Statdisk Online

Student Laboratory Manual

and Workbook

Version 1.0

To Accompany

The Triola Statistics Series

Elementary Statistics, 14th Edition

Essentials of Statistics, 7th Edition

Elementary Statistics Using Excel, 7th Edition

Mario F. Triola

Pearson

Statdisk

Content Development: Robert Carroll
Content Management: Suzanna Bainbridge, Amanda Brands
Content Production: Jean Choe, Peggy McMahon
Product Management: Karen Montgomery
Product Marketing: Alicia Wilson
Rights and Permissions: Tanvi Bhatia /Anjali Singh

Please contact https://support.pearson.com/getsupport/s/contactsupport with any queries on this content.

Pearson

ISBN-10: 0137568797
ISBN-13: 9780137568796

Preface

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

- *Elementary Statistics*, 14th Edition
- *Essentials of Statistics*, 7th Edition
- *Elementary Statistics Using Excel*, 7th Edition

Accessing Statdisk

- Statdisk is available online as a free browser-based application. Visit www.Statdisk.com to create an account and begin using Statdisk.
- No download or install is required to use Statdisk.
- Statdisk can be used with device running a modern web browser, including laptops (Windows, macOS), Chromebooks, tablets and smartphones.

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 an important component of technology 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 technology.
- 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 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.

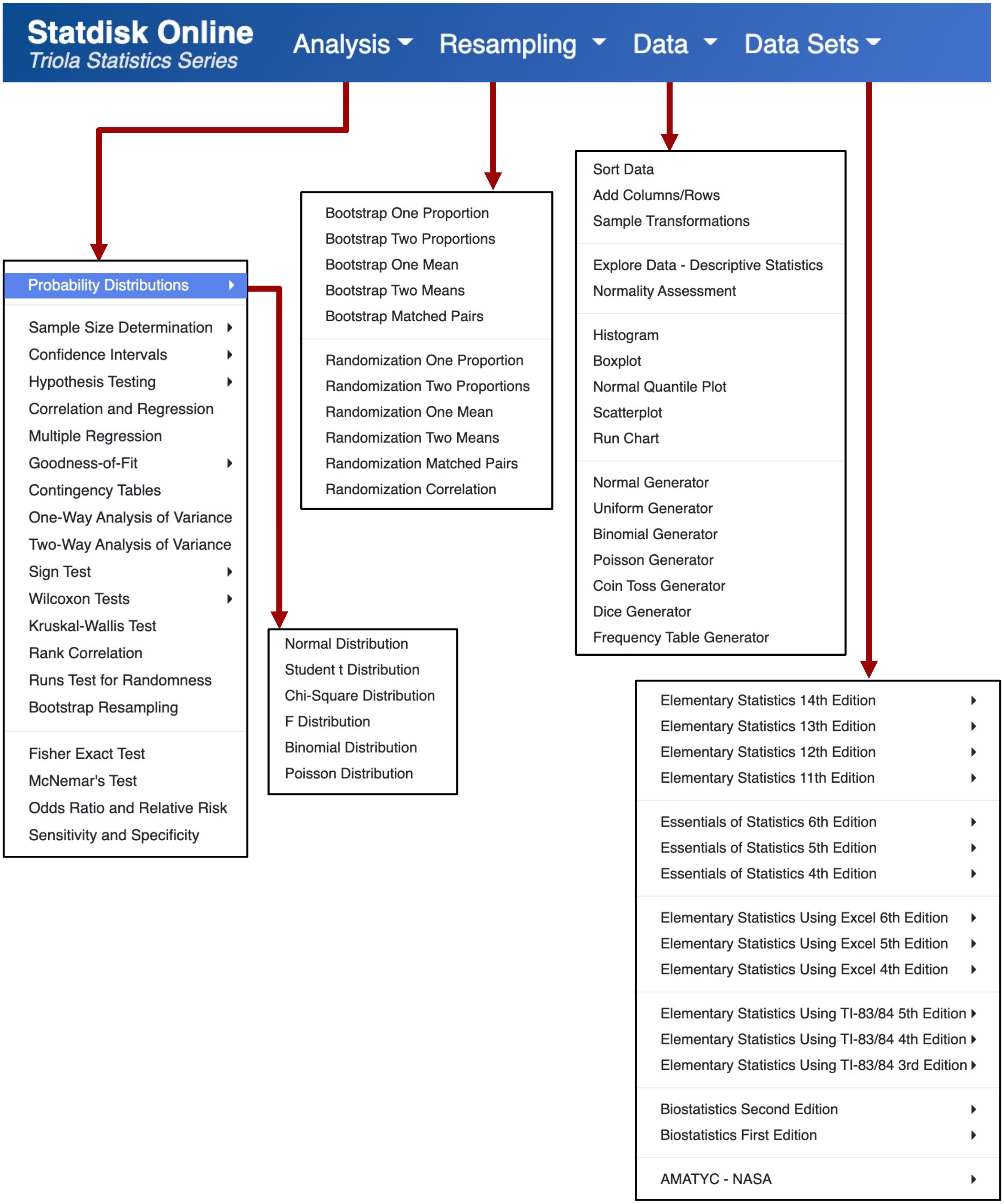
Format Chapter 1 of this supplement describes some important basics for using Statdisk. Chapters 2-14 in the manual/workbook correspond to Chapters 2-14 in *Elementary Statistics*, 14th Edition. However, individual chapter *sections* in this manual/workbook generally *do not* match the sections of the 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 examples, which 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 results to the results 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 14th 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 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, 2021

Statdisk Menu Configuration



Statdisk Online
Triola Statistics Series

Analysis

- Probability Distributions
  - Normal Distribution
  - Student t Distribution
  - Chi-Square Distribution
  - F Distribution
  - Binomial Distribution
  - Poisson Distribution
- 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 for Randomness
- Bootstrap Resampling
- Fisher Exact Test
- McNemar's Test
- Odds Ratio and Relative Risk
- Sensitivity and Specificity

Resampling

- Bootstrap One Proportion
- Bootstrap Two Proportions
- Bootstrap One Mean
- Bootstrap Two Means
- Bootstrap Matched Pairs
- Randomization One Proportion
- Randomization Two Proportions
- Randomization One Mean
- Randomization Two Means
- Randomization Matched Pairs
- Randomization Correlation

Data

- Sort Data
- Add Columns/Rows
- Sample Transformations
- Explore Data - Descriptive Statistics
- Normality Assessment
- Histogram
- Boxplot
- Normal Quantile Plot
- Scatterplot
- Run Chart
- Normal Generator
- Uniform Generator
- Binomial Generator
- Poisson Generator
- Coin Toss Generator
- Dice Generator
- Frequency Table Generator

Data Sets

- Elementary Statistics 14th Edition
- Elementary Statistics 13th Edition
- Elementary Statistics 12th Edition
- Elementary Statistics 11th Edition
- Essentials of Statistics 6th Edition
- Essentials of Statistics 5th Edition
- Essentials of Statistics 4th Edition
- Elementary Statistics Using Excel 6th Edition
- Elementary Statistics Using Excel 5th Edition
- Elementary Statistics Using Excel 4th Edition
- Elementary Statistics Using TI-83/84 5th Edition
- Elementary Statistics Using TI-83/84 4th Edition
- Elementary Statistics Using TI-83/84 3rd Edition
- Biostatistics Second Edition
- Biostatistics First Edition
- AMATYC - NASA

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 data tools (e.g. **Sort**, **Cut**, **Copy** and **Paste** data 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 Accessing Statdisk Online

Statdisk is a full featured statistical analysis package. It includes over 70 functions and tests, dozens of built-in datasets, and graphing. Statdisk is free to users of any Triola Statistics Series textbooks.

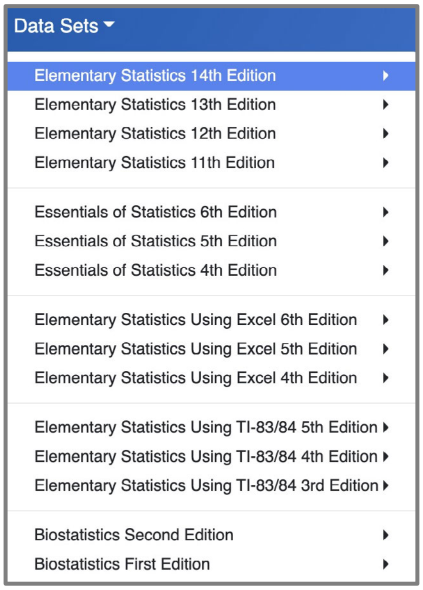
Statdisk is available online as a browser-based application. Visit www.Statdisk.com to create an account and begin using Statdisk.

