the top menu bar.
3. Select **Multiple Regression** from the menu.
4. Select the columns to be included, and identify the column to be used for the dependent variable. (If a column has a check mark but you don't want it included, click on the check mark so that it is removed and its column is excluded.) See the screen below, where columns 1, 3, and 6 are included, with column 1 selected for the dependent variable.
5. Click on **Evaluate**.
6. To use a different combination of variables, simply click on different combinations of columns.

EXAMPLE Consider the Statdisk *12th Edition 7 – Bear Measurements*. We will use Statdisk to obtain the multiple regression equation with WEIGHT (bear weight) as the dependent variable and the measurements of Headlen (head length in inches) and Length (body length in inches) as the two independent variables.

In the Statdisk display, note the selection of columns 4, 7, and 9 at the left, and note the selection of column 9 as the dependent variable.



The results in the above Statdisk display include the intercept b_0 = -424.80 (rounded), the coefficient b_1 = 14.41 (rounded), and the coefficient b_2 = 7.18 (rounded). These values are included in the multiple regression equation as shown here:

$$\hat{y} = -424.80 + 14.41x_1 + 7.18x_2$$

or WEIGHT = -424.80 + 14.41 Headlen + 7.18 Length

The results also include the adjusted coefficient of determination (Adjusted R^2 = 0.7476), as well as the *P*-value of 2.220446e-16. The small *P*-value suggests that the multiple regression equation is a good model for the data.

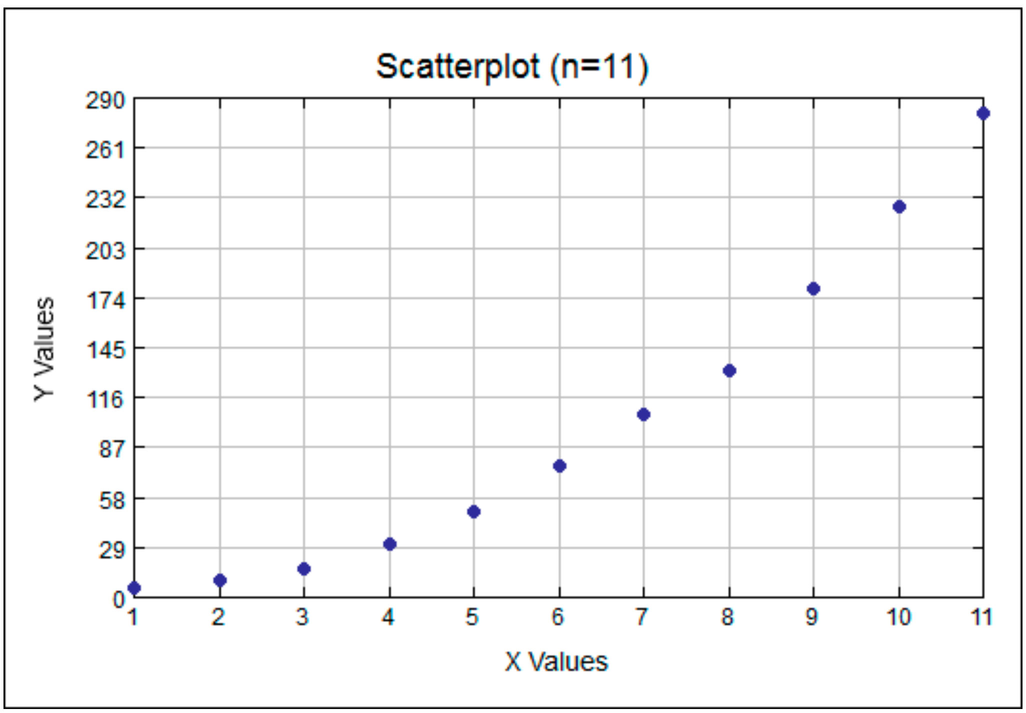
10-4 Nonlinear Regression

Nonlinear regression is discussed in the Triola statistics textbooks (except *Essentials of Statistics*). The objective is to find a mathematical function that "fits" or describes real-world data. Among the models discussed in the textbook, we will describe how Statdisk can be used for the linear, quadratic, logarithmic, exponential, and power models.

Consider the sample data in Table 10-7 from *Elementary Statistics* 12th edition, reproduced below. As in the textbook, we use the coded year values for *x*, so *x* = 1, 2, 3, . . . , 11. The *y* values are the populations (in millions) of 5, 10, 17, . . . , 281. The Statdisk scatterplot is displayed.

Table 10-7 Population (in millions) of the United States

Year	1800	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
Coded Year	1	2	3	4	5	6	7	8	9	10	11
Population	5	10	17	31	50	76	106	132	179	227	281

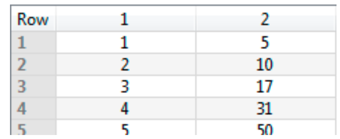


Statdisk

Linear Model: *y* = *a* + *bx*

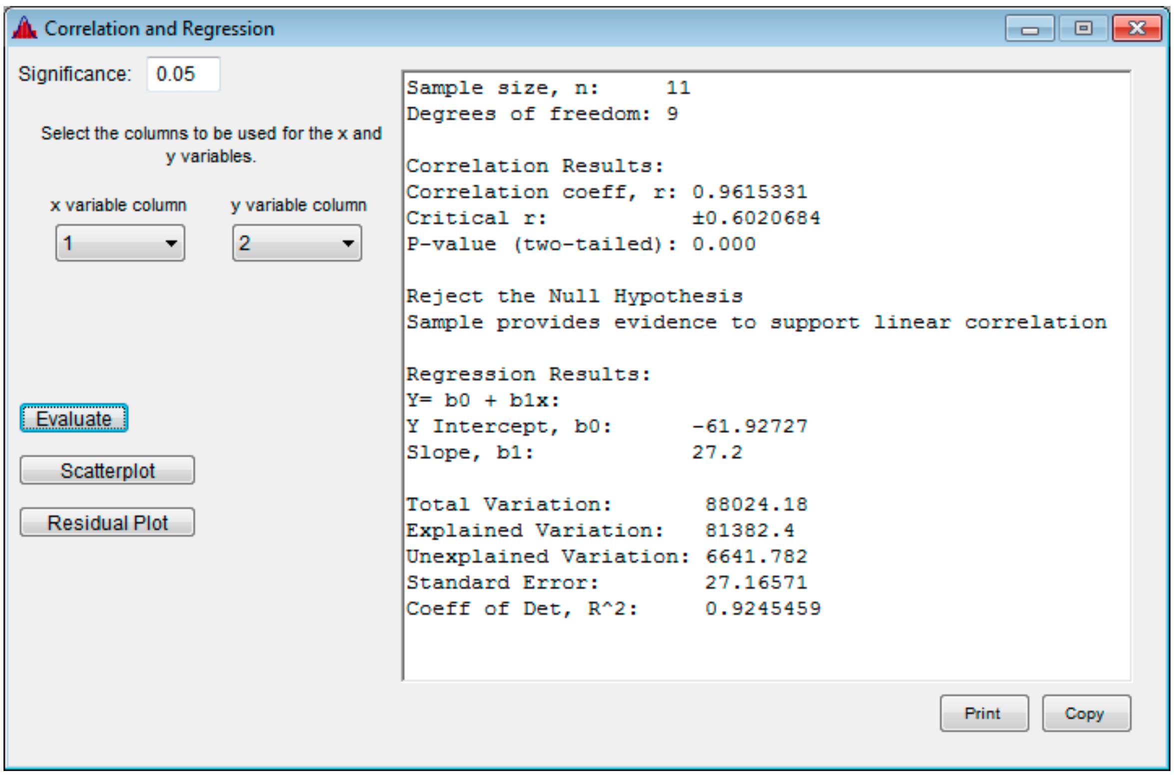
The linear model can be obtained by using Statdisk's correlation and regression module.

1. For the data in Table 10-7, enter the coded year values of 1, 2, 3, . . . , 11 in the first column of the Statdisk Sample Editor and enter the population values of 5, 10, 17, . . . , 281 in the second column.



Row	1	2
1	1	5
2	2	10
3	3	17
4	4	31
5	5	50

2. Use the procedure described in Section 10-2 of this workbook. Select **Analysis**, then **Correlation and Regression**. The result will be as shown below.



The above Statdisk display describes key results for the linear model. The resulting function is

$$y = -61.92727 + 27.2x$$

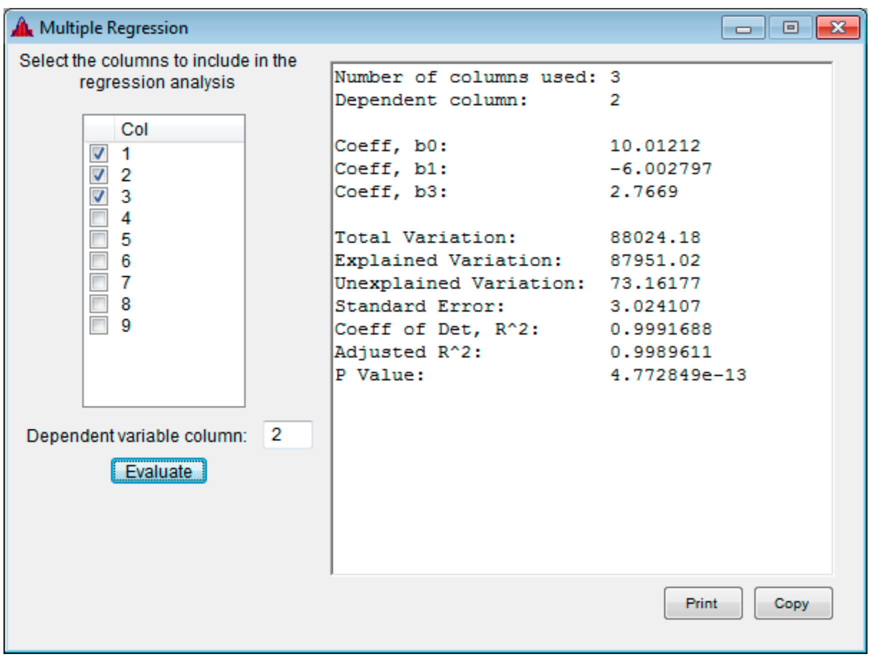
The coefficient of determination is displayed as $r^2 = 0.9245459$. The high value of r^2 suggests that the linear model is a reasonably good fit.

Quadratic Model: $y = ax^2 + bx + c$

The linear model can be obtained by using Statdisk's multiple regression module.

1. For the data in Table 10-7, enter the coded year values (*x*) of 1, 2, 3, . . . , 11 in the first column of the Statdisk Sample Editor and enter the population values (*y*) of 5, 10, 17, . . , 281 in the second column. In the third column, enter the x^2 values of 1, 4, 9,…121.
    - The Advanced Transformation formula of **Col1^2** can be used to generate the x^2 values as explained Section 1-6 of this manual (click **Data** on the top menu, then select **Sample Transformations**).

2. Select **Analysis**, then **Multiple Regression.**
    - When indicating the columns to be used, select columns 1, 2, and 3, and be sure to identify the column containing the *y* values (column 2) as the column for the dependent variable.



The display shown above corresponds to the quadratic model used with the sample data in Table 10-7. Note that there are three columns of data representing *x*, *y*, and x^2. The results show that the function has the form given as

$$y = 10.012 - 6.003x + 2.767x^2$$

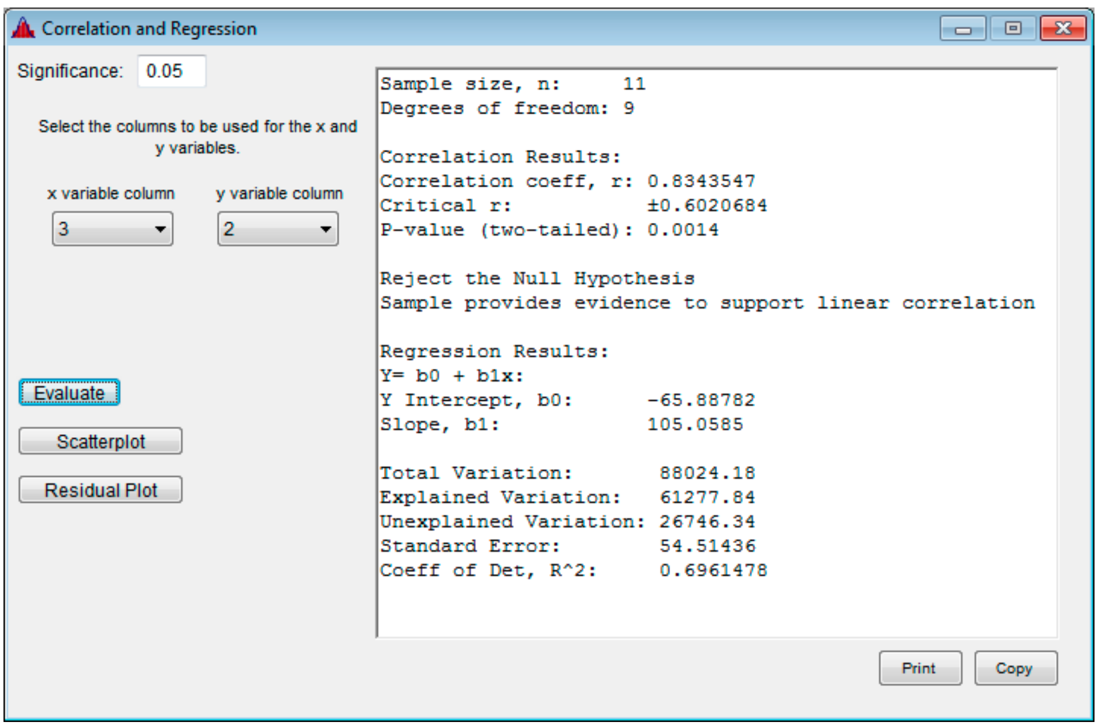
(The coefficient b_1 corresponds to *x* and b_3 corresponds to x^2.) The coefficient of determination is given by R^2 = 0.9991688, suggesting a better fit than the linear model (which has R^2 = 0.9245459).

Logarithmic Model: *y* = *a* + *b* ln *x*

The linear model can be obtained by using Statdisk's correlation and regression module.

1. For the data in Table 10-7, enter the coded year values (*x*) of 1, 2, 3, . . . , 11 in the first column of the Statdisk Sample Editor and enter the population values (*y*) of 5, 10, 17, . . , 281 in the second column. In the third column, enter the ln *x* values of 0, 0.6931472, 1.098612,..,2.397895.
   - The Advanced Transformation formula of **log(Col1)** can be used to generate the ln *x* values as explained Section 1-6 of this manual (click **Data** on the top menu, then select **Sample Transformations**).

2. Select **Analysis**, then **Correlation and Regression** to obtain the results shown below.
   - Select ln *x* (column 3) as the x variable and *y* (column 2) as the *y* variable.



The display shown above results from the logarithmic model used with the sample data in Table 10-7. The function is given by

$$y = -65.9 + 105 \ln x$$

Also, R^2 = 0.6961478, suggesting that this model does not fit as well as the linear or quadratic models. Of the three models considered so far, the quadratic model appears to be best (because it has the highest value of R^2).

Statdisk

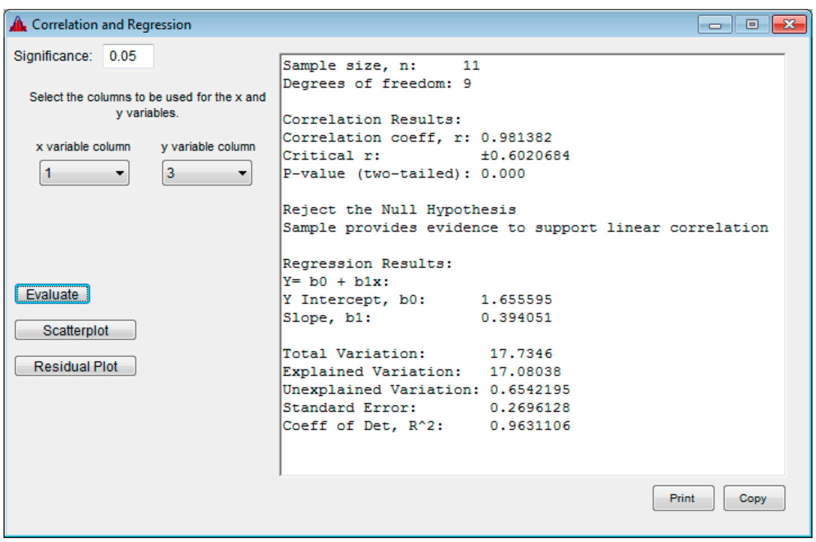
Exponential Model: $y = ab^x$

The exponential model is tricky, but it can be obtained using Statdisk.

1. For the data in Table 10-7, enter the coded year values (*x*) of 1, 2, 3, . . . , 11 in the first column of the Statdisk Sample Editor and enter the population values (*y*) of 5, 10, 17, . . , 281 in the second column. In the third column, enter the ln *y* values of 1.609438, 2.302585, 2.833213,..,5.638355.
   - The Advanced Transformation formula of **log(Col2)** can be used to generate the ln *y* values as explained Section 1-6 of this manual (click **Data** on the top menu, then select **Sample Transformations**).

2. Select **Analysis**, then **Correlation and Regression** to obtain the results shown below.
   - Select *x* (column 1) as the x variable and ln *y* (column 3) as the *y* variable.

3. The value of the coefficient of determination provided by Statdisk is correct, but the values of *a* and *b* in the exponential model must be computed as follows:
   - To find the value of *a*: Evaluate e^{b0} where b_0 is given by Statdisk.
   - To find the value of *b*: Evaluate e^{b1} where b_1 is given by Statdisk.



The value of R^2 = 0.9631106 is OK as is, but the values of *a* and *b* must be computed from the Statdisk results as shown below:

$$a = e^{b0} = e^{1.655595} = 5.2362$$
$$b = e^{b1} = e^{0.3944051} = 1.4830$$

Using these values of *a* and *b*, we express the exponential model as

$$\mathbf{y = 5.2362(1.4830^x)}$$

Power Model: $y = ax^b$

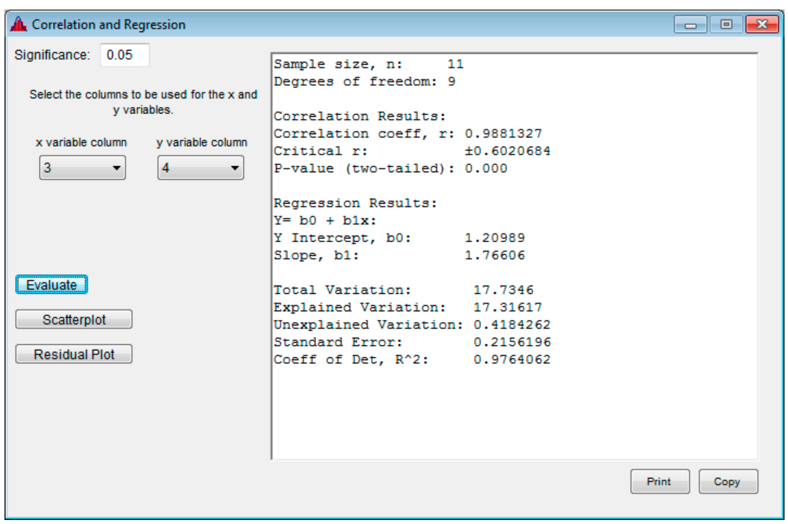
The power model is also tricky, but it too can be obtained using Statdisk.