Statdisk is simple to use on a wide variety of devices:

- No download or install is required to use Statdisk.
- Statdisk can be used with any device running a modern web browser, including laptops (Windows, macOS), Chromebooks, tablets and smartphones.

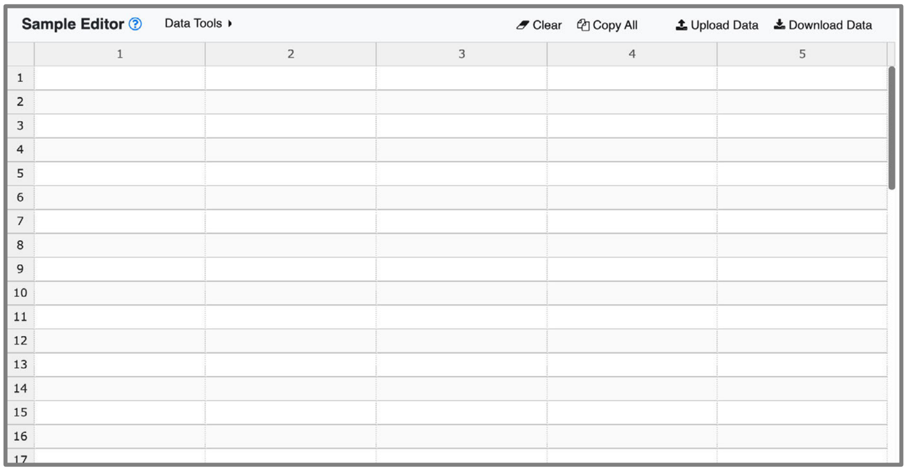
1-2 Entering Data

Your first use of Statdisk is likely to occur with topics from Chapter 2 of your Triola Statistics Series 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 textbooks are already available in Statdisk, and they can be opened by clicking **Data Sets** in the top menu. It is not necessary to manually enter these Appendix B data sets. See display below.) To manually enter a set of sample data, use the procedure on the next page.


Data Sets
Elementary Statistics 14th Edition
Elementary Statistics 13th Edition
Elementary Statistics 12th Edition
Elementary Statistics 11th Edition
Essentials of Statistics 6th Edition
Essentials of Statistics 5th Edition
Essentials of Statistics 4th Edition
Elementary Statistics Using Excel 6th Edition
Elementary Statistics Using Excel 5th Edition
Elementary Statistics Using Excel 4th Edition
Elementary Statistics Using TI-83/84 5th Edition
Elementary Statistics Using TI-83/84 4th Edition
Elementary Statistics Using TI-83/84 3rd Edition
Biostatistics Second Edition
Biostatistics First Edition

Statdisk Procedure for Entering Data

1. When logging into Statdisk, the Sample Editor appears as shown.

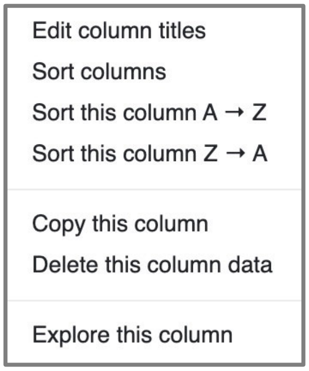


2. Click the cell in row 1 of column 1 and type your first data value, and then press the **Enter** key. Next, type the second data value and press the **Enter** key again. Continue to enter all of your sample values. You can use your mouse or the keyboard arrow keys to navigate to, and select individual cells in the Sample Editor.

   Note: If you see that you have made a mistake, simply click 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 a column name (e.g. column "*1*") you can select from the following options:

Edit column titles: Allows you to enter or modify the *names* of the columns of data



Sort columns: Allows you to sort data in one or all columns in ascending or descending order. See Section 2-5 for detailed sorting instructions.

Sort this column A→Z & Sort this column Z→A: Sorts the values only in the selected column in ascending or descending order.

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

Delete this column data: Deletes the column of data.

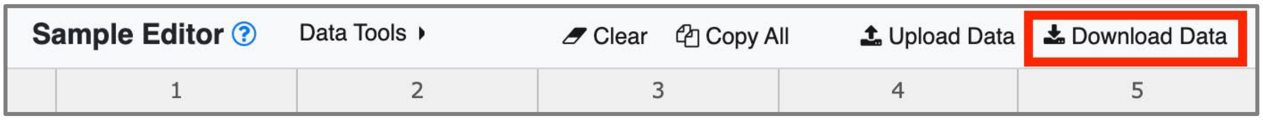
Explore this column: Allows you to obtain one screen that displays key statistics and graphs for data in a specific column.

1-3 Downloading and Saving Data

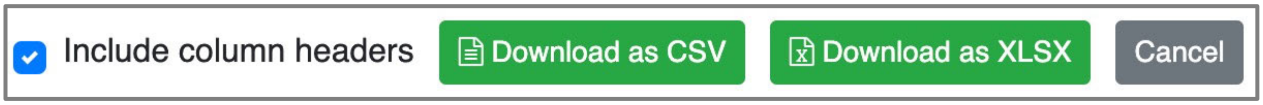
After entering a data set as described on the previous page, you can download and save these data to your computer for future use using the following procedure.

Statdisk Procedure for Downloading and Saving Data

1. After entering all of the values in the Sample Editor, select **Download Data** in the top menu of the Sample Editor.



2. The download toolbar will appear as shown below. Select whether or not you want to include column headers in the data download, and select the desired data format (.csv or .xlsx).

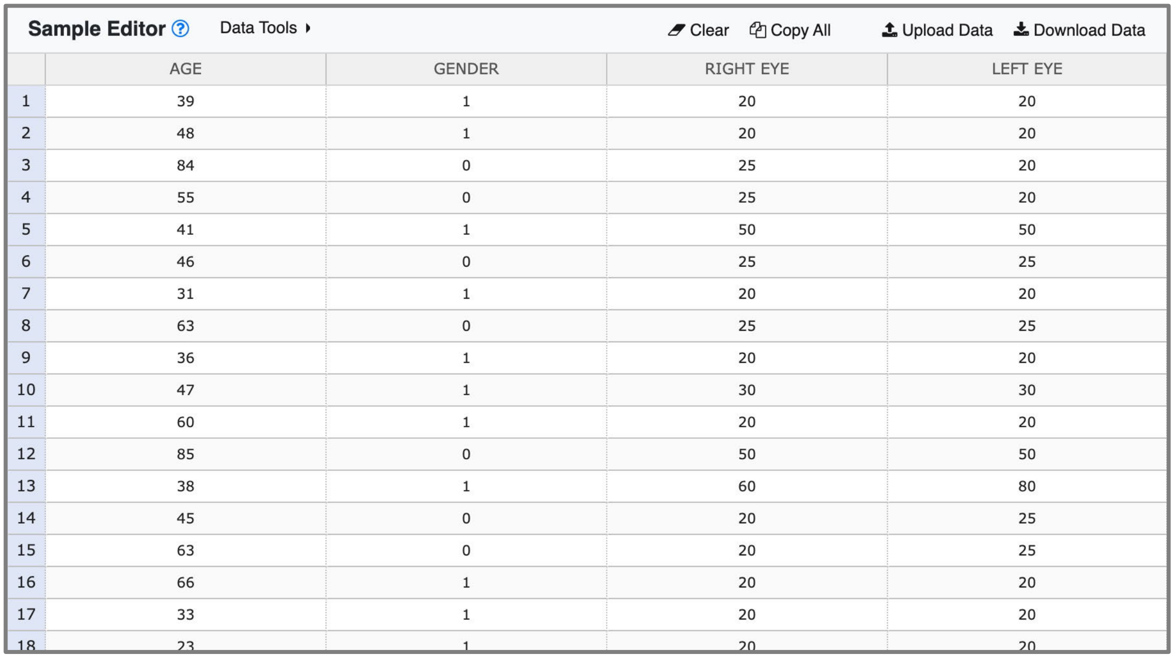


The data will be downloaded by your web browser and saved in the default download location for that browser. (In Google Chrome, you can view or change the download location by selecting **Settings-Advanced-Downloads**.)

1-4 Opening Saved Data

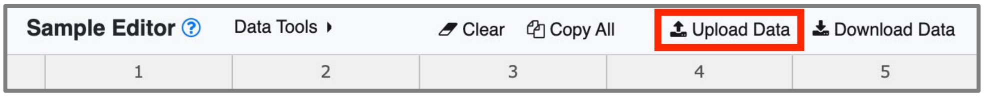
Appendix B Data sets: Statdisk includes all Appendix B data sets from the Triola Statistics Series. Many workbook exercises in the textbook require that you use these data sets. To retrieve one of the stored Appendix B data sets, follow these steps.

1. Click **Data Sets** in the top menu.
2. Select the Triola textbook that you are using.
3. Click the name of the desired data set to open. For example, to select Data Set 8 “Vision” from *Elementary Statistics* 14th Edition, scroll down to the menu item of **08 - Vision**, and then click that name. The data set will be inserted in the Sample Editor, as shown below. The full data set has 300 rows, and you can see the additional rows by scrolling down in the Sample Editor.



	AGE	GENDER	RIGHT EYE	LEFT EYE
1	39	1	20	20
2	48	1	20	20
3	84	0	25	20
4	55	0	25	20
5	41	1	50	50
6	46	0	25	25
7	31	1	20	20
8	63	0	25	25
9	36	1	20	20
10	47	1	30	30
11	60	1	20	20
12	85	0	50	50
13	38	1	60	80
14	45	0	20	25
15	63	0	20	25
16	66	1	20	20
17	33	1	20	20
18	23	1	20	20

Retrieving Your Own Dataset: To open a data set that has been previously saved on your computer, click **Upload Data** in the Sample Editor menu, and then browse and select the file containing the data to be uploaded into Statdisk. A preview of the data in the Sample Editor will be shown. Click the **Proceed with loading data** button to load the data into the Sample Editor.



Note: Uploaded data are only loaded into your browser's current Sample Editor. They are not sent to Statdisk nor permanently stored.

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 other content.**

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

See Section 1-8 of this manual 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.

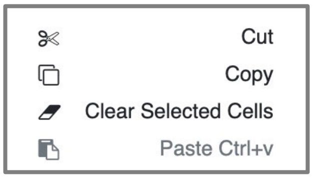
Statdisk Procedure for Using Copy and Paste

There are two ways to copy data: selected data can be copied or all data in the Sample Editor can be copied. See below for more detail on these two options. Note that only data can be copied in Statdisk, column names cannot be copied.

Copy Selected Data– Copies only the data you select.

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

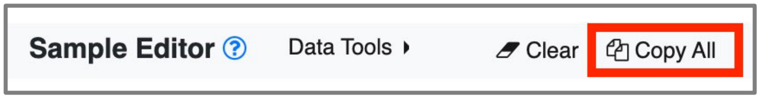
2. To copy the selected data, use **Control+c** on Windows or **Command+c** on macOS; or right-click and use the **Copy** item from the menu (shown on right).



3. To paste the copied data in Statdisk, click the first cell in the column where you want to insert the data. Next, use **Control+v** on Windows or **Command+v** on macOS to paste the data into the Sample Editor.

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

1. Select **Copy All** in the Sample Editor menu bar. All data in the Sample Editor is copied.



2. To paste the copied data, click the first cell in the column where you want to insert the data. Next, use **Control+v** in Windows or **Command+v** in macOS to paste the data into the Sample Editor.

1-6 Editing and Transforming Data

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

- *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.
- *Sort* data in ascending or descending order by clicking on **Data Tools** in the Sample Editor menu bar and selecting **Sort Data**. See Section 2-5 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 Temperatures* data set in Appendix B of the textbook) and you want to transform the values to the Celsius scale, you can use the equation:

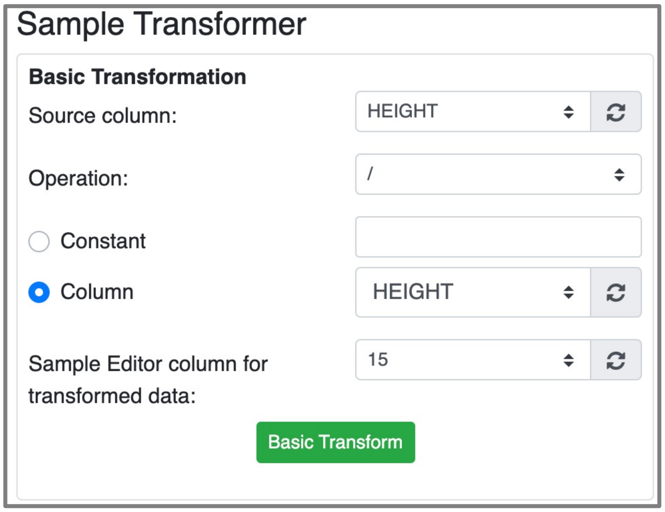
$$C = \frac{5}{9}(F - 32)$$

Statdisk Procedure for Transforming Data

1. Enter the data in one or more columns of the Sample Editor.
2. Click the top menu item of **Data**.
3. Click **Sample Transformations** from the dropdown menu and the *Sample Transformer* dialog box appears in Statdisk..

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. Select the column in which to place the transformed data.
8. Click **Basic Transform** and the transformed values will appear in the selected column. *Note*: *Any existing data in the selected column will be overwritten.*



Sample Transformer

Basic Transformation

Source column: HEIGHT

Operation: /

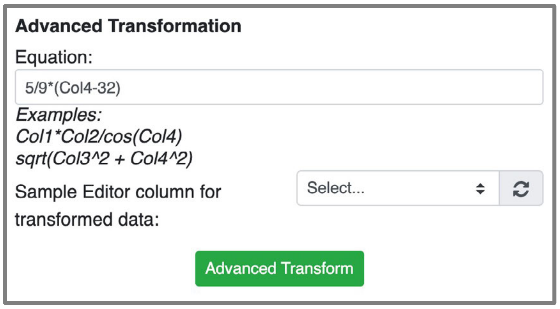
○ Constant

● Column HEIGHT

Sample Editor column for transformed data: 15

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 Fahrenheit to Celsius, enter the following equation in the **Equation** box, select the column for the transformed data, and then click the **Advanced Transform** button.



Advanced Transformation

Equation:

5/9*(Col4-32)

Examples:
*Col1*Col2/cos(Col4)*
sqrt(Col3^2 + Col4^2)

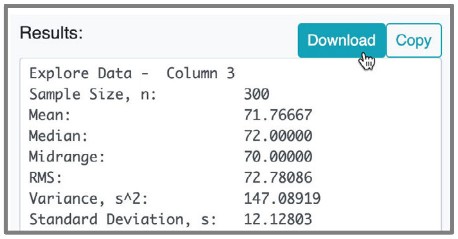
Sample Editor column for transformed data: Select...

Note: *For advanced transformation equations, use the column number as shown (e.g. Col4) instead of the column name.*

1-7 Downloading Results

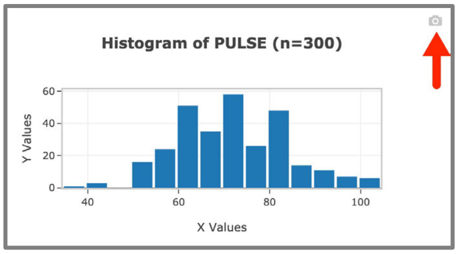
After you have obtained results from Statdisk, such as a graph or a listing of statistics, you can save those results to your computer.

Numerical Results Numerical and text results can be downloaded as a text file by clicking the Download button above the results box.



```
Explore Data -  Column 3
Sample Size, n:           300
Mean:                     71.76667
Median:                   72.00000
Midrange:                 70.00000
RMS:                      72.78086
Variance, s^2:            147.08919
Standard Deviation, s:    12.12803
```

Graphs Graph images can be downloaded as a .png file by clicking the camera icon in the upper right corner of the graph.



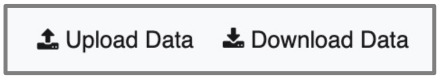
1-8 Exchanging Data with Other Applications

There may be times when you want to move data from Statdisk to another application (such as Excel or Minitab or Word) or move data from another application into 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.