1. For the data in Table 10-7, enter the coded year values (*x*) of 1, 2, 3, . . . , 11 in the first column of the Statdisk Sample Editor and enter the population values (*y*) of 5, 10, 17, . . , 281 in the second column. In the third column, enter the ln *x* values of 1.609438, 2.302585, 2.833213,..,5.638355. In the fourth column enter the ln *y* values.
   - The Advanced Transformation formula of **log(Col1)** can be used to generate the ln *x* values and the formula **log(Col2)** can be used to generate the ln *y* values as explained Section 1-6 of this manual (click **Data** on the top menu, then select **Sample Transformations**).

.

2. Select **Analysis**, then **Correlation and Regression** to obtain the results shown below.
   - Select ln *x* (column 3) as the x variable and ln *y* (column 4) as the *y* variable.

3. The value of the coefficient of determination provided by Statdisk is correct, but the values of *a* and *b* in the power model are found as follows:
   - To find the value of *a*: Evaluate e^{b0} where b_0 is given by Statdisk.
   - The value of *b* is the same as the value of b_1 given by Statdisk.



The value of R^2 = 0.9764062 is OK as is, and it suggests that the power model is not as good as the quadratic model. The values of *a* and *b* are found from the Statdisk results as shown below:

$$a = e^{b0} = e^{1.20989} = 3.3531$$
$$b = b_1 \text{ from Statdisk} = 1.76606$$

Using these values of *a* and *b*, we express the power model as

$$\mathbf{y = 3.3531(x^{1.76606})}$$

The rationale underlying the methods for the exponential and power models is based on transformations of equations. In the exponential model of $y = ab^x$, for example, taking natural logarithms of both sides yields $\ln y = \ln a + x (\ln b)$, which is the equation of a straight line. Statdisk can be used to find the equation of this straight line that fits the data best; the intercept will be ln *a* and the slope will be ln *b*, but we need the values of *a* and *b*, so we solve for them as described above. Similar reasoning is used with the power model.

CHAPTER 10 EXPERIMENTS: Correlation and Regression

10-1 **Bear Weights and Ages** Refer to Appendix B in the Triola textbook for data set *7 – Bear Measurements*. Use the values for WEIGHT (x) and the values for AGE (y) to find the following.

a. Display the scatter diagram of the paired WEIGHT/AGE data. Based on that scatter diagram, does there appear to be a relationship between the weights of bears and their ages? If so, what is it?

____

b. Find the value of the linear correlation coefficient r. ______________

c. Assuming a 0.05 level of significance, what do you conclude about the correlation between weights and ages of bears?

____

d. Find the equation of the regression line. (Use WEIGHT as the x predictor variable, and use AGE as the y response variable.) ______________________________________

e. What is the best predicted age of a bear that weighs 300 lb?____________________

10-2 **Effect of Transforming Data** The ages used in Experiment 10–1 are in months. Convert them to days by multiplying each age by 30 and find the following:

a. Display the scatter diagram of the paired WEIGHT/AGE data. Based on that scatter diagram, does there appear to be a relationship between the weights of bears and their ages? If so, what is it?

____

b. Find the value of the linear correlation coefficient r.___________________________

c. Assuming a 0.05 level of significance, what do you conclude about the correlation between weights and ages of bears?

____

d. Find the equation of the regression line. (Use WEIGHT as the x predictor variable, and use AGE as the y response variable.) ______________________________________

e. What is the best predicted age of a bear that weighs 300 lb?____________________

f. After comparing the responses obtained in Experiment 10-1 to those obtained here, describe the general effect of changing the scale for one of the variables.

____

____

10-3 **Bear Weights and Chest Sizes** Refer to Appendix B in the Triola textbook for data set *7 – Bear Measurements*. Use the values for CHEST (*x*) and the values for WEIGHT (*y*) to find the following.

a. Display the scatter diagram of the paired CHEST/WEIGHT data. Based on that scatter diagram, does there appear to be a relationship between the chest sizes of bears and their weights? If so, what is it?

____

b. Find the value of the linear correlation coefficient *r*. ______________________________

c. Assuming a 0.05 level of significance, what do you conclude about the correlation between chest sizes and weights of bears?

____

d. Find the equation of the regression line. (Use CHEST as the *x* predictor variable, and use WEIGHT as the *y* response variable.) _____________________

e. What is the best predicted weight of a bear with a chest size of 36.0 in?_______

f. When trying to obtain measurements from an anesthetized bear, what is a practical advantage of being able to predict the bear's weight by using its chest size?

____

____

10-4 **Effect of No Variation for a Variable** Use the following paired data and obtain the indicated results.

x	1	2	3	4	5	7	7	9
y	5	5	5	5	5	5	5	5

a. Print a scatterplot of the paired *x* and *y* data. Based on the result, does there appear to be a relationship between *x* and *y*? If so, what is it?

____

b. What happens when you try to find the value of *r*? Why?

____

c. What do you conclude about the correlation between *x* and *y*? What is the equation of the regression line?

____

10-5 **Town Courts** Listed below are amounts of court income and salaries paid to the town justices (based on data from the *Poughkeepsie Journal*). All amounts are in thousands of dollars, and all of the towns are in Dutchess County, New York.

Court Income	65	404	1567	1131	272	252	111	154	32
Justice Salary	30	44	92	56	46	61	25	26	18

Is there a correlation between court incomes and justice salaries? Explain.

____

What is the equation of the regression line? ___________________________________

Find the best predicted justice salary for a court with income of $83,941 (or $83.941 thousand). ________________________

10-6 **Car Repair Costs** **L**isted below are repair costs (in dollars) for cars crashed at 6 mi/h in full-front crash tests and the same cars crashed at 6 mi/h in full-rear crash tests (based on data from the Insurance Institute for Highway Safety). The cars are the Toyota Camry, Mazda 6, Volvo S40, Saturn Aura, Subaru Legacy, Hyundai Sonata, and Honda Accord.

Front	936	978	2252	1032	3911	4312	3469
Rear	1480	1202	802	3191	1122	739	2767

Is there a correlation between front and rear repair costs? Explain.

____

What is the equation of the regression line? ___________________________________

Find the best predicted repair cost for a full-rear crash for a Volkwagon Passat, given that its repair cost from a full-front crash is $4594. _______________

Statdisk

10-7 **Movie Data** Refer to the data from movies in the table below, where all amounts are in millions of dollars. Let the movie gross amount be the dependent *y* variable.

Budget	62	90	50	35	200	100	90
Gross	65	64	48	57	601	146	47

Is there a correlation between the budget amount and the gross amount? Explain.

____

What is the equation of the regression line? ______________________________

Find the best predicted gross amount for a movie with a budget of $150 million.__________

10-8 **Crickets and Temperature** The numbers of chirps in one minute were recorded for different crickets, and the corresponding temperatures were also recorded. The results are given in the table below.

Chirps in one minute	882	1188	1104	864	1200	1032	960	900
Temperature (^{o}F)	69.7	93.3	84.3	76.3	88.6	82.6	71.6	79.6

Is there a correlation between the number of chirps and the temperature? Explain.

____

What is the equation of the regression line? ______________________________

Find the best predicted temperature when a cricket chirps 1000 times in one minute._____

10-9 **Blood Pressure** Refer to the systolic and diastolic blood pressure measurements from randomly selected subjects. Let the dependent variable *y* represent diastolic blood pressure.

Systolic	138	130	135	140	120	125	120	130	130	144	143	140	130	150
Diastolic	82	91	100	100	80	90	80	80	80	98	105	85	70	100

Is there a correlation between systolic and diastolic blood pressure? Explain.

____

What is the equation of the regression line? ______________________________

Find the best predicted measurement of diastolic blood pressure for a person with a systolic blood pressure of 123. __________________

Statdisk

10-10 **Garbage Data for Predicting Household Size** Statdisk *12th Edition 23 - Weights of Discarded Garbage for one Week* consists of data from the Garbage Project at the University of Arizona. Use household size (HHSize) as the response *y* variable. For each given predictor *x* variable, find the value of the linear correlation coefficient, the equation of the regression line, and the value of the coefficient of determination r^2. Enter the results in the spaces below.

	r	**Equation of regression line**	r^2
Metal	______	__________________________	______
Paper	______	__________________________	______
Plastic	______	__________________________	______
Glass	______	__________________________	______
Food	______	__________________________	______
Yard	______	__________________________	______
Text	______	__________________________	______
Other	______	__________________________	______
Total	______	__________________________	______

Based on the above results, which single independent variable appears to be the best predictor of household size? Why?

____

10-11 **Multiple Regression** Use the same data set described in Experiment 10-10. Let household size (HHSize) be the dependent *y* variable and use the given predictor *x* variables to enter the results below.

	Multiple regression equation	R^2	**Adj.** R^2
Metal and Paper	____________________________	_____	_______
Plastic and Food	____________________________	_____	_______
Metal, Paper, Glass	____________________________	_____	_______
Metal, Paper, Plastic, Glass	____________________________	_____	_______

Based on the above results, which of the multiple regression equations appears to best fit the data? Why?

____

Statdisk

10-12 **Statdisk Data Set: Predicting Nicotine in Cigarettes** Refer to Appendix B in the Triola textbook for data set *10 – Cigarette Tar, Nicotine and Carbon Monoxide* and use the tar, nicotine, and CO amounts for the cigarettes that are 100 mm long, filtered, non-menthol, and non-light (columns FLTar, FLNic, FLCO). Find the best regression equation for predicting the amount of nicotine in a cigarette. Is the best regression equation a good regression equation for predicting the nicotine content?

____

____

____

10-13 **Statdisk Data Set: Predicting IQ Score** Refer to Appendix B in the Triola textbook for data set *6 – IQ and Brain Size* and find the best regression equation with IQ score as the response variable. Use predictor variables of brain volume and/or weight. Why is this equation best? Based on these results, can we predict someone's IQ score if we know the volume and weight of their brain? Based on these results, does it appear that people with larger brains have higher IQ scores?

____

____

____

10-14 **Statdisk Data Set: Full IQ Score** Refer to Appendix B in the Triola textbook for data set *5 – IQ and Lead Exposure* and find the best regression equation with IQF (full IQ score) as the response variable. Use predictor variables of IQV (verbal IQ score) and IQP (performance IQ score). Why is this equation best? Based on these results, can we predict someone's full IQ score if we know their verbal IQ score and their performance IQ score? Is such a prediction likely to be very accurate?

____

____

____

10-15 **Manatee Deaths from Boats** Listed below are the numbers of Florida manatee deaths related to encounters with watercraft (based on data from *The New York Times*). The data are listed in order, beginning with the year 1980 and ending with the year 2000.

16 24 20 15 34 33 33 39 43 50 47 53 38 35 49 42 60 54 67 82 78

Use the given data to find equations and coefficients of determination for the indicated models.

	Equation	R^2
Linear	___________	___________
Quadratic	___________	___________
Logarithmic	___________	___________
Exponential	___________	___________
Power	___________	___________

Based on the above results, which model appears to best fit the data? Why?

____

____

What is the best predicted value for 2001? In 2001, there were 82 watercraft–related manatee deaths. How does the predicted value compare to the actual value?

____

____

Statdisk

11 Goodness-of-Fit and Contingency Tables

Statdisk

11-1 Goodness-of-Fit

In Section 11-2 of the Triola textbook, we deal with frequency counts from qualitative data that have been separated into different categories. The main objective is to determine whether the distribution of the sample data agrees with or "fits" some claimed distribution.

Statdisk Procedure for Goodness-of-Fit

1. Enter the *observed* frequencies in a column of the Statdisk Sample Editor. If the expected frequencies are not all the same, you must also enter a column consisting of *one* of these lists of values:
   - *Expected frequencies*
   - *Expected proportions*

2. Select **Analysis** from the main menu.

3. Select **Goodness-of-Fit**.

4. You now are presented with the following two options:
   - Equal Expected Frequencies
   - Unequal Expected Frequencies

   If you want to test the claim that the different categories are all equally likely, select "Equal Expected Frequencies." If you want to test the claim that the different categories occur with some claimed proportions (not all equal), select the second item of "Unequal Expected Frequencies."

5. In the dialog box that now appears, enter a significance level, such as 0.05.

6. Select the column containing the observed frequencies. If Unequal Expected Frequencies was chosen, also select the column containing the expected frequencies or the expected proportions.

7. Click the **Evaluate** button.

8. Click on **Plot** to obtain a graph of the χ^2 distribution that includes the test statistic and critical value.

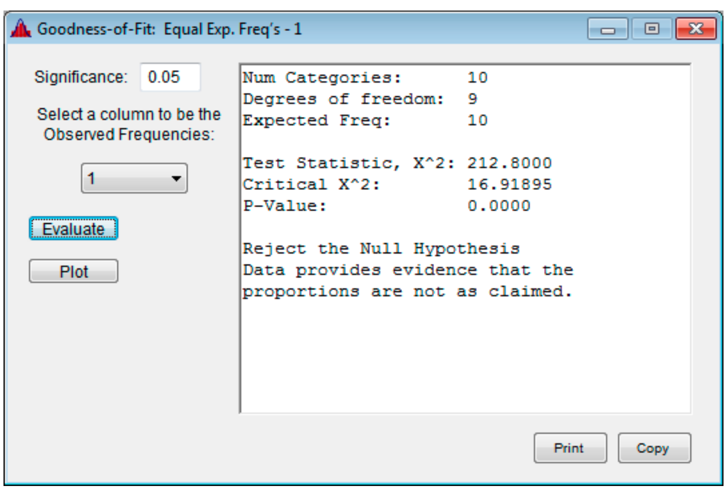
Let's consider Example 1 in Section 11-2 of the textbook.

> **EXAMPLE** A random sample of 100 weights of Californians is obtained and the frequency counts of the 100 last digits of the reported weights are listed in the table below. We want to test the claim that the observed digits are from a population of weights in which the last digits do not occur with the same frequency.

Table 11-2
Last Digits of Weights

Last Digit	Frequency
0	46
1	1
2	2
3	3
4	3
5	30
6	4
7	0
8	8
9	3

As in Section 11–2 of the textbook, we will test the claim that the sample is from a population of weights in which the last digits do *not* occur with the same frequency. Using the Statdisk procedure, we get the results shown below.



The *P*-value of 0.000 suggests that we reject the null hypothesis that the digits occur with the same frequency. There is sufficient evidence to support the claim that the digits do not occur with the same frequency.

11-2 Contingency Tables

A **contingency table** (or **two-way frequency table**) is a table in which frequencies correspond to two variables. One variable is used to categorize rows, and a second variable is used to categorize columns. Let's consider the data in the contingency table shown below (from Example 1 in Section 11-3 of the textbook).

Table 11-6 Study of Success with Different Treatments for Stress Fracture

	Success	Failure
Surgery	54	12
Weight-Bearing Cast	41	51
Non-Weight-Bearing Cast for 6 Weeks	70	3
Non-Weight-Bearing Cast Less Than 6 Weeks	17	5

Statdisk Procedure for Contingency Tables

1. Enter the observed frequencies in columns of the Statdisk Sample Editor. Enter the data in rows and columns as they appear in the contingency table.

2. Select **Analysis** from the main menu.

3. Select **Contingency Tables**.

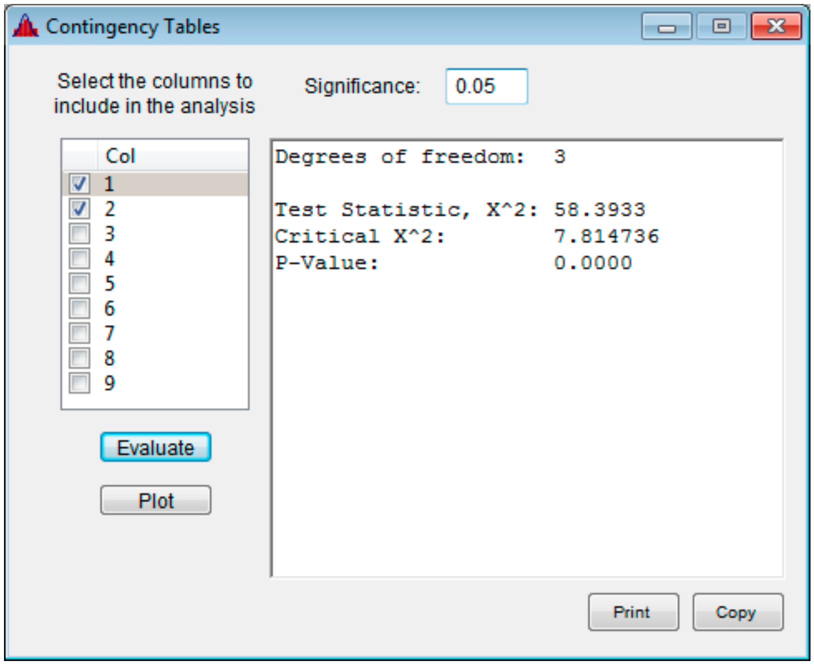
4. In the dialog box that appears, enter a significance level such as 0.05 or 0.01.

5. Select the columns that include the data from the contingency table.

6. Click on the **Evaluate** button.

7. Click on **Plot** to obtain a graph of the χ^2 distribution that includes the test statistic and critical value.

Shown below are the Statdisk results from the data in Table 11-6. The Statdisk display includes the important elements we need to make a decision. The test statistic, critical value, and *P*-value are all provided. The *P*-value of 0.000 is less than the 0.05 significance level, so we reject the null hypothesis of independence between the row and column variables. It appears that success or failure is dependent on the treatment.



11-3 Fisher's Exact Test

Fisher's exact test can be used for two-way tables, and it is used mostly for 2 x 2 tables. This test uses an *exact* distribution instead of an approximating chi-square distribution. It is particularly helpful when the approximating chi-square distribution cannot be used because of *expected* cell frequencies that are less than 5. Consider the sample data in the table below, with expected frequencies shown in parentheses. Note that the first cell has an expected frequency of 3, which is less than 5, so the chi-square distribution should not be used.