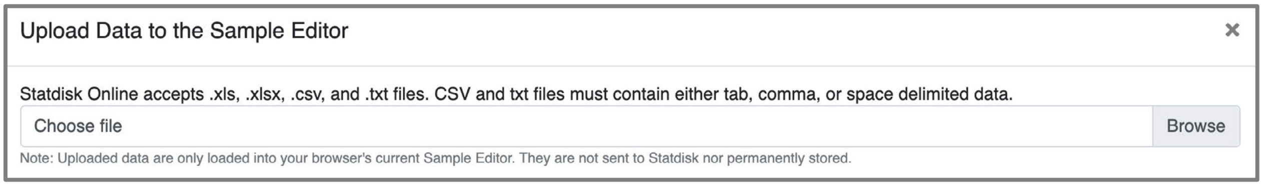
Method 1: Use Copy and Paste

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

Method 2: Upload/Download Data



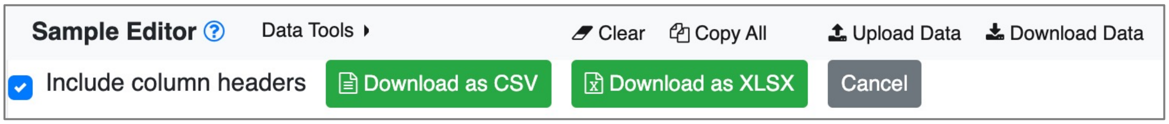
Upload Data to Statdisk: To upload a data file to Statdisk, click the **Upload Data** button in the Sample Editor menu bar and choose the file to upload from your computer. (Statdisk accepts .xls, .xlsx, .csv, and .txt files.)


Upload Data to the Sample Editor

Statdisk Online accepts .xls, .xlsx, .csv, and .txt files. CSV and txt files must contain either tab, comma, or space delimited data.

Note: Uploaded data are only loaded into your browser's current Sample Editor. They are not sent to Statdisk nor permanently stored.

Download Data from Statdisk: Data in the Sample Editor can be downloaded to your computer for use in other programs. Click the Download Data button and then select the desired file format (.csv or .xlsx). The Statdisk data file will be downloaded to your web browser's default download folder.



CHAPTER 1 WORKBOOK: 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 exercise, we will enter a small data set, save it, and retrieve it. *Elementary Statistics* 14th Edition, Data Set 5 “Body Temperatures” includes these body temperatures, along with others:

98.6 98.6 98.0 98.0 99.0

a. Access Statdisk and enter the above sample temperatures. (See the procedure described in Section 1-2 of this manual/workbook.)

b. Save the data set to your computer.

c. Upload the saved data set back into Statdisk.

1-2. ***Editing Data*** *Elementary Statistics* 14th Edition, Data Set 18 “Bear Measurements” includes measurements from 54 wild bears.

a. Open the data set “Bear Measurements”, and then find the value of the *mean* of the weights by selecting **Data,** then **Explore Data - Descriptive Statistics**. Enter the mean weight. _________

b. Go back to the Sample Editor and change the weight of 34 lb (row 12) 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?

____

1-3. ***Retrieving and Transforming Data*** Open *Elementary Statistics* 14th Edition, Data Set 18 “Bear Measurements”, which includes the weights (in pounds) of a sample of bears. To convert the weights to kilograms, multiply the weights 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.

____

1-4. ***Copy Data Between Applications*** Open *Elementary Statistics* 14th Edition, Data Set 8 “Vision” and complete the following:

a. Copy the data in the column “Right Eye”.

b. Paste the copied data into a spreadsheet in Excel, StatCrunch, or Google Sheets.

c. Clear the Statdisk Sample Editor.

d. Copy the “Right Eye” data in Excel/StatCrunch/Google Sheets and paste it back into Statdisk.

2 Exploring Data With Tables and Graphs

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 manual. Be sure to understand these functions before beginning this chapter.

Section 2-1 in the textbook describes the construction of a table representing the *frequency distribution* for a set of data. Shown below is Table 2-1, which contains 50 reported daily commute times (minutes) in Los Angeles. (The listed commute times are part of *Elementary Statistics* 14th Edition, Data Set 31 "Commute Times.")

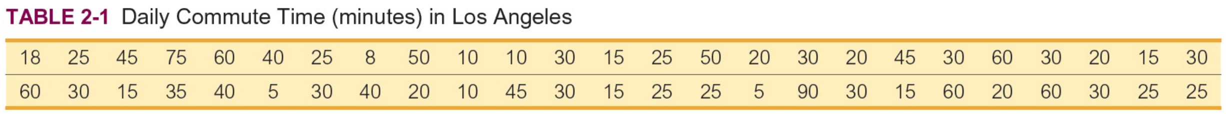


TABLE 2-1 Daily Commute Time (minutes) in Los Angeles

18	25	45	75	60	40	25	8	50	10	10	30	15	25	50	20	30	20	45	30	60	30	20	15	30
60	30	15	35	40	5	30	40	20	10	45	30	15	25	25	5	90	30	15	60	20	60	30	25	25

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

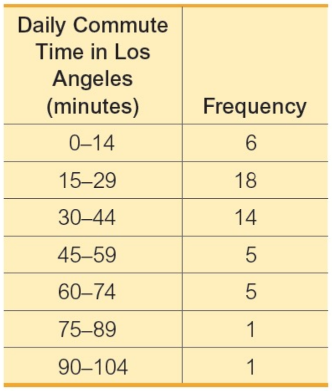
2-1 Histograms

Statdisk can be used to automatically generate a histogram. The basic approach is to enter the data in the Sample Editor, and then use the *Histogram* function 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) in a frequency distribution.

As an example, see Table 2-2 on the next page. Table 2-2 is a frequency distribution summarizing the listed Los Angeles commute times from Table 2-1. The *class width is 15* (the difference between the consecutive lower class limits of 0 and 15).

TABLE 2-2 Daily Commute Time in Los Angeles



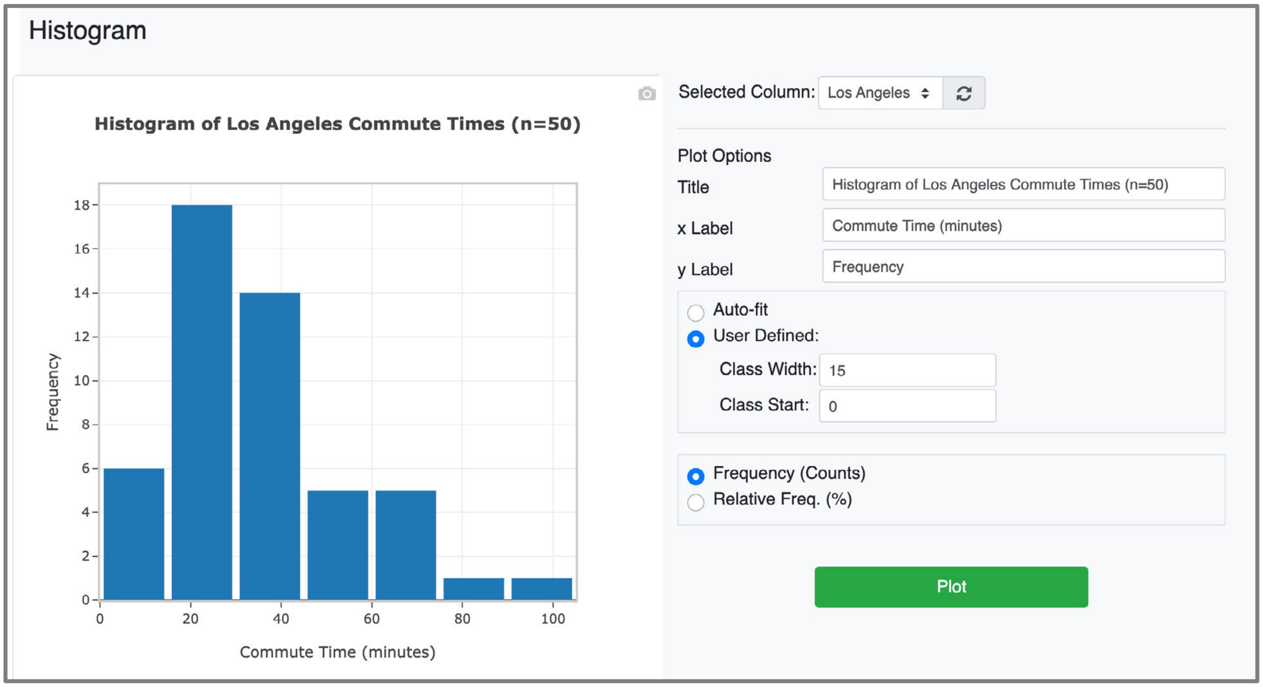
Daily Commute Time in Los Angeles (minutes)	Frequency
0–14	6
15–29	18
30–44	14
45–59	5
60–74	5
75–89	1
90–104	1

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.
   - **Retrieve a data set from those included in Appendix B:** Click the top menu item of **Data Sets** and proceed to select one of the listed textbooks and data sets.
   - **Retrieve a data set that you created:** Use **Upload Data** as described in Section 1-4 of this manual/workbook.
2. Click **Data** in the top menu bar.
3. Click **Histogram** and the *Histogram* dialog box appears in Statdisk.
4. In the *Histogram* dialog box, first select the column to be used. The default is column 1, and it can be changed to any column.
5. **Using the default settings**: Click **Plot** to allow Statdisk to automatically generate a histogram using default settings.

   Using your own settings: Select **User Defined** to the right of the Histogram. Proceed to enter the desired class width and the starting value of the first class.
6. Click **Plot** to update the histogram.

As an example, see the Statdisk display on the following page. The histogram is constructed using the list of Los Angeles commute times from Table 2-1. Instead of using the default settings, **User defined** is selected and we have entered 15 for the *Class Width* and 0 as the *Class Start*. 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 15 and the lower limit of the first class is 0.)



Histogram

Selected Column: Los Angeles

Plot Options

Title: Histogram of Los Angeles Commute Times (n=50)

x Label: Commute Time (minutes)

y Label: Frequency

○ Auto-fit

● User Defined:

Class Width: 15

Class Start: 0

● Frequency (Counts)

○ Relative Freq. (%)

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.

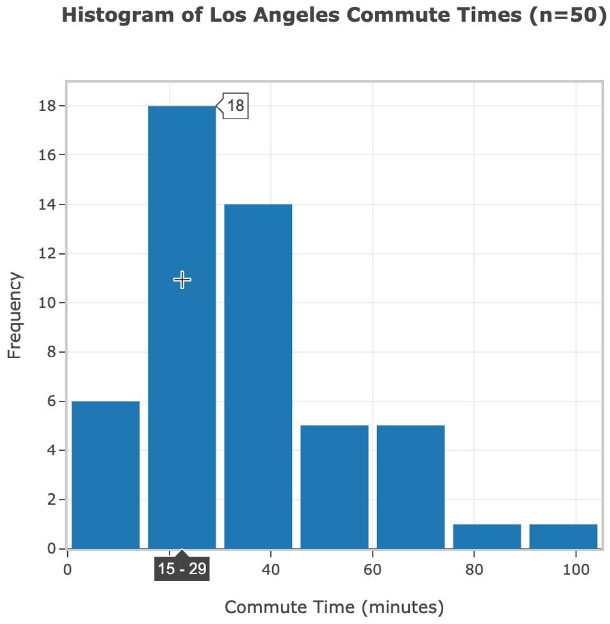
If we examine the Statdisk histogram shown above, we can see that the distribution does *not* appear to be bell–shaped, so the requirement of a normal distribution does *not* appear to be satisfied for this data set. (Statdisk includes a feature of *Assessing Normality* that provides more information. 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 class starting point and class width (based on the range of values and the minimum value).

Moving the cursor over the histogram will display the lower class limit, upper class limit and frequency for each bar in the histogram as shown below. We can see that the second class is 15-29 and it has a frequency of 18. Moving the cursor left/right will reveal the class limits and frequency count for each class. 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.



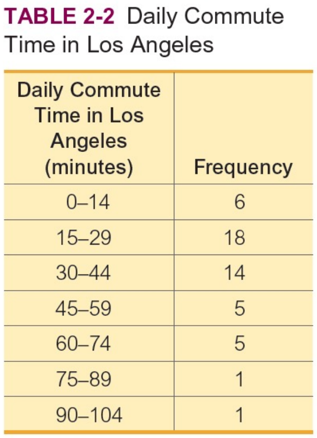


TABLE 2-2 Daily Commute Time in Los Angeles

Daily Commute Time in Los Angeles (minutes)	Frequency
0–14	6
15–29	18
30–44	14
45–59	5
60–74	5
75–89	1
90–104	1

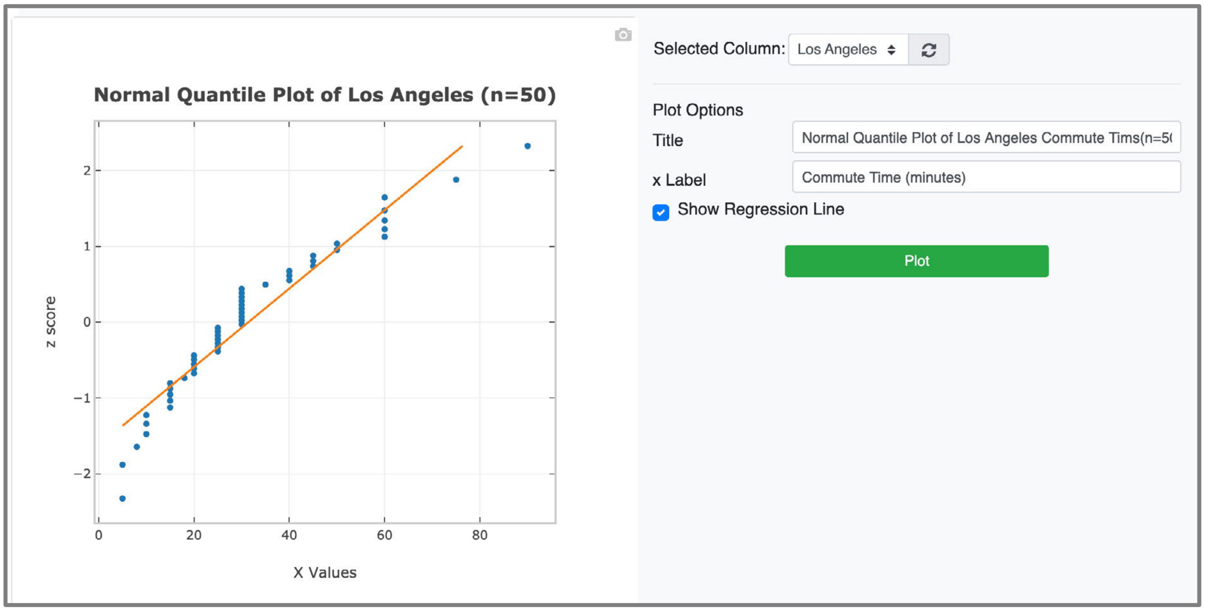
2-3 Normal Quantile Plots

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. Statdisk generates normal quantile plots like the one shown below. This normal quantile plot was generated using the same 50 Los Angeles commute times used to generate the histogram and frequency distributions earlier in this chapter.

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.
   - **Retrieve a data set from those included in Appendix B:** Click the top menu item of **Data Sets** and proceed to select one of the listed textbooks and data sets. **Retrieve a data set that you created:** Use **Upload Data** as described in Section 1-4 of this manual/workbook.
2. Click **Data** in the top menu bar.
3. Click **Normal Quantile Plot** in the dropdown menu. The *Normal Quantile Plot* dialog box appears in Statdisk.
4. Select the data column to be used for the normal quantile plot.
5. Click the **Plot** button.
   - If the *Show Regression Line* box is checkmarked (✓), the graph will include a straight line that best fits the points.

As an example, see the following Statdisk display. The normal quantile plot is constructed using the 50 Los Angeles commute times from Table 2-1 near the beginning of this chapter.