Statdisk Procedure for Fisher's Exact Test

1. Select **Analysis** from the main menu.

2. Select **Fisher Exact Test**.

3. In the dialog box that appears, enter the four frequencies in the cells of the table. (Enter a frequency, press the **Tab** key, then enter another frequency, and so on. Do not try to enter row and column totals; they will be provided by Statdisk.
   - Be careful to enter the four sample frequencies, not the *expected* frequencies)

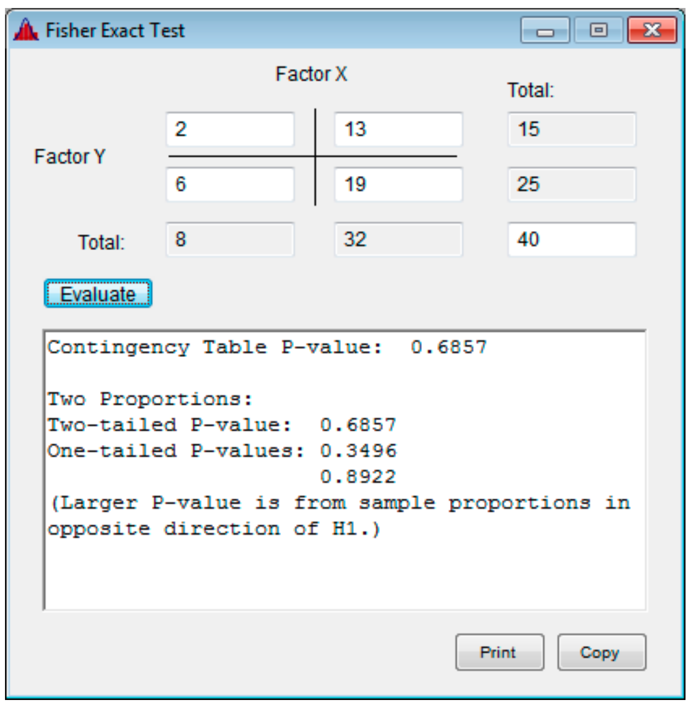
4. Click on the **Evaluate** button.

Consider the sample data in the table below. The expected frequencies are shown in parentheses, and we can see that one of them is less than 5.

Helmets and Facial Injuries in Bicycle Accidents
(Expected frequencies are in parentheses.)

	Helmet Worn	**No Helmet**
Facial injuries received	2 (3)	13 (12)
All injuries nonfacial	6 (5)	19 (20)

The Statdisk display is shown below. Because the *P*-value is large, we fail to reject the null hypothesis that wearing a helmet and receiving facial injuries are independent. There isn't enough evidence to suggest that facial injuries are dependent on whether a helmet was worn.



Statdisk

11-4 McNemar's Test for Matched Pairs

The contingency table procedures in Section 11-3 of the textbook are based on *independent* data. For 2 × 2 tables consisting of frequency counts that result from *matched pairs*, we do not have independence and, for such cases, we can use McNemar's test for matched pairs. We can use McNemar's test for the null hypothesis that frequencies from the discordant (different) categories occur in the same proportion. Here is the Statdisk procedure.

Statdisk Procedure for McNemar's Test

1. Select **Analysis** from the main menu.

2. Select **McNemar's Test**.

3. In the dialog box that appears, enter the four frequencies in the cells of the table. (Enter a frequency, press the **Tab** key, then enter another frequency, and so on. Do not try to enter row and column totals; they will be provided by Statdisk.)

4. Enter a significance level, such as 0.05.

5. Click the **Evaluate** button.

Consider Example 6 form Section 11-3 of the textbook.

EXAMPLE ***Are Hip Protectors Effective?*** A randomized controlled trial was designed to test the effectiveness of hip protectors in preventing hip fractures in the elderly. Nursing home residents each wore protection on one hip, but not the other. Results are summarized in Table 11-8 (based on data from "Efficacy of Hip Protector to Prevent Hip Fracture in Nursing Home Residents," by Kiel et al, *Journal of the American Medical Association*, Vol. 298, No. 4). Using a 0.05 significance level, apply McNemar's test to test the null hypothesis that the following two proportions are the same:

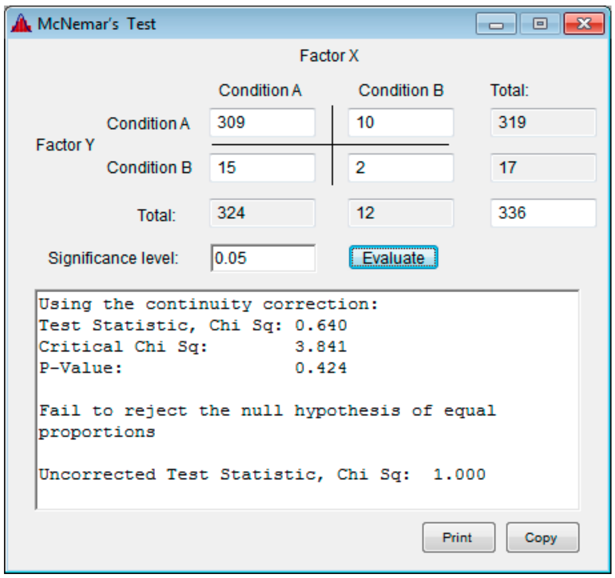
- The proportion of subjects with no hip fracture on the protected hip and a hip fracture on the unprotected hip.
- The proportion of subjects with a hip fracture on the protected hip and no hip fracture on the unprotected hip.

Statdisk

Table 11-8 Randomized Controlled Trial of Hip Protectors

		No Hip Protector Worn	
		No Hip Fracture	**Hip Fracture**
Hip Protector Worn	**No Hip Fracture**	309	10
	Hip Fracture	15	2

Using the Statdisk procedure and data from Table 11-8 the following results are obtained.



The Statdisk results include the chi-square test statistic, the critical value, and the *P*-value. Because the *P*-value of 0.424 is greater than the significance level of 0.05, we fail to reject the null hypothesis of equal proportions. It appears that the proportion of hip fractures with the protectors worn is not significantly different from the proportion of hip fractures without the protectors worn. The hip protectors do not appear to be effective in preventing hip fractures.

Statdisk

CHAPTER 11 EXPERIMENTS: Goodness-of-Fit and Contingency Tables

11-1 **Loaded Die** The author drilled a hole in a die and filled it with a lead weight, then proceeded to roll it 200 times. Here are the observed frequencies for the outcomes of 1, 2, 3, 4, 5, and 6 respectively: 27, 31, 42, 40, 28, 32. Use a 0.05 significance level to test the claim that the outcomes are not equally likely.

Test statistic:____________ Critical value:____________ *P*-value:_________________

Conclusion:____

____

Does it appear that the loaded die behaves differently than a fair die?

____

11-2 **Flat Tire and Missed Class** A classic tale involves four car-pooling students who missed a test and gave as an excuse a flat tire. On the makeup test, the instructor asked the students to identify the particular tire that went flat. If they really didn't have a flat tire, would they be able to identify the same tire? The author asked 41 other students to identify the tire they would select. The results are listed in the following table (except for one student who selected the spare). Use a 0.05 significance level to test the author's claim that the results fit a uniform distribution.

Tire	Left front	Right front	Left rear	Right rear
Number selected	11	15	8	6

Test statistic:____________ Critical value:____________ *P*-value:_________________

Conclusion:____

____

What does the result suggest about the ability of the four students to select the same tire when they really didn't have a flat?

____

11-3 **Births** Records of randomly selected births were obtained and categorized according to the day of the week that they occurred (based on data from the National Center for Health Statistics). Because babies are unfamiliar with our schedule of weekdays, a reasonable claim is that births occur on the different days with equal frequency. Use a 0.01 significance level to test that claim.

Day	Sun	Mon	Tues	Weds	Thurs	Fri	Sat
Birth	77	110	124	122	120	123	97

Test statistic:_____________ Critical value:_____________ *P*-value:__________________

Conclusion:____

____

11-4 **NYC Homicides** For a recent year, the following are the numbers of homicides that occurred each month in New York City: 38, 30, 46, 40, 46, 49, 47, 50, 50, 42, 37, 37. Use a 0.05 significance level to test the claim that homicides in New York City are equally likely for each of the 12 months.

Test statistic:_____________ Critical value:_____________ *P*-value:__________________

Conclusion:____

____

11-5 **Measuring Pulse Rates** According to one procedure used for analyzing data, when certain quantities are measured, the last digits tend to be uniformly distributed, but if they are estimated or reported, the last digits tend to have disproportionately more 0s or 5s. Refer to Statdisk *12th Edition* data set *1 – Body Measurements Male/Female* and use the last digits of the pulse rates of the 80 men and women. Those pulse rates were obtained as part of the National Health and Examination Survey. Test the claim that the last digits of 0, 1, 2, . . . , 9 occur with the same frequency.

Test statistic:_____________ Critical value:_____________ *P*-value:__________________

Conclusion:____

____

Based on the observed digits, what can be inferred about the procedure used to obtain the pulse rates?

____

11-6 **Testing a Normal Distribution** In this experiment we will use Statdisk's ability to generate normally distributed random numbers. We will then test the sample data to determine if they actually do fit a normal distribution.

a. Generate 1000 random numbers from a normal distribution with a mean of 100 and a standard deviation of 15. (IQ scores have these parameters.) Use the **Normal Generator** feature of Statdisk.

b. Arrange the generated data in order. (Use **Copy/Paste** to copy the data to the Statdisk data window, where the **Data tools** button gives you the option of sorting a column of data.)

c. Examine the sorted list and determine the frequency for each of the categories listed below. Enter those frequencies in the spaces provided. (The expected frequencies were found using the methods of Chapter 6 in the textbook.)

	Observed Frequency	Expected Frequency
Below 55:	____________	1
55-70:	____________	22
70-85:	____________	136
85-100:	____________	341
100-115:	____________	341
15-130:	____________	136
130-145:	____________	22
Above 145:	____________	1

d. Use Statdisk to test the claim that the randomly generated numbers actually do fit a normal distribution with mean 100 and standard deviation 15.

Test statistic:__________ Critical value:___________ *P*-value:________________

Conclusion:____

____

11-7 **Instant Replay in Tennis** The table below summarizes challenges made by tennis players in the first U. S. Open that used the Hawk-Eye electronic instant replay system. Use a 0.05 significance level to test the claim that success in challenges is independent of the gender of the player. Does either gender appear to be more successful?

	Was the challenge to the call successful? Yes	No
Men	201	288
Women	126	224

Test statistic:____________ Critical value:____________ *P*-value:_________________

Conclusion:____

____

11-8 **Accuracy of Polygraph Tests** The data in the accompanying table summarize results from tests of the accuracy of polygraphs (based on data from the Office of Technology Assessment). Use a 0.05 significance level to test the claim that whether the subject lies is independent of the polygraph indication.

	Polygraph Indicated Truth	Polygraph Indicated Lie
Subject actually told the truth	65	15
Subject actually told a lie	3	17

Test statistic:____________ Critical value:____________ *P*-value:_________________

Conclusion:____

____

11-9 **Global Warming Survey** A Pew Research poll was conducted to investigate opinions about global warming. The respondents who answered yes when asked if there is solid evidence that the earth is getting warmer were then asked to select a cause of global warming. The results are given the table below. Use a 0.05 significance level to test the claim that the sex of the respondent is independent of the choice for the cause of global warming. Do men and women appear to agree, or is there a substantial difference?

	Human activity	Natural patterns	Don't know or refused to answer
Male	314	146	44
Female	308	162	46

Test statistic:_____________ Critical value:_____________ *P*-value:__________________

Conclusion:____

____

11-10 **Occupational Hazards** Use the data in the table to test the claim that occupation is independent of whether the cause of death was homicide. The table is based on data from the U.S. Department of Labor, Bureau of Labor Statistics.

	Police	Cashiers	Taxi Drivers	Guards
Homicide	82	107	70	59
Cause of Death Other than Homicide	92	9	29	42

Test statistic:_____________ Critical value:_____________ *P*-value:__________________

Conclusion:____

____

Does any particular occupation appear to be most prone to homicides? If so, which one?

____

How are the results affected if the order of the rows is switched?

____

How are the results affected by the presence of an outlier? If we change the first entry from 82 to 8200, are the results dramatically affected?

____

11-11 **Fisher's Exact Test** Refer to Experiment 11-8 in this manual/workbook. Repeat that experiment by using Fisher's exact test instead of using the approximating chi-square distribution. Enter the results below.

P-value obtained by using the approximating chi-square distribution: __________

P-value obtained by using Fisher's exact test: __________

Does the use of the Fisher's exact test have much of an effect on the *P*-value?

____

____

11-12 **Fisher's Exact Test** The U. S. Supreme Court considered a case involving the exam for firefighter lieutenant in the city of New Haven, CT. Results from the exam are shown in the table below. Is there sufficient evidence to support the claim that results from the test should be thrown out because they are discriminatory? Use a 0.01 significance level.

	Passed	Failed
White Candidates	17	16
Minority Candidates	9	25

P-value obtained by using the approximating chi-square distribution: __________

Conclusion (based on results from using the approximating chi-square distribution):

____

P-value obtained by using Fisher's exact test: __________

Conclusion (based on results from using Fisher's exact test):

____

Does the use of the Fisher's exact test have much of an effect on the *P*-value?

____

____

Statdisk

Exercises 11-13 through 11-16 require McNemar's test discussed in Section 11-3 of Elementary Statistics.

11-13 **Treating Athlete's Foot** Assume that subjects are inflicted with athlete's foot on each of their feet. Also assume that for each subject, one foot is treated with a fungicide solution while the other foot is given a placebo. The results are given in the accompanying table. Using a 0.05 significance level, test the effectiveness of the treatment.

		Fungicide Treatment	
		Cure	No Cure
Placebo	Cure	5	12
	No Cure	22	55

Test statistic:____________ Critical value:____________ *P*-value:_________________

Conclusion:____

____

11-14 **Treating Athlete's Foot** Repeat Experiment 11-13 after changing the frequency of 22 to 66.

Test statistic:____________ Critical value:____________ *P*-value:_________________

Conclusion:____

____

11-15 **PET/CT Compared to MRI** In the article "Whole-Body Dual-Modality PET/CT and Whole Body MRI for Tumor Staging in Oncology" (by Antoch et al, *Journal of the American Medical Association*, Vol. 290, No. 24), the authors cite the importance of accurately identifying the stage of a tumor. Accurate staging is critical for determining appropriate therapy. The article discusses a study involving the accuracy of positron emission tomography (PET) and computed tomography (CT) compared to magnetic resonance imaging (MRI). Using the data in the table for 50 tumors analyzed with both technologies, does there appear to be a difference in accuracy? Does either technology appear to be better? Enter the results on the following page.

		PET/CT	
		Correct	Incorrect
MRI	Correct	36	1
	Incorrect	11	2

Statdisk

Test statistic:____________ Critical value:____________ *P*-value:_________________

Conclusion:____

____

11-16 **Testing a Treatment** In the article "Eradication of Small Intestinal Bacterial Overgrowth Reduces Symptoms of Irritable Bowel Syndrome" (by Pimentel, Chow, Lin, *American Journal of Gastroenterology*, Vol. 95, No. 12), the authors include a discussion of whether antibiotic treatment of bacteria overgrowth reduces intestinal complaints. McNemar's test was used to analyze results for those subjects with eradication of bacterial overgrowth. Using the data in the given table, does the treatment appear to be effective against abdominal pain?

		Abdominal pain before treatment?	
		Yes	No
Abdominal pain after treatment?	Yes	11	1
	No	14	3

Test statistic:____________ Critical value:____________ *P*-value:_________________

Conclusion:____

____

12 Analysis of Variance

Statdisk

12-1 One-Way Analysis of Variance

One–way analysis of variance is used to test the claim that three or more populations have the same mean. When the Triola textbook discusses one-way analysis of variance, it is noted that the term "one-way" is used because the sample data are separated into groups according to one characteristic or "factor". For example, in data set *20 - M&M Plain Candy Weights* (grams) shown below, there is one factor used to categorize the data: the candy color.

Red	Orange	Yellow	Brown	Blue	Green
0.751	0.735	0.883	0.696	0.881	0.925
0.841	0.895	0.769	0.876	0.863	0.914
0.856	0.865	0.859	0.855	0.775	0.881
0.799	0.864	0.784	0.806	0.854	0.865
0.966	0.852	0.824	0.840	0.810	0.865
0.859	0.866	0.858	0.868	0.858	1.015
0.857	0.859	0.848	0.859	0.818	0.876
0.942	0.838	0.851	0.982	0.868	0.809

Because the calculations are very complicated, the Triola textbook emphasizes the interpretation of results obtained by using software, so Statdisk is very suitable for this topic. We should understand that a small *P*-value (such as 0.05 or less) leads to rejection of the null hypothesis of equal means. [“If the *P* (value) is low, the null must go.”] With a large *P*-value (such as greater than 0.05), fail to reject the null hypothesis of equal means.

For the data in the above table, the claim of equal means leads to these hypotheses:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

H_1: At least one of the population means is different from the others.

Statdisk Procedure for One-Way Analysis of Variance

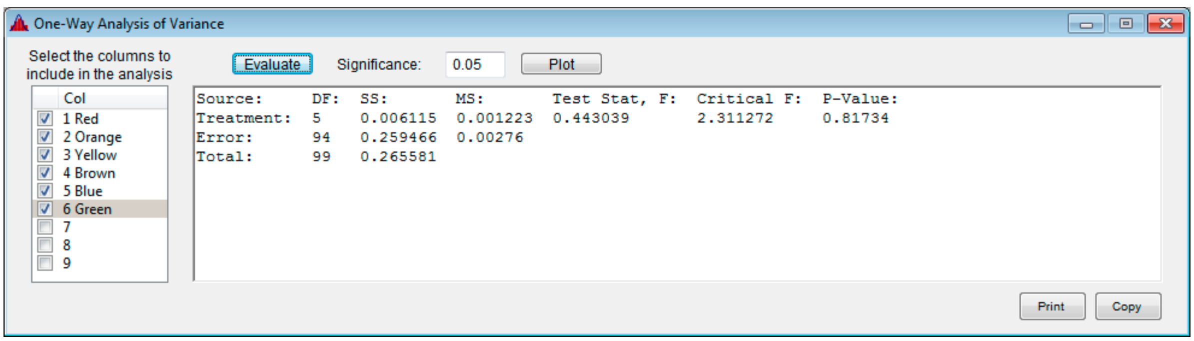
1. Enter the data in separate columns of the Statdisk data window.
2. Select **Analysis** from the top menu bar.
3. Select **One-Way Analysis of Variance** from the submenu.
4. In the dialog box, enter a significance level, such as 0.05 or 0.01.

5. Select the columns to be used for the analysis of variance. If a box already has a check mark and you do not want to include it, click on the box to remove the check mark. If a box does not have a check mark and you want to include it, click on the box to make the check mark appear.