Selected Column: Los Angeles

Plot Options

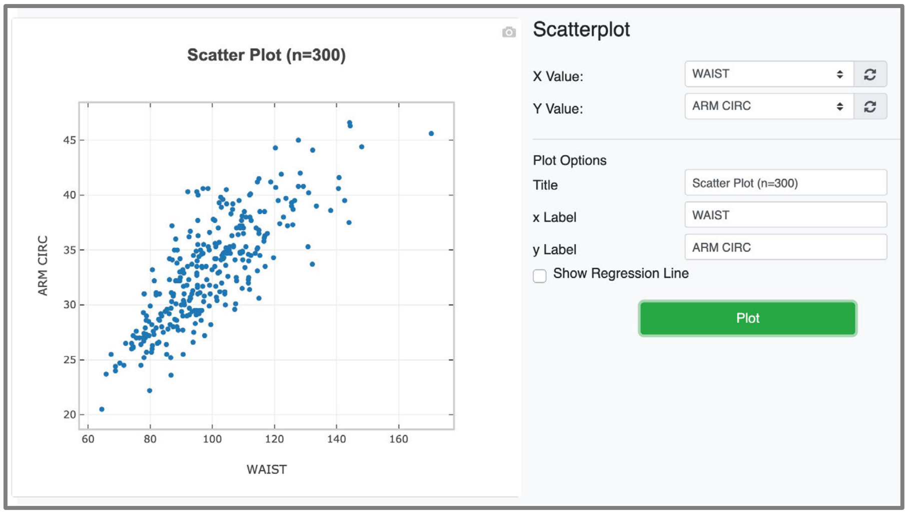
Title: Normal Quantile Plot of Los Angeles Commute Tims(n=5(

x Label: Commute Time (minutes)

☑ Show Regression Line

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 Statdisk scatterplot shown below results from paired data consisting of waist circumference (cm) and arm circumferences (cm) of randomly selected adults. This data set is available in Statdisk. (Click **Data Set** in the top menu bar and select *Elementary Statistics* 14th Edition - “01 - Body Data”.) This scatterplot shows that as the waist circumference (*x* axis) increases, the corresponding arm circumference (*y* axis) tends to be higher.



Scatterplot

X Value: WAIST

Y Value: ARM CIRC

Plot Options

Title: Scatter Plot (n=300)

x Label: WAIST

y Label: ARM CIRC

□ Show Regression Line

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.

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.
   - **Retrieve a data set from those included in Appendix B:** Click the top menu item of **Data Sets** and proceed to select one of the listed textbooks and data sets.
   - **Retrieve a data set that you created:** Use **File** - **Open** as described in Section 1-4 of this manual/workbook.

2. Click **Data** in the top menu bar.

3. Click **Scatterplot** in the dropdown menu.

4. Select the two columns to be used for the scatterplot.

5. Click the **Plot** button.
   - If the *Show Regression Line* box is checkmarked (✓), the graph will include a straight line that best fits the points. In the Scatterplot display shown above, the box is not checked so the line is not included.

2-5 Sorting Data

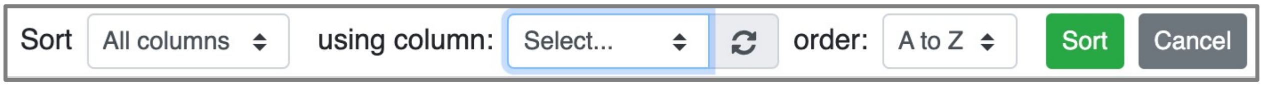
To *sort* data is to arrange them in ascending or descending order. There are several cases in which it becomes necessary to rearrange a data set so that the values are in order. Statdisk features a sorting tool that makes sorting data fast and easy.

Procedure for Sorting Data

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

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

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



4. Select **Sort - All columns** to sort based on a single column while rearranging the other columns so that data in the same row remain in the same row.

   Select **Sort - One column** to 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 WORKBOOK: Graphing Data

Histograms. *In exercises 2-1 through 2-4, use the* Elementary Statistics *14th 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. **Weights** Use the weights (WEIGHT) contained in Data Set 1 "Body Data." Construct a histogram. Does the histogram appear to depict data having a normal distribution? Why or why not?

2-2. **Earthquake Magnitudes** Use the earthquake magnitudes contained in Data Set 24 "Earthquakes." 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?

2-3. **Earthquake Depths** Use the earthquake depths contained in Data Set 24 "Earthquakes." 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?

2-4. **Red Blood Cell Counts** Use the red blood cell counts (RED) contained in Data Set 1 "Body Data." 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?

Scatterplots*.* *In exercises 2-5 through 2-7, use the given paired data from the* Elementary Statistics *14th 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.*

2-5. **President's Heights** Refer to Data Set 22 "Presidents" 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.)

2-6. **Brain Volume and IQ** Refer to Data Set 12 "IQ and Brain Size" and use the brain volumes (cm^3) 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?

2-7. **Bear Chest Size and Weight** Refer to Data Set 18 "Bear Measurements" and use the measured chest sizes and weights of bears. Does there appear to be a correlation between those two variables?

3 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.

3-1 Measures of Center and Variation

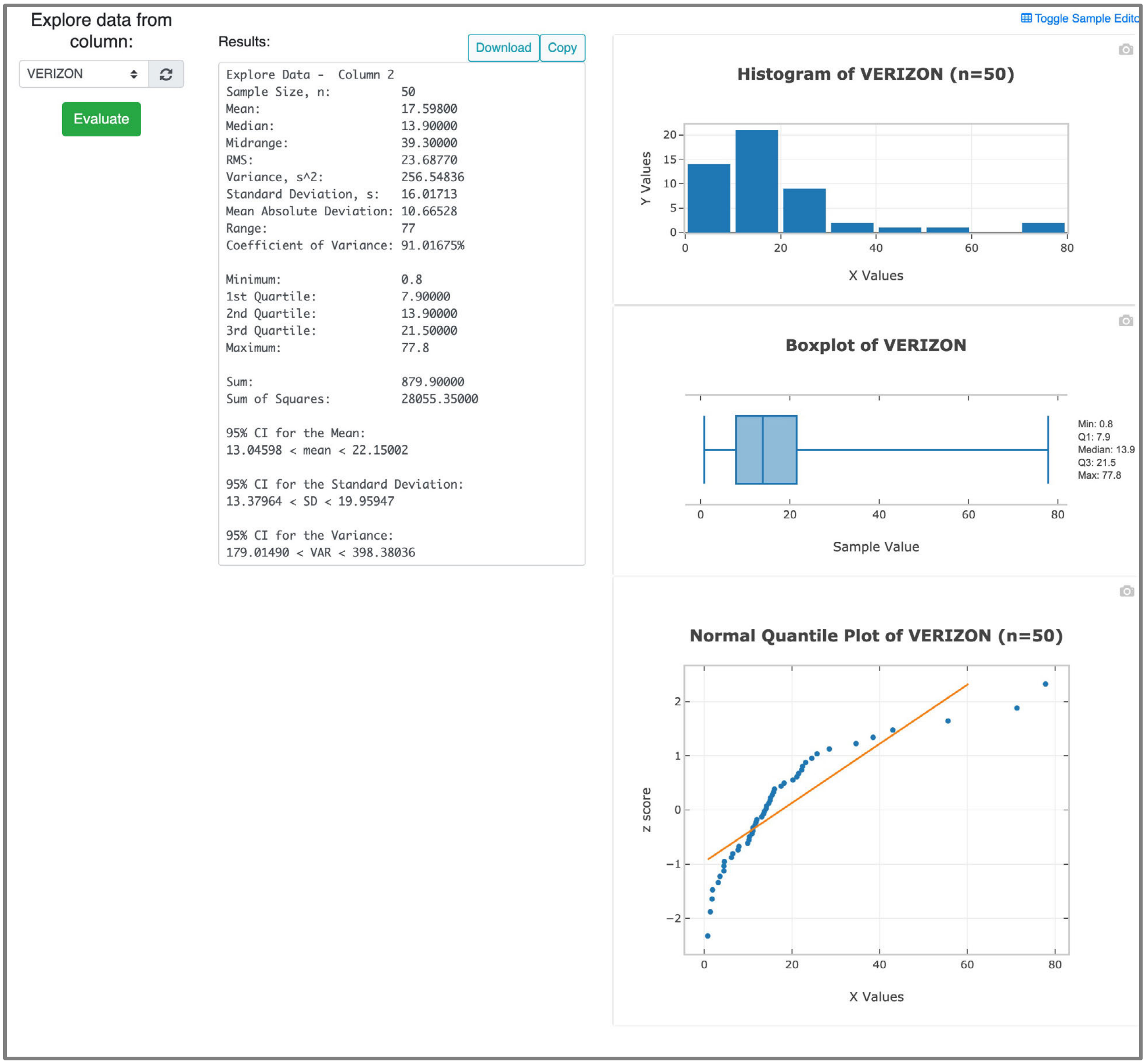
Important measures of center and variation can be obtained by using Statdisk's **Explore Data – Descriptive Statistics** function. To explore a list of data, follow the procedure below.