6. Click on the **Evaluate** button.

7. Click on **Plot** to obtain a graph that includes the critical value and test statistic.

The data in data set *20 - M&M Plain Candy Weights* is included in Statdisk and can be opened by selecting ***Data Sets*** in the top menu, selecting ***Elementary Statistics 12th Edition*** and then ***20 - M and M Plain Candy Weights***. If you use the above steps with this data set and include all colors in the analysis, the Statdisk result will appear as follows.



The *P*-value of 0.81734 indicates that there is not sufficient evidence to warrant rejection of the null hypothesis that the means are equal. The test statistic of F = 0.443039 is also provided along with the critical value of F = 2.311272. The values of the SS and MS components are also provided. If you click on the **Plot** button, you will get a graph showing the test statistic and critical value.

Caution: It is easy to feed Statdisk (or any other software package) data that can be processed quickly and painlessly, but we should *think* about what we are doing. We should consider the assumptions for the test being used, and we should *explore* the data before jumping into a formal procedure such as analysis of variance. Carefully explore the important characteristics of data, including the center (through means and medians), variation (through standard deviations and ranges), distribution (through histograms, boxplots, and normal quantile plots), outliers, and any changing patterns over time.

12-2 Two-Way Analysis of Variance

Two-way analysis of variance involves *two* factors, such as type of gender (male or female) and blood lead level (low, medium high) shown in Table 12-3 form the textbook. The two–way analysis of variance procedure requires that we test for (1) an interaction effect between the two factors; (2) an effect from the row factor; (3) an effect from the column factor.

Table 12-3 Measures of Performance IQ

	Blood Lead Level		
	Low	**Medium**	**High**
Male	85	78	93
	90	107	97
	107	90	79
	85	83	97
	100	101	111
Female	64	97	100
	111	80	71
	76	108	99
	136	110	85
	99	97	93

Statdisk Procedure for Two-Way Analysis of Variance

1. Select **Analysis** from the top menu bar.

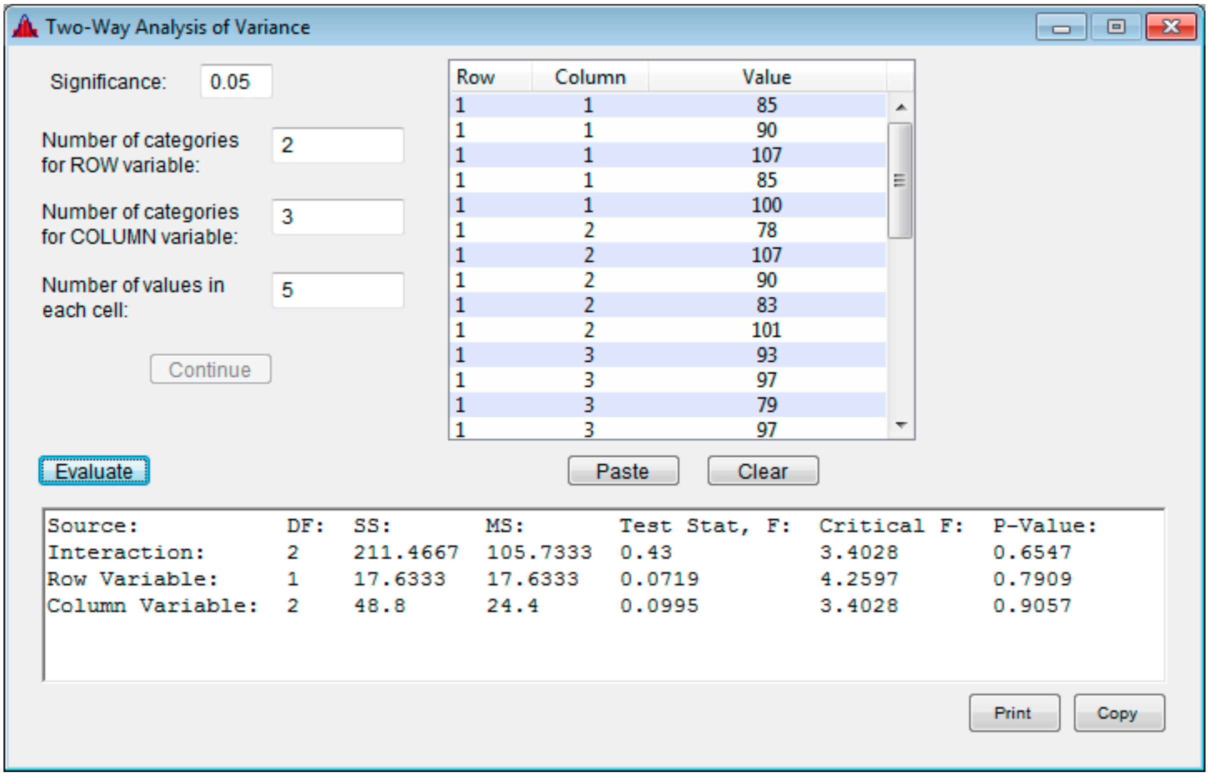
2. Select **Two-Way Analysis of Variance** from the submenu.

3. In the dialog box, enter the significance level, such as 0.05 or 0.01.

4. In the dialog box, enter the number of categories for the row variable, enter the number of categories for the column variable, and enter the number of values in each cell.
   - For the sample data in Table 12-3, enter 2 for the number of categories for the row variable (Male, Female), enter 3 for the number of categories of the column variable (Low, Medium, High), and enter 5 for the number of values in each cell.
   - Click on **Continue** when finished.

5. Statdisk will automatically generate a format for entering the sample data. You will be given row and column numbers, so enter the sample values according to their locations See the Statdisk display below to see how the sample values from Table 12-3 are entered.

6. Click **Evaluate** after all sample values have been entered.



See Section 12–3 in the Triola textbook for the basic procedure for two–way analysis of variance, and note that it involves three distinct components:

1. Test for Interaction

In the results included in the Statdisk display, see that the interaction is associated with a *P*-value of 0.6547, so we fail to reject the null hypothesis of no interaction between the two factors. It does not appear that the performance IQ scores are affected by an interaction between sex (male, female) and blood lead level (low, medium, high). There does not appear to be an interaction effect.

2. Test for Effect from the Row Factor

Our two-way analysis of variance procedure outlined in the textbook indicates that we should now proceed to test the null hypothesis that there are no effects from the row factor (Sex). The corresponding *P*-value is shown in the Statdisk display as 0.7909. Because that *P*-value is greater than the significance level of 0.05, we fail to reject the null hypothesis of no effects from sex. That is, performance IQ scores do not appear to be affected by the sex of the subject.

3. Test for Effect from the Column Factor

Our two-way analysis of variance procedure outlined in the textbook indicates that we should now proceed to test the null hypothesis that there are no effects from the column factor (Blood Lead Level). The corresponding *P*-value is shown in the Minitab display as 0.9057. Because that

P-value is greater than the significance level of 0.05, we fail to reject the null hypothesis of no effects from lead level. Performance IQ scores do not appear to be affected by whether the lead exposure is low, medium, or high.

Special Case: One Observation per Cell and No Interaction

The Triola textbook (excluding *Essentials of Statistics*) includes a subsection describing the special case in which there is only one sample value in each cell. Here's how we proceed when there is one observation per cell: *If it seems reasonable to assume (based on knowledge about the circumstances) that there is no interaction between the two factors, make that assumption and then proceed as before to test the following two hypotheses separately:*

H_0: There are no effects from the row factor.

H_0: There are no effects from the column factor.

Statdisk does work with this special case of only one observation per cell. Simply follow the same procedure given earlier in this section.

CHAPTER 12 EXPERIMENTS: Analysis of Variance

12-1 **Clancy, Rowling, Tolstoy Readability** Pages were randomly selected by the author from *The Bear and the Dragon* by Tom Clancy, *Harry Potter and the Sorcerer's Stone* by J. K. Rowling, and *War and Peace* by Leo Tolstoy. The Flesch Reading Ease scores for those pages are listed below. Use a 0.05 significance level to test the claim that the three samples are from populations with the same mean. Do the books appear to have different reading levels of difficulty?

Clancy	58.2	73.4	73.1	64.4	72.7	89.2	43.9	76.3	76.4	78.9	69.4	72.9
Rowling	85.3	84.3	79.5	82.5	80.2	84.6	79.2	70.9	78.6	86.2	74.0	83.7
Tolstoy	69.4	64.2	71.4	71.6	68.5	51.9	72.2	74.4	52.8	58.4	65.4	73.6

SS(treatment):________ MS(treatment): ________ Test statistic *F*: ____________

SS(error): ________ MS(error): ________ *P*-value: ____________

SS(total): ________

Conclusion:____

12-2 **Poplar Tree Weights** Weights (kg) of poplar trees were obtained from trees planted in a sandy and dry region. The trees were given different treatments identified in the table below. The data are from a study conducted by researchers at Pennsylvania State University, and the data were provided by Minitab, Inc. Use a 0.05 significance level to test the claim that the four treatment categories yield poplar trees with the same mean weight. Is there a treatment that appears to be most effective in the sandy and dry region?

No Treatment	Fertilizer	Irrigation	Fertilizer and Irrigation
1.21	0.94	0.07	0.85
0.57	0.87	0.66	1.78
0.56	0.46	0.10	1.47
0.13	0.58	0.82	2.25
1.30	1.03	0.94	1.64

SS(treatment):________ MS(treatment): ________ Test statistic *F*: ____________

SS(error): ________ MS(error): ________ *P*-value: ____________

SS(total): ________

Conclusion:____

12-3 **Archeology: Skull Breadths from Different Epochs** The values in the table are measured maximum breadths of male Egyptian skulls from different epochs (based on data from Ancient Races of the Thebaid, by Thomson and Randall-Maciver). Changes in head shape over time suggest that interbreeding occurred with immigrant populations. Use a 0.05 significance level to test the claim that the different epochs do not all have the same mean.

4000 b.c.	1850 b.c.	150 a.d.
131	129	128
138	134	138
125	136	136
129	137	139
132	137	141
135	129	142
132	136	137
134	138	145
138	134	137

SS(treatment):________ MS(treatment): ________ Test statistic *F*: ____________

SS(error): ________ MS(error): ________ *P*-value: ____________

SS(total): ________

Conclusion:____

12-4 **Mean Weights of M&Ms** Refer to Statdisk *12th Edition* data set *20 – M and M Plain Candy Weights*. At the 0.05 significance level, test the claim that the mean weight of M&Ms is the same for each of the six different color populations.

SS(treatment):________ MS(treatment): ________ Test statistic *F*: ____________

SS(error): ________ MS(error): ________ *P*-value: ____________

SS(total): ________

Conclusion:____

If it is the intent of Mars, Inc., to make the candies so that the different color populations have the same mean weight, do these results suggest that the company has a problem requiring corrective action?

____

12-5 **Nicotine in Cigarettes** Refer to Statdisk *12th Edition* data set *10 – Cigarette Tar, Nicotine, and Carbon Monoxide* and use the amounts of nicotine (mg per cigarette) in the king size cigarettes (KgNic), the 100 mm menthol cigarettes (MnNic), and the 100 mm non-menthol cigarettes (FLNic). The king size cigarettes are non-filtered, non-menthol, and non-light. The 100 mm menthol cigarettes are filtered and non-light. The 100 mm non-menthol cigarettes are filtered and non-light. Use a 0.05 significance level to test the claim that the three categories of cigarettes yield the same mean amount of nicotine. Given that only the king size cigarettes are not filtered, do the filters appear to make a difference?

SS(treatment):________ MS(treatment): ________ Test statistic *F*: ____________

SS(error): ________ MS(error): ________ *P*-value: ____________

SS(total): ________

Conclusion:____

____

12-6 **Tar in Cigarettes** Refer to Statdisk *12th Edition* data set *10 – Cigarette Tar, Nicotine, and Carbon Monoxide* and use the amounts of tar (mg per cigarette) in the three categories of cigarettes described in Experiment 12-5 (KgTar, MnTar, FLTar). Use a 0.05 significance level to test the claim that the three categories of cigarettes yield the same mean amount of tar. Given that only the king size cigarettes are not filtered, do the filters appear to make a difference?

SS(treatment):________ MS(treatment): ________ Test statistic *F*: ________

SS(error): ________ MS(error): ________ *P*-value: ________

SS(total): ________

Conclusion:____

Statdisk

12-7 **Simulations** Use Statdisk to randomly generate three different samples of 500 values each. (Select **Data**, then select **Normal Generator**.) For the first two samples, use a normal distribution with a mean of 100 and a standard deviation of 15. For the third sample, use a normal distribution with a mean of 105 and a standard deviation of 15. We know that the three populations have different means, but do the methods of analysis of variance allow you to conclude that the means are different? Explain.

____

____

____

12-8 ***Pulse Rate*** The following table lists pulse rates obtained from data set *1 – Body Measurements Male/Female*. Use a 0.05 significance level and apply the methods of two-way analysis of variance. What do you conclude?

	Under 30 Years of Age	Over 30 Years of Age
Female	78 104 78 64 60 98 82 98 90 96	76 76 72 66 72 78 62 72 74 56
Male	60 80 56 68 68 74 74 68 62 56	46 70 62 66 90 80 60 58 64 60

Are pulse rates affected by an interaction between gender and age? Explain.

____

Are pulse rates affected by gender? Explain.

____

Are pulse rates affected by age? Explain.

____

____

12-9 **Smoking, Body Temperature, Gender** The table below lists body temperatures obtained from randomly selected subjects (based on data set *3 – Body Temperature of Healthy Adults*). The temperatures are categorized according to gender and whether the subject smokes. Using a 0.05 significance level, test for an interaction between gender and smoking, test for an effect from gender, and test for an effect from smoking. What do you conclude?

	Smokes				Does not smoke			
Male	98.4	98.4	99.4	98.6	98.0	98.0	98.8	97.0
Female	98.8	98.0	98.7	98.4	97.7	98.0	98.2	99.1

Are body temperatures affected by an interaction between sex and smoking? Explain.

____

Are body temperatures affected by sex? Explain.

____

Are body temperatures affected by smoking? Explain.

____

____

Statdisk

13 Nonparametric Statistics

Statdisk

13-1 Nonparametric Methods

Statdisk includes a wide variety of nonparametric procedures and can perform all of the nonparametric methods described in Chapter 13 of the Triola textbook (except *Essentials of Statistics*). The sections of this chapter correspond to those in the textbook.

Section 13-1 in the textbook introduces some basic principles of nonparametric methods. The textbook notes that prior to Chapter 13, most of the methods of inferential statistics are called *parametric* methods because they are based on sampling from a population with specific parameters (such as the mean μ, standard deviation σ, or proportion p). Those parametric methods usually have some fairly strict conditions, such as a requirement that the sample data must come from a normally distributed population. Because nonparametric methods do not require specific distributions (such as the normal distribution), these nonparametric methods are often called **distribution-free tests**. The following sections describe some of the more important and commonly used nonparametric methods.

13-2 Sign Test

Section 13-2 in *Elementary Statistics* includes the following definition.

> **DEFINITION** The **sign test** is a nonparametric (distribution-free) test that uses plus and minus signs to test different claims, including:
> 1. Claims involving matched pairs of sample data
> 2. Claims involving nominal data with two categories
> 3. Claims about the median of a single population

Statdisk makes it possible to work with all three of the above cases. We first describe the Statdisk procedure, then we illustrate it with an example.

Statdisk Procedure for the Sign Test

1. Either determine the number of positive signs and the number of negative signs, or enter the paired data in columns of the Sample Editor window.

2. Select **Analysis** from the top menu bar.

3. Select **Sign Tests** from the submenu.

4. If you already know the number of positive and negative signs (as in cases involving nominal data), select the option of **Given Number of Signs**. If you have sample *paired* data, select the option of **Given Pairs of Values**.

5. The content of the dialog box will depend on the choice made in step 4. Both cases require that you select the form of the claim being tested and a significance level, such as 0.05 or 0.01. You must then enter the numbers of positive and negative signs, or the columns containing the original pairs of data.

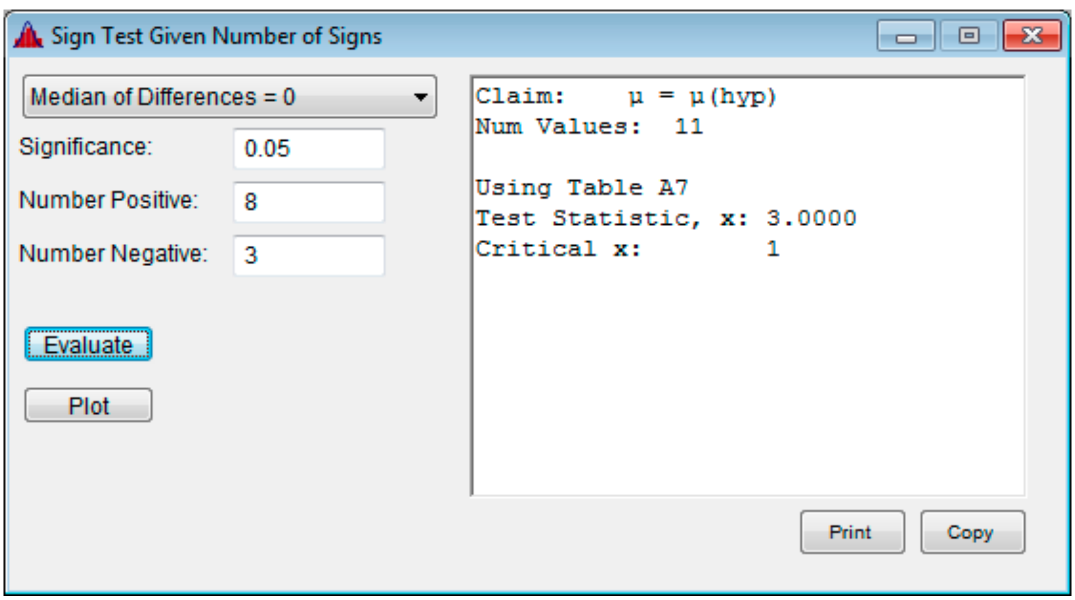
6. Click on the **Evaluate** button.
   - Click on **Plot** to obtain a graph that includes the test statistic and critical value. The plot will be generated only if the normal approximation is used (because $n > 25$).

EXAMPLE Let's consider the matched data in Table 13-3 below, from Example 2 in Section 13-2 of the textbook. (The data are matched, because each pair of values is from the same flight.)

Table 13-3 Taxi-Out Times and Taxi-In Times for American Airlines Flight 21

Taxi-out time (min)	13	20	12	17	35	19	22	43	49	45	13	23
Taxi-in time (min)	13	4	6	21	29	5	27	9	12	7	36	12
Sign of difference	0	+	+	−	+	+	−	+	+	+	−	+

From Table 13-3 we see that there are 8 positive signs and 3 negative signs. The Statdisk result is shown below. We can see from this display that the test statistic is $x = 3$ and the critical value is 1, so we fail to reject the null hypothesis of no difference. There does not appear to be significant difference between the taxi-out times and the taxi-in times.