Procedure for Exploring Data and Obtaining Descriptive Statistics

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.
   - **Retrieve a data set from those included in Appendix B:** Click the top menu item of **Data Sets** and proceed to select one of the listed textbooks and data sets.
   - **Retrieve a data set that you created:** Use **File** - **Open** as described in Section 1-4 of this manual/workbook.
2. Click **Data** in the top menu bar.
3. Click **Explore Data – Descriptive Statistics** from the dropdown menu.
4. Select the column to be used for the calculations and graphs.
5. Click the **Evaluate** button.

As an example, consider the Verizon airport data speeds contained in Column 2 of *Elementary Statistics* 14th Edition, Data Set 34 “Airport Data Speeds.”

From the display on the next page we see that there are n = 50 sample values, the sample mean is 17.598 Mbps, the median is 13.9 Mbps, the midrange is 39.3 Mbps. The value of "RMS" is the value of the *root mean square* (or quadratic mean) described in the textbook. The variance is s^2 = 256.55 Mbps^2 (rounded), and the standard deviation is 16.02 (rounded). The value listed as "Mean Absolute Deviation" 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.



```
Explore Data -  Column 2
Sample Size, n:          50
Mean:                    17.59800
Median:                  13.90000
Midrange:                39.30000
RMS:                     23.68770
Variance, s^2:           256.54836
Standard Deviation, s:   16.01713
Mean Absolute Deviation: 10.66528
Range:                   77
Coefficient of Variance: 91.01675%

Minimum:                 0.8
1st Quartile:            7.90000
2nd Quartile:            13.90000
3rd Quartile:            21.50000
Maximum:                 77.8

Sum:                     879.90000
Sum of Squares:          28055.35000

95% CI for the Mean:
13.04598 < mean < 22.15002

95% CI for the Standard Deviation:
13.37964 < SD < 19.95947

95% CI for the Variance:
179.01490 < VAR < 398.38036
```

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 Statdisk results on the previous page. Here is the 5-number summary:

Minimum:	0.8 Mbps
1st Quartile Q_1:	7.9 Mbps
2nd Quartile Q_2:	13.9 Mbps
3rd Quartile Q_3:	21.5 Mbps
Maximum:	77.8 Mbps

Important Note: There is not universal agreement on a single procedure for calculating quartiles, and different computer programs might yield different results

3-3 Boxplots and Modified 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 – Descriptive Statistics* 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.

Procedure for Generating Boxplots and Modified Boxplots

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.
   - **Retrieve a data set from those included in Appendix B:** Click the top menu item of **Data Sets** and proceed to select one of the listed textbooks and data sets.
   - **Retrieve a data set that you created:** Use **Upload Data** as described in Section 1-4 of this manual/workbook.

2. Click **Data** in the top menu bar.

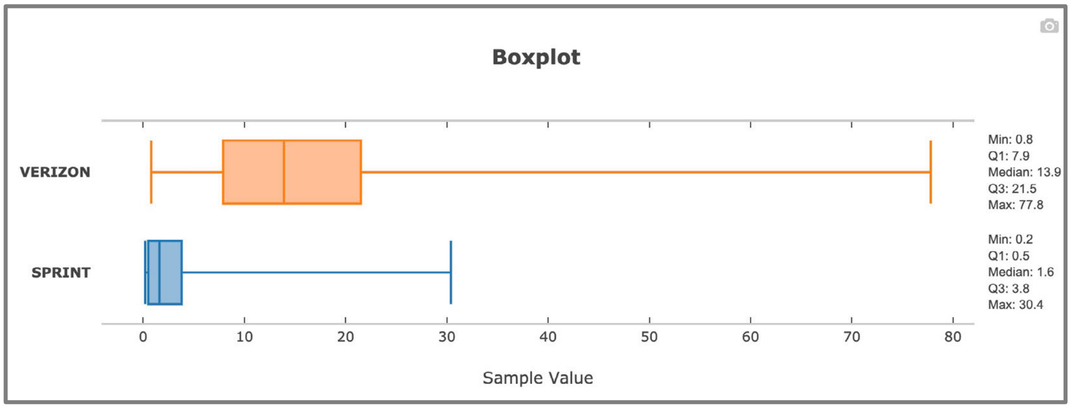
3. Click **Boxplot** in the dropdown menu.

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

5. Click the **Boxplot** button or **Modified Boxplot** button based on your preference.

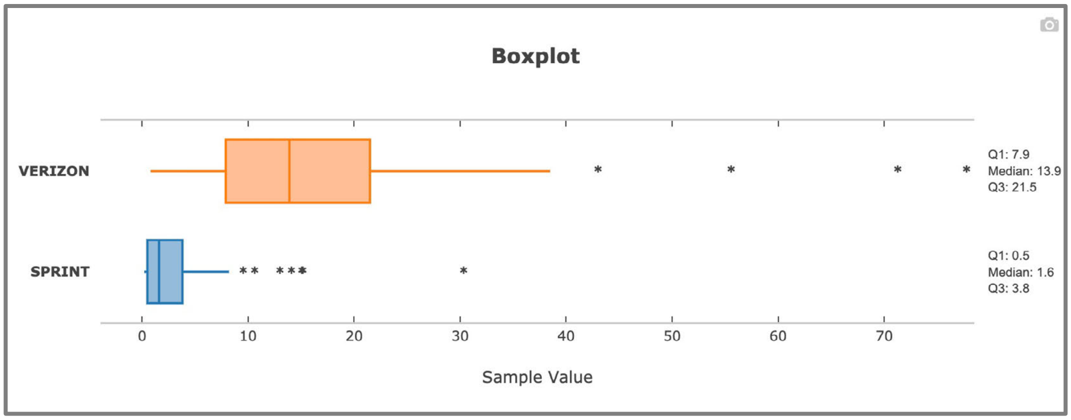
One important advantage of boxplots is that they are very useful in comparing data sets. Shown below is the Statdisk boxplot display showing the two boxplots representing the airport data speeds (Mbps) for Verizon and Sprint from *Elementary Statistics* 14th Edition, Data Sets 34 "Airport Data Speeds." Because the two boxplots are constructed on the same scale, a comparison becomes easier. The boxplots suggests that Verizon data speeds are generally faster.

Boxplot



Shown below is the Statdisk modified boxplot display.

Modified Boxplot



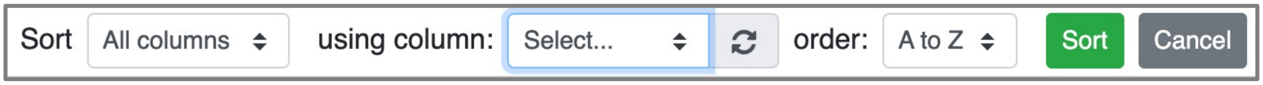
Important note: Statdisk generates boxplots based on the minimum, maximum, and three quartiles. Statdisk determines the values of the quartiles by following the same procedure described in the textbook, but other programs may use different procedures, so there may be some differences in boxplot results.

3-4 Outliers

Statdisk's sort feature is useful for identifying outliers and organizing data. When analyzing data, it is important to identify outliers because they can have a dramatic effect on many results. 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.

Procedure for Sorting Data

1. Enter or retrieve a set of sample data so that the sample values are listed in the Sample Editor.
2. Click **Data** in the top menu, or **Data Tools** in the Statdisk Sample Editor menu bar.
3. Click **Sort Data** in the dropdown menu and the following toolbar will appear above the Sample Editor:



4. Select **Sort - All columns** to sort based on a single column while rearranging the other columns so that data in the same row remain in the same row.

   Select **Sort - One column** to 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**.