13-3 Wilcoxon Signed-Ranks Test

The Triola textbook (excluding *Essentials of Statistics*) describes the Wilcoxon signed-ranks test, and the following definition is given.

> **DEFINITION** The **Wilcoxon signed-ranks test** is a nonparametric test that uses ranks for these applications:
> 1. Testing a claim that the population of matched pairs has the property that the matched pairs have differences with median equal to zero.
> 2. Testing a claim that a single population of individual values has median equal to some claimed value.

The textbook makes this important point: The Wilcoxon signed-ranks test and the sign test can both be used with sample data consisting of matched pairs, but the sign test uses only the *signs* of the differences and not their actual magnitudes (how large the numbers are). The Wilcoxon signed-ranks test uses *ranks*, so the magnitudes of the differences are taken into account. Because the Wilcoxon signed-ranks test incorporates and uses more information than the sign test, it tends to yield conclusions that better reflect the true nature of the data.

First we describe the Statdisk procedure for conducting a Wilcoxon signed-ranks test, then we illustrate it with an example.

Statdisk Procedure for the Wilcoxon Signed-Ranks Test

1. Enter the paired data in columns of the Statdisk Sample Editor window.

2. Select **Analysis** from the top menu bar.

3. Select **Wilcoxon Tests** from the submenu.

4. You must now choose between the following two options.
   - **Wilcoxon (Matched Pairs)**
   - **Wilcoxon (Indep. Samples)**

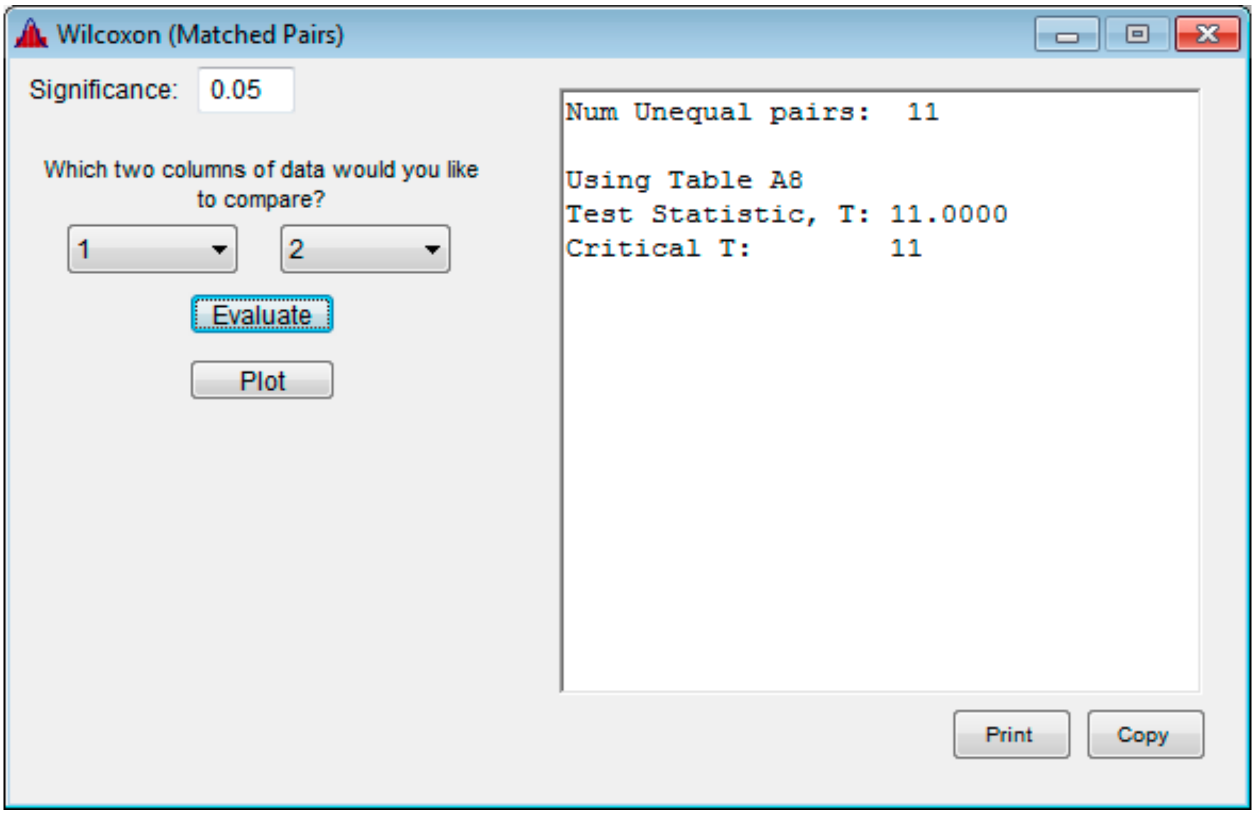
5. For the Wilcoxon Signed-Ranks test, select **Wilcoxon (Matched Pairs)**, then proceed to enter a significance level, such as 0.05 or 0.01.

6. Select the columns of the Sample Editor window that contain the paired data.

7. Click on the **Evaluate** button.

8. Click on the **Plot** button to see a graph that includes the test statistic and critical values.
   - The graph will be displayed only if the normal approximation is used (because $n > 30$).

Using the same matched data from Table 13-3 (or Example 1 in Section 13-3 of the textbook), the Statdisk results for the Wilcoxon signed-ranks test will be as shown below. Based on this display we see that the test statistic T = 11, the critical value is T = 6, and the conclusion is to reject the null hypothesis that the matched pairs have differences with a median equal to 0. There does appear to be a significant difference between the taxi-in times and the corresponding taxi-out times. (Note that the sign test in Section 13-2 led to the conclusion of no difference. By using only positive and negative signs, the sign test did not use the magnitudes of the differences, but the Wilcoxon signed-ranks test was more sensitive to those magnitudes through its use of ranks.)



13-4 Wilcoxon Rank-Sum Test

The Triola textbook (excluding *Essentials of Statistics*) discusses the Wilcoxon rank-sum test and includes the following definition.

> **DEFINITION** The **Wilcoxon rank-sum test** is a nonparametric test that uses ranks of sample data from two independent populations to test this null hypothesis: H_0: The two independent samples come from populations with equal medians. (The alternative hypothesis H_1 can be any one of the following three possibilities: The two populations have different medians, or the first population has a median *greater than* the median of the second population, or the first population has a median *less than* the median of the second population.)

First we describe the Statdisk procedure for conducting a Wilcoxon rank-sum test, then we illustrate it with an example.

Statdisk Procedure for the Wilcoxon Rank-Sum Test

1. Enter the two lists of values from the two independent samples in columns of the Statdisk Sample Editor window.

2. Select **Analysis** from the top menu bar.

3. Select **Wilcoxon Tests** from the subdirectory.

4. You must now choose between the following two options.
   - **Wilcoxon (Matched Pairs)**
   - **Wilcoxon (Indep. Samples)**

   For the Wilcoxon Rank-Sum test, select **Wilcoxon (Indep. Samples)**.

5. Proceed to enter a significance level, such as 0.05 or 0.01.

6. Select the columns of the Sample Editor window that contain the two sets of independent sample data.

7. Click on the **Evaluate** button.

8. Click on **Plot** to display a graph that shows the test statistic and critical values.

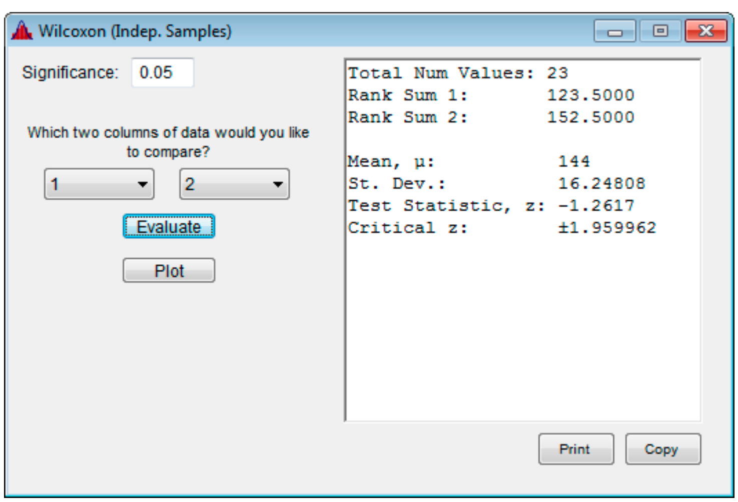
Statdisk

EXAMPLE Shown below are pulse rates of males and females (from Statdisk data set *1 – Body Measurements Male/Female*). (Ranks are shown in parentheses.) We want to use a 0.05 significance level to test the claim that males and females have the same median pulse rate (as in Example 1 in Section 13-4 of the textbook).

Table 13-5 Pulse Rates
(Ranks in parentheses)

Males	Females
60 (**4.5**)	78 (**14**)
74 (**11**)	80 (**17**)
86 (**19**)	68 (**9**)
54 (**1**)	56 (**2.5**)
90 (**20.5**)	76 (**12**)
80 (**17**)	78 (**14**)
66 (**7**)	78 (**14**)
68 (**9**)	90 (**20.5**)
68 (**9**)	96 (**22**)
56 (**2.5**)	60 (**4.5**)
80 (**17**)	98 (**23**)
62 (**6**)	
$n_1 = 12$	$n_2 = 11$
$R_1 = 123.5$	$R_2 = 152.5$

Shown below are the Statdisk results using the data from Table 13-5. Based on the Statdisk display, we fail to reject the null hypothesis. Based on the available sample data, it appears that there is not sufficient evidence to warrant rejection of the claim that males and females have pulse rates with the same median.



Statdisk

13-5 Kruskal-Wallis Test

The Triola textbook (excluding *Essentials of Statistics*) discusses the Kruskal-Wallis test and includes the following definition.

> **DEFINITION** The **Kruskal-Wallis Test** (also called the ***H* test**) is a nonparametric test that uses ranks of simple random samples from three or more independent populations to test the null hypothesis that the populations have the same median. (The alternative hypothesis is the claim that the populations have medians that are not all equal.)

We describe the Statdisk procedure for the Kruskal-Wallis test, then we illustrate it with an example.

Statdisk Procedure for the Kruskal-Wallis Test

1. Enter the samples of data in columns of the Statdisk Sample Editor window.

2. Select **Analysis** from the top menu bar.

3. Select **Kruskal-Wallis test** from the submenu.

4. In the dialog box, enter a significance level, such as 0.05 or 0.01.

5. Select the columns containing the samples of data. Click on the boxes to insert or delete check marks. Columns with check marks are included in the calculations.

6. Click on the **Evaluate** button.

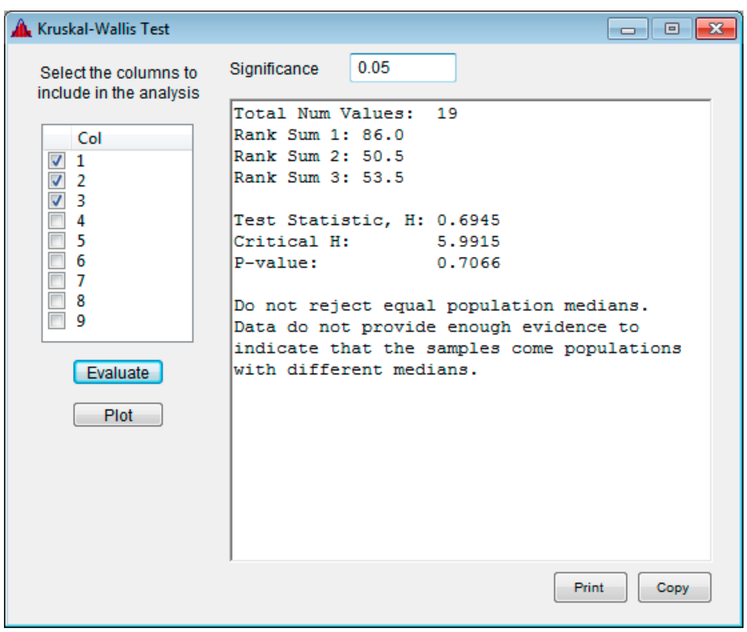
7. Click on **Plot** to display a graph that shows the test statistic and critical values.

> **EXAMPLE** Example 1 in Section 13-5 of the textbook illustrates the Kruskal-Wallis test with Performance IQ Score data in Table 13-6 on the following page.

Table 13-6 Performance IQ Scores

Low Lead Level	Medium Lead Level	High Lead Level
85 (**6.5**)	78 (**2**)	93 (**10**)
90 (**8.5**)	97 (**12.5**)	100 (**15.5**)
107 (**18.5**)	107 (**18.5**)	97 (**12.5**)
85 (**6.5**)	80 (**4**)	79 (**3**)
100 (**15.5**)	90 (**8.5**)	97 (**12.5**)
97 (**12.5**)	83 (**5**)	
101 (**17**)		
64 (**1**)		
$n_1 = 8$	$n_2 = 6$	$n_3 = 5$
$R_1 = 86$	$R_2 = 50.5$	$R_3 = 53.5$

If we use Statdisk with a 0.05 significance level to test the claim that the three samples come from populations with medians that are all equal, we get the display shown below. Important elements of the display include the rank sums of 86.0, 50.5, and 53.5, the test statistic of H = 0.6945, the critical value of H = 5.9915, and the P-value of 0.7066. Remember that the Kruskal-Wallis test is a *right-tailed* test. Because the P-value of 0.7066 is greater than the significance level of 0.05, we fail to reject the null hypothesis. There is not sufficient evidence to reject the claim that the performance IQ scores from subjects with low lead exposure, medium lead exposure, and high lead exposure all have the same median. The population medians do not appear to be different.



13-6 Rank Correlation

The Triola textbook introduces *rank correlation*, which uses ranks in a procedure for determining whether there is some relationship between two variables.

> **DEFINITION** The **rank correlation test** (or Spearman's rank correlation test) is a nonparametric test that uses ranks of sample data consisting of matched pairs. It is used to test for an association between two variables.

First we describe the Statdisk procedure, then we illustrate it with an example.

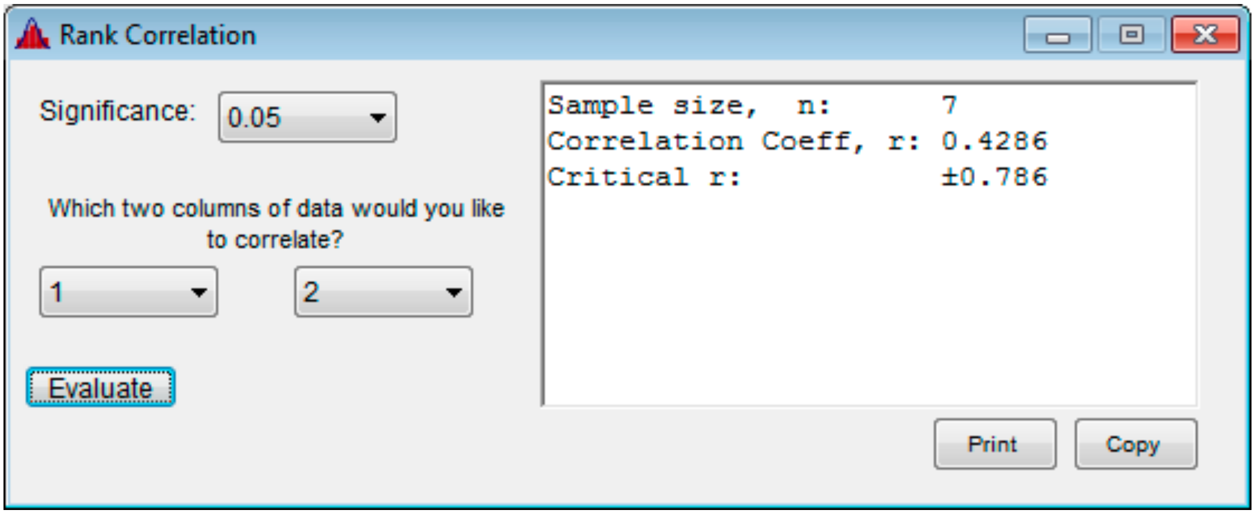
Statdisk Procedure for Rank Correlation

1. Enter the paired sample data in columns of the Sample Editor window.
2. Select **Analysis** from the top menu bar.
3. Select **Rank Correlation** from the submenu.
4. Enter a significance level, such as 0.05 or 0.01.
5. Select the columns containing the paired data to be used for the calculations.
6. Click on the **Evaluate** button.
7. Click on the **Plot** button to obtain a graph that shows the test statistic and critical values. The graph will be displayed only if the normal approximation is used (because $n > 30$).

Consider the sample data in the Table 13-7 below (from Example 1 in Section 13-6 of the textbook). The Statdisk display obtained from this data follows.

Table 13-7 Overall Quality Scores and Prices of LCD Televisions

Quality rank	1	2	3	4	5	6	7
Price (dollars)	1900	1200	1300	2000	1700	1400	2700



Because the test statistic (r = 0.4286) is between the critical r values ±0.786, we fail to reject the null hypothesis. There is not sufficient evidence to support a claim of a correlation between quality and price.

13-7 Runs Test for Randomness

The Triola textbook (excluding *Essentials of Statistics*) discusses the runs test for randomness and includes these definitions.

> **DEFINITION** After characterizing each data value as one of two separate categories, a **run** is a sequence of data having the same characteristic; the sequence is preceded and followed by data with a different characteristic or by no data at all.

The **runs test** uses the number of runs in a sequence of sample data to test for randomness in the order of the data.

First we describe the Statdisk procedure for the runs test for randomness, then we illustrate it with an example.

Statdisk Procedure for the Runs Test for Randomness

1. Using the original data, count the number of runs, the number of elements of the first type, and the number of elements of the second type.

2. Select **Analysis** from the top menu bar.

3. Select **Runs Test** from the submenu.

4. Make these entries in the dialog box:
    - Enter a significance level, such as 0.05 or 0.01.
    - Enter the number of runs.
    - Enter the number of elements of the first type.
    - Enter the number of elements of the second type.

5. Click on the **Evaluate** button.

6. Click on **Plot** to display a graph with the test statistic and critical values.

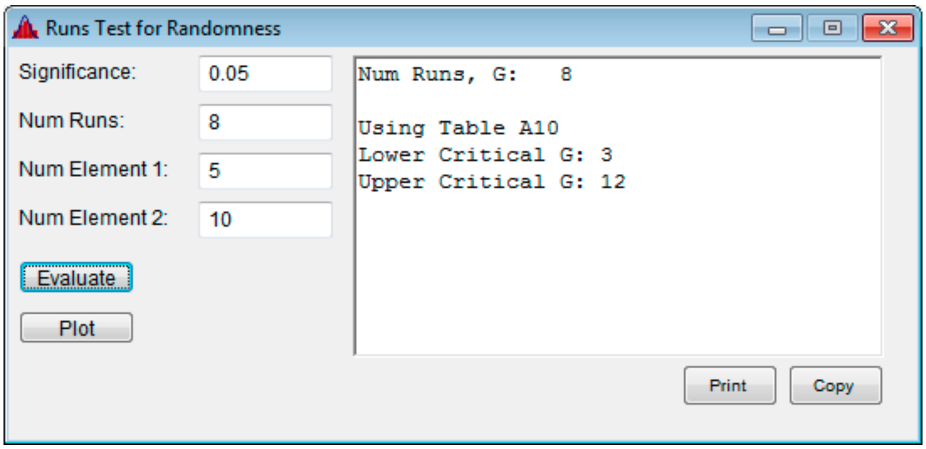
Consider the winners of the NBA basketball championship game, with W denoting a winner from the Western Conference and E denoting a winner from the Eastern Conference (as in Example 1 in Section 13-7 of the textbook). We get the following sequence of E's and W's, which we then convert to 0's and 1's as shown below.

E	**E**	**W**	**W**	**W**	**W**	**W**	**E**	**W**	**E**	**W**	**E**	**W**	**W**	**W**
0	**0**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**1**	**1**

The textbook describes the procedure for examining the above sequence to find these results:

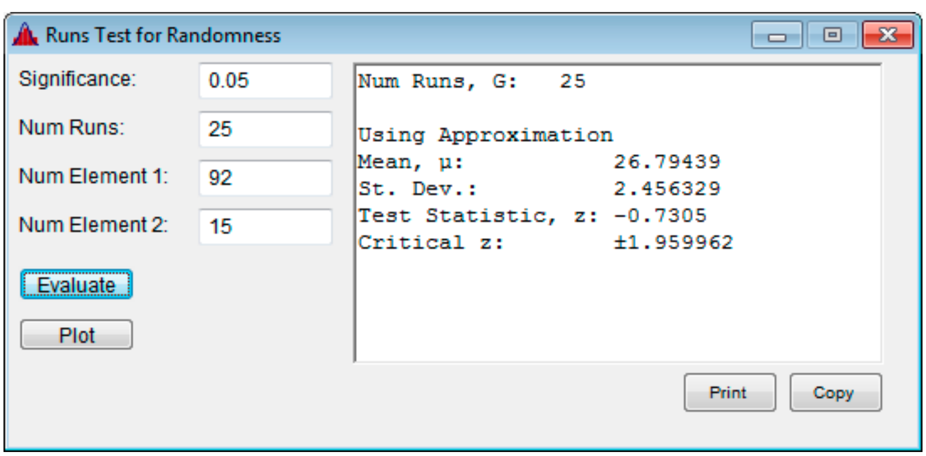
G = number of runs = 8
n_1 = number of Es = 5
n_2 = number of Ws = 10

Using Statdisk for the runs test for randomness, we obtain the display shown below. We fail to reject the null hypothesis. There is not sufficient evidence to reject randomness in the sequence of winners. Randomness cannot be rejected.



Runs Test with Large Samples

Example 3 in the textbook involves a large sample resulting in 25 runs with n_1 = 92 and n_2 = 15. Instead of performing complicated calculations for μ_G and σ_G and the test statistic, Statdisk automatically displays results of μ_G = 26.79439, σ_G = 2.456329, the test statistic z = −0.7305, and the critical values of z = ±1.959962.



CHAPTER 13 EXPERIMENTS: Nonparametric Statistics

*In Experiments 13–1 through 13–7, use Statdisk's **Sign Test** program.*

13-1 **Oscar Winners** Listed below are ages of actresses and actors at the times that they won Oscars. The data are paired according to the years that they won. Use a 0.05 significance level to test the claim that there is no difference between the ages of best actresses and the ages of best actors at the time that the awards were presented.

Best Actresses	33	35	35	28	30	29	61	32	33	45
Best Actors	36	47	29	43	37	38	45	50	48	60

Test statistic:_____ Critical value:_____

Conclusion:

____

____

13–2. **Oscar Winners** Repeat the preceding exercise using the 82 pairs of ages listed in Statdisk *12th Edition* data set *11 – Ages of Oscar Winner*.

Test statistic:_____ Critical value:_____

Conclusion:

____

____

13-3 **Flight Data** Use the following times for American Airlines Flight 19 from New York (JFK) to Los Angeles (LAX). Use a 0.05 significance level to test the claim that there is no difference between taxi-out times and taxi-in times.

Taxi-Out Time (min)	15	12	19	18	21	20	13	15	43	18	17	19
Taxi-In Time (min)	10	10	16	13	9	8	4	3	8	16	9	5

Test statistic:_____ Critical value:_____

Conclusion:

____

____

Statdisk

13-4 **Flight Data** The preceding example uses flight data from American Airlines Flight 19, and Example 2 in Section 13-2 of the textbook uses the data from Flight 21. Statdisk *12th Edition* data set *15 – Flight Data* includes sample data for 48 flights from Flights 1, 3, 19, and 21. Open this data set in Statdisk and use the paired taxi-out times and taxi-in times from those 48 flights. Use a 0.05 significance level to test the claim that there is no difference between taxi-out times and taxi-in times.

Test statistic:_____ Critical value:_____

Conclusion:

____

____

13-5 **Gender Selection** The Genetics and IVF Institute conducted a clinical trial of its methods for gender selection. As of this writing, 239 of 291 babies born to parents using the YSORT method were boys. Use a 0.01 significance level to test the claim that the YSORT method has no effect.

Test statistic:_____ Critical value:_____

Conclusion:

____

____

13-6 **Earthquake Magnitudes** Refer to Statdisk 12th Edition data set *16 – Earthquake Measurements* for the earthquake magnitudes. Use a 0.01 significance level to test the claim that the median is equal to 1.00.

Test statistic:___________ Critical value: ___________

Conclusion:

____

____

Statdisk

13-7 **Testing for Median Weight of Quarters** Refer to Statdisk *12th Edition* data set *21 – Coin Weights* for the weights (in g) of randomly selected quarters that were minted after 1964. The quarters are supposed to have a median weight of 5.670 g. Use a 0.01 significance level to test the claim that the median is equal to 5.670 g. Do quarters appear to be minted according to specifications?

Test statistic:__________ Critical value: __________

Conclusion:

____

____

Experiments 13–8 through 13–17, use Statdisk's ***Wilcoxon Tests*** *program.*

13-8 **Sign Test vs. Wilcoxon Signed–Ranks Test** Repeat Experiment 13-1 by using the Wilcoxon signed-ranks test for matched pairs. Enter the Statdisk results below, and compare them to the sign test results obtained in Experiment 13-1. Specifically, how do the results reflect the fact that the Wilcoxon signed-ranks test uses more information?

Test statistic:__________ Critical value: __________

Conclusion:

____

____

Comparison: ____

____

13-9 **Sign Test vs. Wilcoxon Signed–Ranks Test** Repeat Experiment 13-2 by using the Wilcoxon signed-ranks test for matched pairs. Enter the Statdisk results below, and compare them to the sign test results obtained in Experiment 13-2. Specifically, how do the results reflect the fact that the Wilcoxon signed-ranks test uses more information?

Test statistic:__________ Critical value: __________

Conclusion:

____

____

Comparison: ____

____

13-10 **Sign Test vs. Wilcoxon Signed–Ranks Test** Repeat Experiment 13-3 by using the Wilcoxon signed-ranks test for matched pairs. Enter the Statdisk results below, and compare them to the sign test results obtained in Experiment 13-3. Specifically, how do the results reflect the fact that the Wilcoxon signed-ranks test uses more information?

Test statistic:___________ Critical value: ___________

Conclusion:

____

____

Comparison: ____

____

13-11 **Sign Test vs. Wilcoxon Signed–Ranks Test** Repeat Experiment 13-4 by using the Wilcoxon signed-ranks test for matched pairs. Enter the Statdisk results below, and compare them to the sign test results obtained in Experiment 13-4. Specifically, how do the results reflect the fact that the Wilcoxon signed-ranks test uses more information?

Test statistic:___________ Critical value: ___________

Conclusion:

____

____

Comparison: ____

____

13-12 **Earthquakes** Refer to Statdisk *12th Edition* data set *16 – Earthquake Measurements* for the sample of paired earthquake magnitudes and depths. Use a 0.10 significance level with the Wilcoxon signed-ranks test to test the claim of no difference between the magnitudes and the depths.

Test statistic:__________ Critical value: __________

Conclusion:

____

____

What is fundamentally wrong with this analysis?

____

13-13 **Coke Contents** Refer to Statdisk *12th Edition* data set *19 – Cola Weights and Volumes* for the amounts (in oz) in cans of regular Coke. The cans are labeled to indicate that the contents are 12 oz of Coke. Use a 0.05 significance level with the Wilcoxon signed-ranks test to test the claim that cans of Coke are filled so that the median amount is 12 oz.

Test statistic:__________ Critical value: __________

Conclusion:

____

____

13-14 **IQ and Lead Exposure** Statdisk *12th Edition* data set *5 – IQ and Lead Exposure* lists full IQ scores for a random sample of subjects with medium lead levels in their blood and another random sample of subjects with high lead levels in their blood. Use a 0.05 significance level with the Wilcoxon rank-sum test to test the claim that subjects with medium lead levels have full IQ scores with a higher median than the median full IQ score for subjects with high lead levels.

Test statistic:__________ Critical value: __________

Conclusion:

____

____

13-15 **Weights of Coke** Statdisk *12th Edition* data set *19 – Cola Weights and Volumes* lists weights (lb) of the cola in cans of regular Coke and diet Coke. Use a 0.05 significance level with the Wilcoxon rank-sum test to test the claim that the samples are from populations with the same median.

Test statistic:__________ Critical value: __________

Conclusion:

____

____

13-16 **Cigarettes** Refer to Statdisk *12th Edition* data set *10 – Cigarette Tar, Nicotine, and Carbon Monoxide* for the amounts of nicotine (in mg per cigarette) in the sample of king size cigarettes (KgNic), which are non-filtered, non-menthol, and non-light, and for the amounts of nicotine in the 100 mm cigarettes (FLNic), which are filtered, non-menthol, and non-light. Use a 0.01 significance level with the Wilcoxon rank-sum test to test the claim that the median amount of nicotine in the non-filtered king size cigarettes is greater than the median amount of nicotine in the 100 mm filtered cigarettes.

Test statistic:__________ Critical value: __________

Conclusion:

____

____

13-17 **Cigarettes** Refer to Statdisk *12th Edition* data set *10 – Cigarette Tar, Nicotine, and Carbon Monoxide* for the amounts of tar in the sample of king size cigarettes (KgTar), which are non-filtered, non-menthol, and non-light, and for to the amounts of tar in the 100 mm cigarettes (FLTar), which are filtered, non-menthol, and non-light. Use a 0.01 significance level with the Wilcoxon rank-sum test to test the claim that the median amount of tar in the non-filtered king size cigarettes is greater than the median amount of tar in the 100 mm filtered cigarettes.

Test statistic:__________ Critical value: __________

Conclusion:

____

____

In Experiments 13–18 through 13–22, use Statdisk's ***Kruskal-Wallis Test*** *program.*

13-18 **Do All Colors of M&Ms Weigh the Same?** Refer to the Statdisk *12th Edition* data set *20 – M and M Plain Candy Weights*. Use a 0.05 significance level with the Kruskal-Wallis test to test the claim that the weights of M&Ms have the same median for each of the six different color populations.

Test statistic:__________ Critical value: __________

Conclusion:

____

____

13-19 **Nicotine in Cigarettes** Refer to Statdisk *12th Edition* data set *10 – Cigarette Tar, Nicotine, and Carbon Monoxide* and use the amounts of nicotine (mg per cigarette) in the king size cigarettes (KgNic), the 100 mm menthol cigarettes (MnNic), and the 100 mm non-menthol cigarettes (FLNic). The king size cigarettes are non-filtered, non-menthol, and non-light. The 100 mm menthol cigarettes are filtered and non-light. The 100 mm non-menthol cigarettes are filtered and non-light**.** Use a 0.05 significance level with the Kruskal-Wallis test to test the claim that the three categories of cigarettes yield the same median amount of nicotine.

Test statistic:__________ Critical value: __________

Conclusion:

____

____

13-20 **Tar in Cigarettes** Refer to Statdisk *12th Edition* data set *10 – Cigarette Tar, Nicotine, and Carbon Monoxide* and use the amounts of tar (mg per cigarette) in the three categories of cigarettes described in the preceding exercise (KgTar, MnTar, FLTar)**.** Use a 0.05 significance level with the Kruskal-Wallis test to test the claim that the three categories of cigarettes yield the same median amount of tar.

Test statistic:__________ Critical value: __________

Conclusion:

____

____

13-21 **Carbon Monoxide from Cigarettes** Refer to Statdisk *12th Edition* data set *10 – Cigarette Tar, Nicotine, and Carbon Monoxide* and use the amounts of carbon monoxide (mg per cigarette) in the three categories of cigarettes described in Exercise 19 (KgCO, MnCO, FLCO)**.** Use a 0.05 significance level with the Kruskal-Wallis test to test the claim that the three categories of cigarettes yield the same median amount of carbon monoxide.

Test statistic:__________ Critical value: __________

Conclusion:

____

____

13-22 **Passive and Active Smoke** Statdisk *12th Edition* data set *9 – Passive and Active Smoke* lists measured cotinine levels from a sample of subjects who smoke, another sample of subjects who do not smoke but are exposed to environmental tobacco smoke, and a third sample of subjects who do not smoke and are not exposed to environmental tobacco smoke. Cotinine is produced when the body absorbs nicotine. Use a 0.01 significance level to test the claim that the three samples are from populations with the same median.

Test statistic:__________ Critical value: __________

Conclusion:

____

____

*In Experiments 13–23 through 13–28, use Statdisk's **Rank Correlation** program.*

13-23 **Judges of Marching Bands** Two judges ranked seven bands in the Texas state finals competition of marching bands (Coppell, Keller, Grapevine, Dickinson, Poteet, Fossil Ridge, Heritage), and their rankings are listed below (based on data from the University Interscholastic League). Test for a correlation between the two judges. Do the judges appear to rank about the same or are they very different?

Band	Cpl	Klr	Grp	Dck	Ptt	FR	Her
First Judge	1	3	4	7	5	6	2
Second Judge	6	4	5	1	3	2	7

Rank Correlation Coefficient: __________

Critical Values: __________

Conclusion: ____

____

13-24 **Judges of Marching Bands** In the same competition described in the preceding experiment, a third judge ranked the bands with the results shown below. Test for a correlation between the first and third judges. Do the judges appear to rank about the same or are they very different?

Band	Cpl	Klr	Grp	Dck	Ptt	FR	Her
First Judge	1	3	4	7	5	6	2
Third Judge	3	4	1	5	7	6	2

Rank Correlation Coefficient: __________

Critical Values: __________

Conclusion: ____

____

13-25 **Blood Pressure** Refer to the measured systolic and diastolic blood pressure measurements of 40 randomly selected males in Statdisk *12th Edition* data set *1b – Body Measurements Male* and test the claim that among men, there is a correlation between systolic blood pressure and diastolic blood pressure.

Rank Correlation Coefficient: __________

Critical Values: __________

Conclusion: ____

____

13-26 **Blood Pressure** Refer to the measured systolic and diastolic blood pressure measurements of 40 randomly selected females in Statdisk *12th Edition* data set *1a – Body Measurements Female* and test the claim that among women, there is a correlation between systolic blood pressure and diastolic blood pressure.

Rank Correlation Coefficient: __________

Critical Values: __________

Conclusion: ____

____

13-27 **IQ and Brain Volume** Refer Statdisk *12th Edition* data set *6 – IQ and Brain Size* and test the claim that there is a correlation between brain volume and IQ score.

Rank Correlation Coefficient: __________

Critical Values: __________

Conclusion: ____

____

13-28 **Earthquakes** Refer to Statdisk *12th Edition* data set *16 – Earthquake Measurements* and test the claim that there is a correlation between magnitudes and depths of earthquakes.

Rank Correlation Coefficient: __________

Critical Values: __________

Conclusion: ____

____

In Experiments 13–23 through 13–28, use Statdisk's ***Runs Test*** *program.*

13-29 **Oscar Winners** Listed below are the genders of the younger winner in the Academy Awards categories of Best Actor and Best Actress for recent and consecutive years. Do the genders of the younger winners appear to occur randomly?

F F F M M F F F F F F M F F F M F F F

Number of Runs: ____________

Critical Values: __________

Conclusion: ____

13-30 **Testing for Randomness of Presidential Election Winners** The political parties of the winning candidates for a recent sequence of presidential elections are listed below, where D denotes Democratic party and R denotes Republican party. Does it appear that we elect Democrat and Republican candidates in a random sequence?

R R D R D R R R R D D R R R D D D
D D R R D D R R D R R R D D R R D

Number of Runs: ____________

Critical Values: __________

Conclusion: ____

Statdisk

13-31 **Baseball World Series Victories** Test the claim that the sequence of World Series wins by American League and National League teams is random. Given below are recent results, with American League and National League teams represented by A and N, respectively.

A N A N N N A A A A N A A A A N A N N A A N N A A A A N A N
N A A A A A N A N A N A N A A A A A A A N N A N A N N A A N
N N A N A N A N A A A N N A A N N N N A A A N A N A N A A A
N A N A A A N A N A A N A N A N

Number of Runs: ____________

Critical Values: __________

Conclusion: ____

13-32 **Stock Market** Listed below are the annual high values of the Dow Jones Industrial Average for a recent sequence of years (as of this writing). Test for randomness below and above the median.

969	995	943	985	969	842	951	1036	1052	892
882	1015	1000	908	898	1000	1024	1071	1287	1287
1553	1956	2722	2184	2791	3000	3169	3413	3794	3978
5216	6561	8259	9374	11568	11401	11350	10635	10454	10855
10941	12464	14198	13279	10580	11625				

Number of Runs: ____________

Critical Values: __________

Conclusion: ____

Statdisk

14 Statistical Process Control

Statdisk

The major topics of Chapter 14 from *Elementary Statistics* are run charts, control charts for variation, control charts for mean, and control charts for attributes. **Statdisk is not programmed to generate run charts or control charts**, but there are ways to obtain them by using Statdisk's scatterplot feature. There are other software packages, such as Minitab or StatCrunch, that are much easier to use when constructing run charts or control charts, but this chapter focuses on the use of Statdisk.

14-1 Run Charts

In the Statistical Process Control chapter (Chapter 14) in the Triola textbook (excluding *Essentials of Statistics*) we define **process data** to be data arranged according to some time sequence, such as the data in Table 14-1 below. The Chapter Problem for Chapter 14 in the textbook states that the measurements are from a new minting process. On each of 20 consecutive days of production, a quarter is selected during each of the first five hours of production and the quarter is weighed. The results are listed in the table.

Table 14-1 Weights (grams) of Minted Quarters

Day	Hour 1	Hour 2	Hour 3	Hour 4	Hour 5	$\bar{x}$	s	Range
1	5.543	5.698	5.605	5.653	5.668	5.6334	0.0607	0.155
2	5.585	5.692	5.771	5.718	5.720	5.6972	0.0689	0.186
3	5.752	5.636	5.660	5.680	5.565	5.6586	0.0679	0.187
4	5.697	5.613	5.575	5.615	5.646	5.6292	0.0455	0.122
5	5.630	5.770	5.713	5.649	5.650	5.6824	0.0581	0.140
6	5.807	5.647	5.756	5.677	5.761	5.7296	0.0657	0.160
7	5.686	5.691	5.715	5.748	5.688	5.7056	0.0264	0.062
8	5.681	5.699	5.767	5.736	5.752	5.7270	0.0361	0.086
9	5.552	5.659	5.770	5.594	5.607	5.6364	0.0839	0.218
10	5.818	5.655	5.660	5.662	5.700	5.6990	0.0689	0.163
11	5.693	5.692	5.625	5.750	5.757	5.7034	0.0535	0.132
12	5.637	5.628	5.646	5.667	5.603	5.6362	0.0235	0.064
13	5.634	5.778	5.638	5.689	5.702	5.6882	0.0586	0.144
14	5.664	5.655	5.727	5.637	5.667	5.6700	0.0339	0.090
15	5.664	5.695	5.677	5.689	5.757	5.6964	0.0359	0.093
16	5.707	5.890	5.598	5.724	5.635	5.7108	0.1127	0.292
17	5.697	5.593	5.780	5.745	5.470	5.6570	0.1260	0.310
18	6.002	5.898	5.669	5.957	5.583	5.8218	0.1850	0.419
19	6.017	5.613	5.596	5.534	5.795	5.7110	0.1968	0.483
20	5.671	6.223	5.621	5.783	5.787	5.8170	0.2380	0.602

A *run chart* is a sequential plot of *individual* data values over time. To use Statdisk for generating a run chart, pair the data with the consecutive positive integers, then generate a scatterplot. The table below shows the first eight weights paired with 1, 2, 3,…,8.

x	1	2	3	4	5	6	7	8
y	5.543	5.698	5.605	5.653	5.688	5.585	5.692	5.771

This table includes only the first eight values in Table 14-1, but it can extended to include all of the data. Continue with this procedure to enter the consecutive table values matched with the positive integers.

Statdisk Procedure for Run Chart

To generate a run chart in Statdisk, we generate a scatterplot using the following procedure.

1. Enter the process data in a column of the Statdisk Sampled Editor. Also enter the consecutive positive integers 1, 2, 3, . . . (stopping when the column has the same length as the column of the process data).

2. Select **Analysis** from the top menu bar.

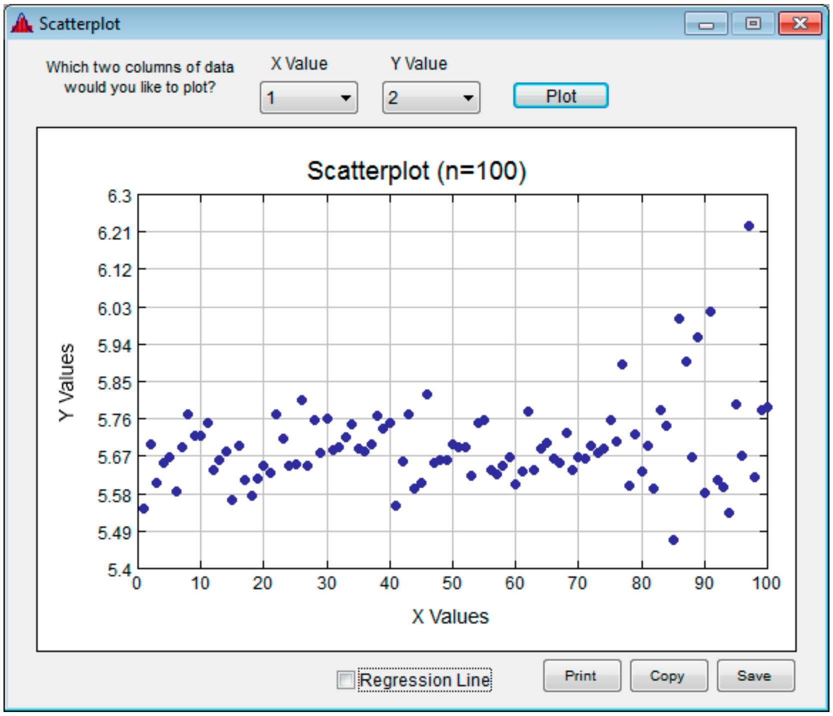
3. Select the menu item of **Data**, then select **Scatterplot**.

4. Click on the box next to "Regression Line" so that there is no check mark in that box, then click on the **Plot** button.

5. Manually connect the points in order from left to right, as shown in the Minitab – generated run chart.

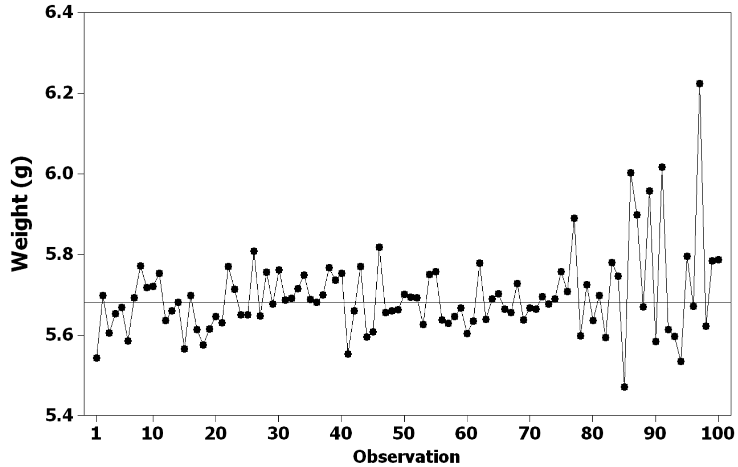
Shown below is the scatterplot of the 100 sample values from Table 14-1 paired with the positive integers 1, 2, 3, . . . , 100.

Statdisk Run Chart



Statdisk does not connect the points as shown in the above display, but these points can be manually connected to get a run chart such as the Minitab–generated run chart shown below. Go ahead and connect the dots in the above Statdisk display; it's more fun than human beings should be allowed to have. In so doing, you will effectively create the same run chart included in Section 14-2 of the textbook.

Minitab– Generated Run Chart



Examine the run chart and note that it reveals this problem: As time progresses from left to right, the heights of the points appear to show a pattern of increasing variation. See how the points at the left fluctuate considerably less than the points farther to the right. It appears that the manufacturing process started out well, but deteriorated as time passed. If left alone, this minting process will continue to deteriorate because the run chart suggests that the process is not **statistically stable** (or **within statistical control**) because it has a pattern of increasing variation.

14-2 Control Charts for Variation

The textbook describes ***R* charts** as sequential plots of ranges. Using the data from the preceding table, for example, the *R* chart is a plot of the ranges 0.155, 0.186, 0.187, . . . , 0.602. *R* charts are used to monitor the *variation* of a process. The procedure for obtaining an *R* chart is as follows.

Statdisk Procedure for *R* Chart to Monitor Process Variation

1. Find the value of the range for each individual sample (as in Table 14-1).
2. Using the methods described in Section 14-2 of *Elementary Statistics,* determine the values to be used for the upper control limit, the centerline, and the lower control limit.
3. Enter the following values in the first column of the Statdisk Data Window:
   0, 0, 0, 1, 2, 3, . . . (stopping when you reach the number of sample ranges).
4. In the second column of the Statdisk Data Window, enter the value of the upper control limit, then the value for the centerline, then the value for the lower control limit, followed by the values of the sample ranges. That is, follow the general format shown below.

Column 1	Column 2
0	(Value of upper control limit)
0	(Value of $\overline{R}$ for the centerline)
0	(Value of lower control limit)
1	(Value of first sample range)
2	(Value of second sample range)
3	(Value of third sample range)
and so on	

Example 3 in Section 14–2 of *Elementary Statistics* shows how to determine the values of the upper control limit (0.4342), the centerline (0.2054), and the lower control limit (0.0000). Enter those three values in column two followed by the list of sample ranges (0.155, 0.186, 0.187, . . . , 0.602) as shown in the following list and the Statdisk Sample Editor window.

Column 1	Column 2	
0	0.4342	←Upper control limit
0	0.2054	←Centerline
0	0.0000	←Lower control limit
1	0.155	←First sample range
2	0.186	←Second sample range
3	0.187	←Third sample range
and so on		

Statdisk Sample Editor

Row	1	2
1	0	.4342
2	0	0.2054
3	0	0.0000
4	1	0.155
5	2	0.186
6	3	0.187
7	4	0.122
8	5	0.14
9	6	0.16
10	7	0.062
11	8	0.086
12	9	0.218
13	10	0.163
14	11	0.132
15	12	0.064
16	13	0.144
17	14	0.09
18	15	0.093
19	16	0.292
20	17	0.31
21	18	0.419
22	19	0.483
23	20	0.602

5. Click on **Data**, then select **Scatterplot**.

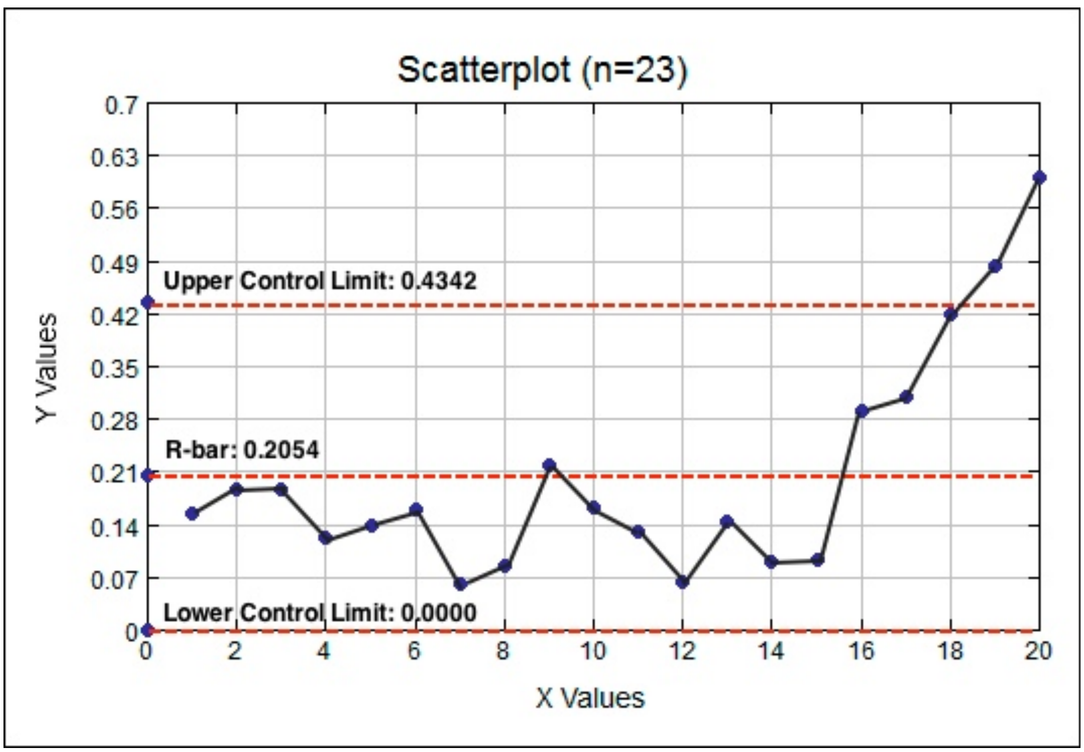
6. Select the columns that contain the data, then click on **Plot**.

7. Click the "Regression Line" box so that there is no checkmark and no regression line appears on the chart.

8. The three leftmost points will be stacked above 0. Use those points to position the upper control limit, the centerline, and the lower control limit. For the remaining points beginning above 1, connect the points in order from left to right.

The Statdisk *R* chart for the data in Table 14-1 is shown below. The graph was manually modified to include lines and labels for the upper control limit, the centerline, and the lower control limit. Note that lines for the upper control limit, centerline, and lower control limit, were located by including these three points as the first three points in the list of paired data:
(0, 0.6712), (0, 0.3777), (0, 0.0842).

Statdisk *R* Chart



Statdisk

We can interpret the above *R* chart by applying these three criteria for determining whether a process is within statistical control:

1. There is no pattern, trend, or cycle that is obviously not random.
2. No point lies beyond the upper or lower control limits.
3. There are not 8 consecutive points all above or all below the center line.

Analyzing the above *R* chart leads to the conclusion that the process *is out of statistical control* because the second criterion is violated: There is a point beyond the upper control limit. Also, the third criterion is violated because there is a pattern of an upward trend, indicating that the process is experiencing increasing variation.

14-3 Control Charts for Mean

Control charts for $\bar{x}$ can be constructed by using the same methods described for *R* charts. You must first find the sample means and the values to be used for the upper control limit, the centerline, and the lower control limit. See Section 14-2 in *Elementary Statistics* for the procedures that can be used to find these values.

Statdisk Procedure for $\bar{x}$ Chart to Monitor Process Mean

1. Find the value of the mean for each individual sample.

2. Determine the values to be used for the upper control limit, the centerline, and the lower control limit. (See Section 14-2 in *Elementary Statistics*.)

3. Enter the following values in the first column of the Statdisk data window:
   0, 0, 0, 1, 2, 3, . . . (stopping when reaching the number of sample means).

4. In the second column of the Statdisk data window, enter the value of the upper control limit, then the value for the centerline, then the value for the lower control limit, followed by the values of the sample means. Using the sample means in Table 14-1 along with the locations of the upper control limit (5.8140), the centerline (5.6955) and the lower control limit (5.5770), the Statdisk Sample Editor window should appear as shown below. (Example 5 in Section 14-2 of *Elementary Statistics* shows how the values of 5.8140, 5.6955, and 5.5770 are obtained.)

Statdisk Sample Editor

Row	1	2
1	0	5.8140
2	0	5.6955
3	0	5.5770
4	1	5.6334
5	2	5.6972
6	3	5.6586
7	4	5.6292
8	5	5.6824
9	6	5.7296
10	7	5.7056
11	8	5.727
12	9	5.6364
13	10	5.699
14	11	5.7034
15	12	5.6362
16	13	5.6882
17	14	5.67
18	15	5.6964
19	16	5.7108
20	17	5.657
21	18	5.8218
22	19	5.711
23	20	5.817

5. Click on **Data**, then select **Scatterplot**.

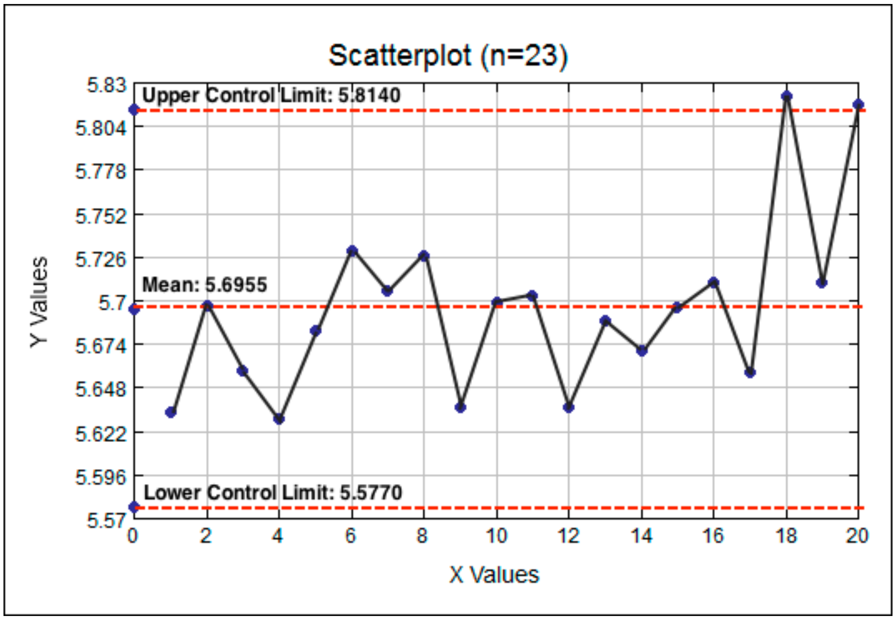
6. Select the columns that contain the data, then click on **Plot**.

7. Click the "Regression Line" box so that there is no checkmark and no regression line appears on the chart.

8. The three leftmost points will be stacked above 0. Use those points to position the upper control limit, the centerline, and the lower control limit. For the remaining points beginning above 1, manually connect the points in order from left to right.

Shown below is the Statdisk $\bar{x}$ chart for the data in Table 14-1. The graph was manually modified to include lines and labels for the upper control limit, the centerline, and the lower control limit. Note that the lines for the upper control limit, centerline, and lower control limit were located by including these three points as the first three points in the list of paired data: (0, 5.8140), (0, 5.6955), (0, 5.5770).

Statdisk $\bar{x}$ Chart



We can interpret the $\bar{x}$ chart by applying the three out-of-control criteria given in the textbook. We conclude that the mean in this process is *out of statistical control* because there is at least one point lying beyond the upper control limit.

14-4 Control Charts for Attributes

A control chart for attributes (or *p* chart) can also be constructed by using the same procedure for *R* charts and $\bar{x}$ charts. A *p* chart is very useful in monitoring some process proportion, such as the proportions of defects over time. See Example 1 in Section 14-3 in the textbook. Listed below are the numbers of defects in batches of 10,000 randomly selected quarters each day from a new manufacturing process that is being tested.

Defects: 8 7 12 9 6 10 10 5 15 14 12 14 9 6 16 18 20 19 18 24

Follow these steps to generate a Statdisk *p* chart:

Statdisk Procedure for *p* Charts

1. Collect the list of sample proportions. The above numbers of defects are from batches of 10,000, so the proportions of defects are:
    0.0008, 0.0007, 0.0012, 0.0009, 0.0006, 0.0010, 0.0010, 0.0005, 0.0015, 0.0014, 0.0012, 0.0014, 0.0009, 0.0006, 0.0016, 0.0018, 0.0020, 0.0019, 0.0018, and 0.0024.

2. Determine the values to be used for the upper control limit, the centerline, and the lower control limit. (See Section 14-3 in *Elementary Statistics*.) For the above sample data, the upper control limit is at 0.002324, the centerline is at 0.00126, and the lower control limit is at 0.000196.

3. Enter the following values in the first column of the Statdisk Sample Editor window:
    0, 0, 0, 1, 2, 3, ... (stop when reaching the number of sample proportions).

4. In the second column of the Statdisk data window, enter the value of the upper control limit, then the value for the centerline, then the value for the lower control limit, followed by the values of the sample proportions. For Example 1 in Section 14-3 of the textbook the entries in the dialog box should appear as shown below.

Statdisk Sample Editor

Row	1	2
1	0	0.002324
2	0	0.00126
3	0	0.000196
4	1	0.0008
5	2	0.0007
6	3	0.0012
7	4	0.0009
8	5	0.0006
9	6	0.0010
10	7	0.0010
11	8	0.0005
12	9	0.0015
13	10	0.0014
14	11	0.0012
15	12	0.0014
16	13	0.0009
17	14	0.0006
18	15	0.0016
19	16	0.0018
20	17	0.0020
21	18	0.0019
22	19	0.0018
23	20	0.0024

5. Click on **Data**, then select **Scatterplot**.

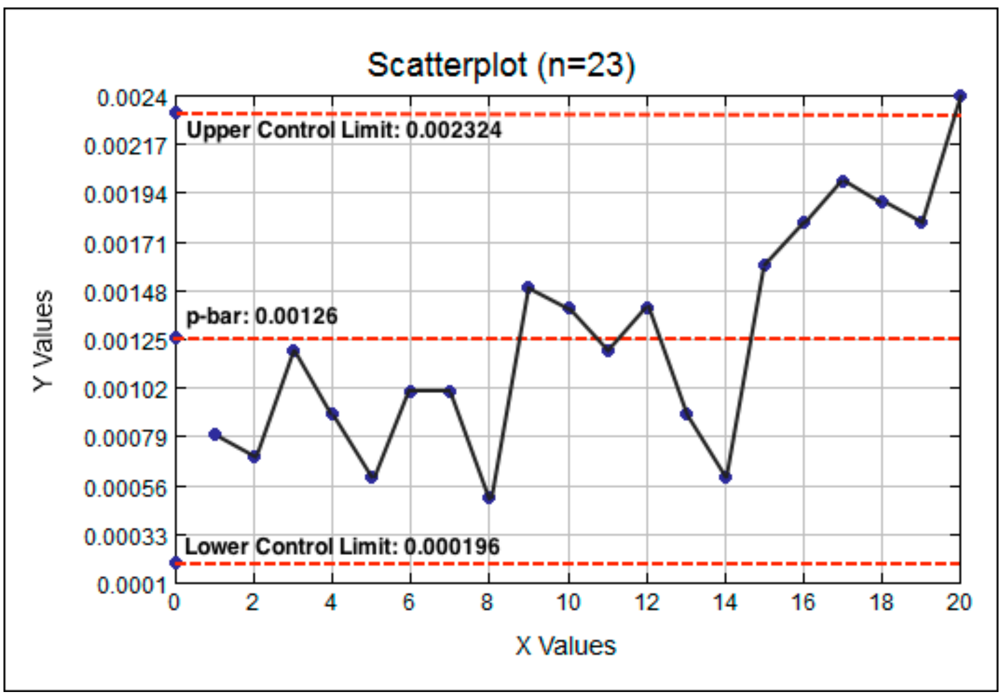
6. Select the columns that contain the data, then click on **Plot**.

7. Click the "Regression Line" box so that there is no checkmark and no regression line appears on the chart.

8. The three leftmost points will be stacked above 0. Use those points to position the upper control limit, the centerline, and the lower control limit. For the remaining points beginning above 1, manually connect the points in order from left to right.

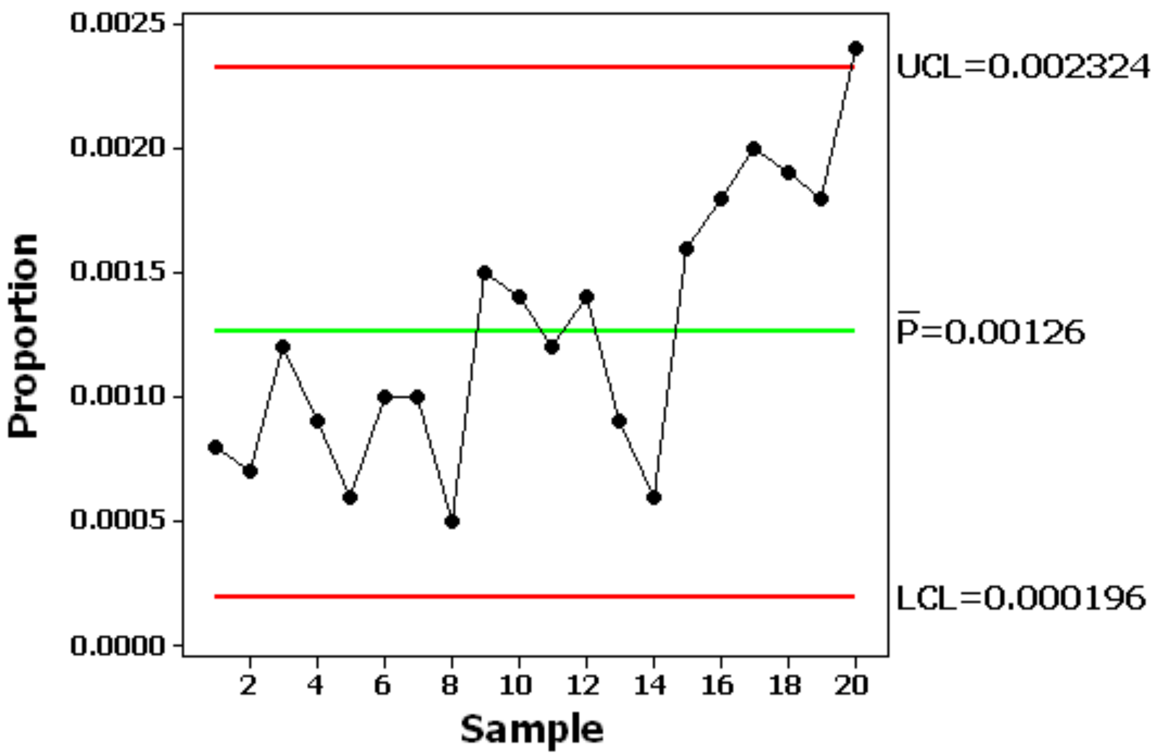
The following control chart for *p* can be interpreted by using the same three out-of-control criteria listed in Section 14-2 of the textbook. Using those criteria, we conclude that this process is out of statistical control for these reasons: There appears to be a downward trend. Also, there are 8 consecutive points lying above the centerline, and there are also 8 consecutive points lying below the centerline. Although the process is out of statistical control, it appears to have been somehow improved, because the proportion of defects has dropped. The company would be wise to investigate the process so that the cause of the lowered rate of defects can be understood and continued in the future.

Statdisk *p* Chart



Minitab and some other statistical software automatically generate run charts and control charts that are so important for monitoring process data over time. Shown below is the Minitab *p* chart that is automatically generated with lines for the upper control limit, centerline, and lower control limit. The use of such charts is increasing as more businesses recognize that this statistical tool can be effective in increasing quality and lowering costs.

Minitab *p* Chart



Statdisk

CHAPTER 14 EXPERIMENTS: Statistical Process Control

Constructing Control Charts for Aluminum Cans Experiments 1 and 2 are based on the axial loads(in pounds) of aluminum cans that are 0.0109 in. thick, as listed in the Statdisk *12th Edition* data set *22 – Axial Loads of Aluminum Cans*. An axial load of a can is the maximum weight supported by its side, and it is important to have an axial load high enough so that the can isn't crushed when the top lid is pressed into place. The data are from a real manufacturing process, and they were provided by a student who used an earlier version of this book.

14–1 ***R* Chart** On each day of production, seven aluminum cans with thickness 0.0109 in. were randomly selected and the axial loads were measured. The ranges for the different days are listed below, but they can also be found from the values given in Statdisk 12th Edition data set *22 – Axial Loads of Aluminum Cans*. Construct an *R* chart and determine whether the process variation is within statistical control. If it is not, identify which of the three out-of-control criteria lead to rejection of statistically stable variation.

78 77 31 50 33 38 84 21 38 77 26 78 78
17 83 66 72 79 61 74 64 51 26 41 31

14–2. $\bar{x}$ **Chart** On each day of production, seven aluminum cans with thickness 0.0109 in. were randomly selected and the axial loads were measured. The means for the different days are listed below, but they can also be found from the values given in Statdisk 12th Edition data set *22 – Axial Loads of Aluminum Cans*. Construct an $\bar{x}$ chart and determine whether the process mean is within statistical control. If it is not, identify which of the three out-of-control criteria lead to rejection of statistically stable variation.

252.7 247.9 270.3 267.0 281.6 269.9 257.7 272.9 273.7 259.1
275.6 262.4 256.0 277.6 264.3 260.1 254.7 278.1 259.7 269.4
266.6 270.9 281.0 271.4 277.3

Energy Consumption In Experiments 14-3 through 14-5, refer to Statdisk *12th Edition* data set *18 – Voltage Measurements from a Home* and use the measured voltage amounts for the power supplied directly to the author's home. Let each subgroup consist of the five amounts within the business days of a week, so the first five voltages constitute the first subgroup, the second five voltages constitute the second subgroup, and so on. The result is eight subgroups with five values each.

14-3 **Home Voltage: $\bar{x}$ Chart** Using subgroups of five voltage amounts, construct an $\bar{x}$ chart and determine whether the process mean is within statistical control. If it is not, identify which of the three out-of-control criteria lead to rejection of a statistically stable mean.

14-4 **Home Voltage: Run Chart** Construct a run chart for the 40 voltage amounts. Does there

Statdisk

appear to be a pattern suggesting that the process is not within statistical control?

14-5 ***Home Voltage: R Chart*** Using subgroups of five voltage amounts, construct an *R*chart and determine whether the process variation is within statistical control. If it is not, identify which of the three out-of-control criteria lead to rejection of statistically stable variation.

Energy Consumption In Experiments 14-6 through 14-8, use the following amounts of electricity consumed (in kWh) in the author's home. Let each subgroup consist of the six amounts within the same year, so that there are eight subgroups with six amounts in each subgroup.

Year 1	3637	2888	2359	3704	3432	2446
Year 2	4463	2482	2762	2288	2423	2483
Year 3	3375	2661	2073	2579	2858	2296
Year 4	2812	2433	2266	3128	3286	2749
Year 5	3427	578	3792	3348	2937	2774
Year 6	3016	2458	2395	3249	3003	2118
Year 7	4261	1946	2063	4081	1919	2360
Year 8	2853	2174	2370	3480	2710	2327

14-6 **Energy Consumption: *R* Chart** Let each subgroup consist of the 6 values within a year. Construct an *R* chart and determine whether the process variation is within statistical control. If it is not, identify which of the three out-of-control criteria lead to rejection of statistically stable variation.

14-7 **Energy Consumption: $\bar{x}$ Chart** Let each subgroup consist of the 6 values within a year. Construct an $\bar{x}$ chart and determine whether the process mean is within statistical control. If it is not, identify which of the three out-of-control criteria lead to rejection of a statistically stable mean.

14-8 **Energy Consumption: Run Chart** Construct a run chart for the 48 values. Does there appear to be a pattern suggesting that the process is not within statistical control?

14-9 ***p Chart for Defective Defibrillators*** Consider a process that includes careful testing of each manufactured defibrillator. Listed below are the numbers of defective defibrillators in successive batches of 10,000. Construct a control chart for the proportion *p* of defective defibrillators and determine whether the process is within statistical control. If not, identify which of the three out-of-control criteria apply.

Defects: 20 14 22 27 12 12 18 23 25 19 24 28 21 25 17 19 17 22 15 20

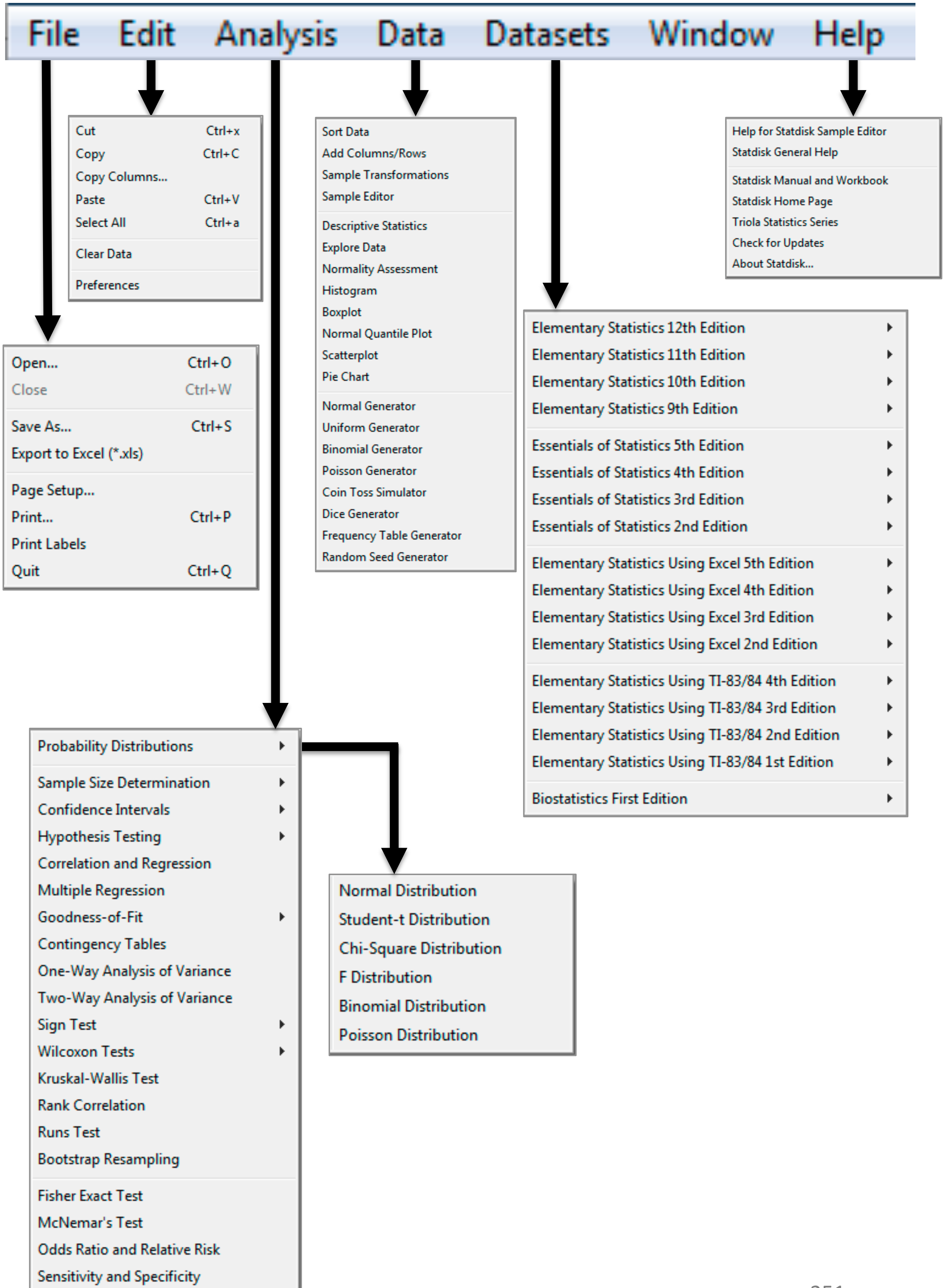
14-10 ***p Chart for Defective Defibrillators*** Repeat the preceding exercise assuming that the size of each batch is 100 instead of 10,000. Compare the control chart to the one found for the preceding exercise. Comment on the general quality of the manufacturing process described in the preceding exercise compared to the manufacturing process described in this exercise.

14-11 **Violent Crimes** In each of recent and consecutive years, 100,000 people in the United States were randomly selected and the number who were victims of violent crime was determined, with the results listed below. Does the rate of violent crime appear to exhibit acceptable behavior? (The values are based on data from the U.S. Department of Justice, and they are the most recent values available at the time of this writing.) How does the result affect us?

685 637 611 566 523 507 505 494 476 463 469 474 458 429

14-12 **Cola Cans** In ach of several consecutive days of production of cola cans, 500 cans are tested and the numbers of defects each day are listed below. What action should be taken?

20 22 19 17 19 15 16 13 14 14 11 13 12 11 10 9 9 10 7 7


File
Edit
Analysis
Data
Datasets
Window
Help
Cut Ctrl+x
Copy Ctrl+C
Copy Columns...
Paste Ctrl+V
Select All Ctrl+a
Clear Data
Preferences
Open... Ctrl+O
Close Ctrl+W
Save As... Ctrl+S
Export to Excel (*.xls)
Page Setup...
Print... Ctrl+P
Print Labels
Quit Ctrl+Q
Sort Data
Add Columns/Rows
Sample Transformations
Sample Editor
Descriptive Statistics
Explore Data
Normality Assessment
Histogram
Boxplot
Normal Quantile Plot
Scatterplot
Pie Chart
Normal Generator
Uniform Generator
Binomial Generator
Poisson Generator
Coin Toss Simulator
Dice Generator
Frequency Table Generator
Random Seed Generator
Help for Statdisk Sample Editor
Statdisk General Help
Statdisk Manual and Workbook
Statdisk Home Page
Triola Statistics Series
Check for Updates
About Statdisk...
Elementary Statistics 12th Edition
Elementary Statistics 11th Edition
Elementary Statistics 10th Edition
Elementary Statistics 9th Edition
Essentials of Statistics 5th Edition
Essentials of Statistics 4th Edition
Essentials of Statistics 3rd Edition
Essentials of Statistics 2nd Edition
Elementary Statistics Using Excel 5th Edition
Elementary Statistics Using Excel 4th Edition
Elementary Statistics Using Excel 3rd Edition
Elementary Statistics Using Excel 2nd Edition
Elementary Statistics Using TI-83/84 4th Edition
Elementary Statistics Using TI-83/84 3rd Edition
Elementary Statistics Using TI-83/84 2nd Edition
Elementary Statistics Using TI-83/84 1st Edition
Biostatistics First Edition
Probability Distributions
Sample Size Determination
Confidence Intervals
Hypothesis Testing
Correlation and Regression
Multiple Regression
Goodness-of-Fit
Contingency Tables
One-Way Analysis of Variance
Two-Way Analysis of Variance
Sign Test
Wilcoxon Tests
Kruskal-Wallis Test
Rank Correlation
Runs Test
Bootstrap Resampling
Fisher Exact Test
McNemar's Test
Odds Ratio and Relative Risk
Sensitivity and Specificity
Normal Distribution
Student-t Distribution
Chi-Square Distribution
F Distribution
Binomial Distribution
Poisson Distribution

Index