3-5 Statistics from a Frequency Distribution

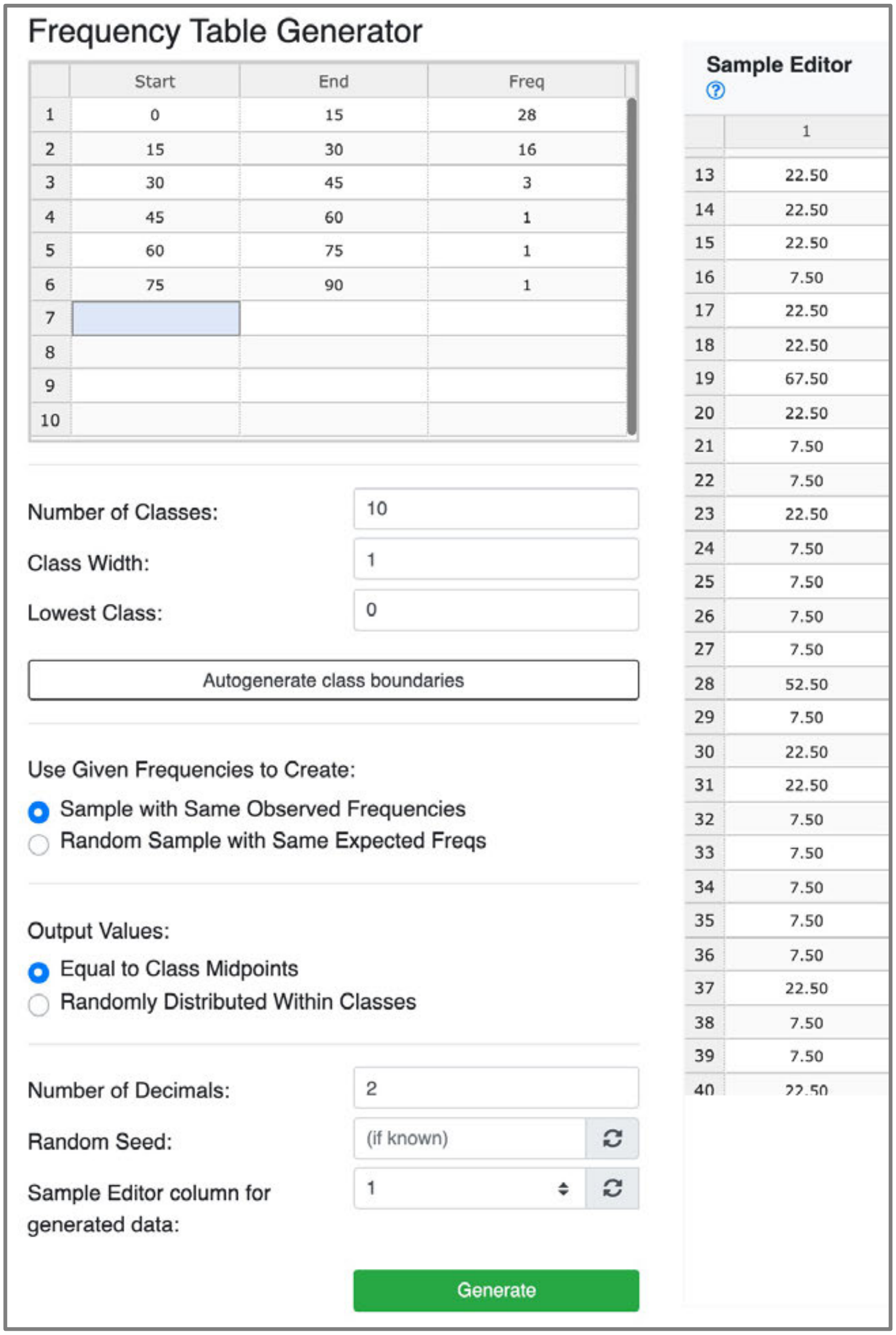
If sample data are summarized in the form of a table representing a frequency distribution such as the one shown on the next page, Statdisk can be used to obtain the important measures of center and variation. The basic idea is to use Statdisk's *Frequency Table Generator* to generate a list of sample values based on the table. Given the table on the next page, for example, Statdisk can generate a list of 100 values containing 56 values of 7.5 (the midpoint of the first class), 32 values of 22.5 (the midpoint of the second class), 6 values of 37.5 (the midpoint of the third class), and so on. The result will be a list of 100 sample values that correspond to the frequency distribution.

Verizon Airport Data Speed (Mbps)	Frequency
0-15	28
15-30	16
30-45	3
45-60	1
60-75	1
75-90	1

Procedure for Obtaining Statistics From a Frequency Distribution

1. First identify the frequency distribution to be used. (See the table above.)
2. Click **Data** in the top menu bar.
3. Click **Frequency Table Generator** in the dropdown menu.
4. See the Statdisk display on the next page showing the entries corresponding to the frequency distribution on the previous page.
   - 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.
   - Select the desired column in which to paste the generated data
5. Click the **Generate** button to get a list of sample values in the selected column.
6. Once the list of sample values is generated, **Data - Explore Data – Descriptive Statistics** can be used 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 Verizon airport data speeds, the mean is found to be 17.6 Mbps when rounded, but using the generated values from the frequency distribution results in a mean of 17.7 Mbps when rounded.



Frequency Table Generator

	Start	End	Freq
1	0	15	28
2	15	30	16
3	30	45	3
4	45	60	1
5	60	75	1
6	75	90	1
7			
8			
9			
10			

Number of Classes: 10

Class Width: 1

Lowest Class: 0

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

Number of Decimals: 2

Random Seed: (if known)

Sample Editor column for generated data: 1

Sample Editor

	1
13	22.50
14	22.50
15	22.50
16	7.50
17	22.50
18	22.50
19	67.50
20	22.50
21	7.50
22	7.50
23	22.50
24	7.50
25	7.50
26	7.50
27	7.50
28	52.50
29	7.50
30	22.50
31	22.50
32	7.50
33	7.50
34	7.50
35	7.50
36	7.50
37	22.50
38	7.50
39	7.50
40	22.50

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

3-1 **Comparing Heights of Men and Women** Use the sample data for all 300 adults included in *Elementary Statistics* 14th Edition, Data Set 1 "Body Data" included in Statdisk. Remember, instead of manually entering the 300 individual heights (which would be no fun at all), open the data set in Statdisk by selecting **Data Sets** from the top menu.

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.

____

____

3-2 **Boxplots** Use *Elementary Statistics* 14th Edition, Data Set 34 "Airport Data Speeds" and generate boxplots for the data speeds for Verizon, Sprint, AT&T and T-Mobile. Include all four boxplots in the same window so that they can be compared. Do the boxplots suggest any notable differences in the four sets of sample data?

____

____

Generate a modified boxplot for the Verizon data speeds and Interpret the *asterisks* that appears in the Statdisk display.

____

____

3–3 **Comparing Data** Open the Statdisk data set *Elementary Statistics* 14th Edition, Data Set 37 "Cola Weights and Volumes" and use Statdisk to compare the weights of regular Coke and the weights of diet Coke. Obtain relevant results. What do you conclude? Can you explain any substantial difference?

____

____

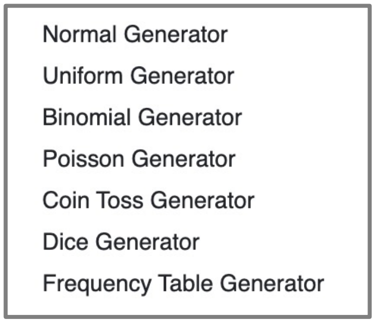
____

4 Probabilities Through Simulations

Statdisk

4-1 Statdisk Simulation Tools

Statdisk includes several different tools that can be used for simulations. Clicking **Data** in the top menu displays these items at the bottom of the dropdown menu:

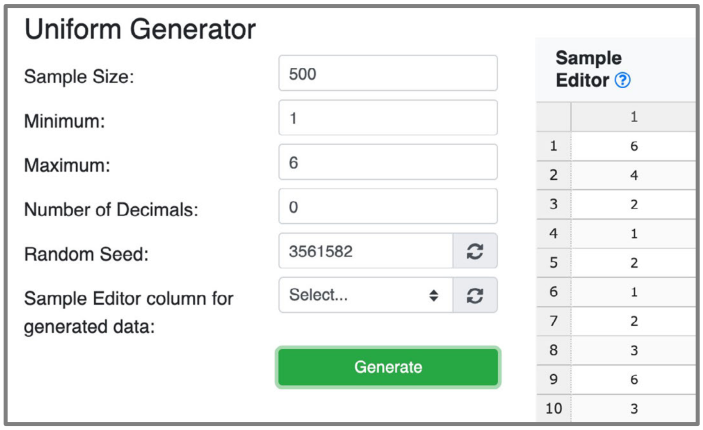


Normal Generator
Uniform Generator
Binomial Generator
Poisson Generator
Coin Toss Generator
Dice Generator
Frequency Table Generator

Here are descriptions of these menu items:

- **Normal Generator**: Generates a sample of data randomly selected from a population having a normal distribution. The desired sample size, population mean and standard deviation must be entered. 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. The number of decimal places can be specified, 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 entering 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 Statdisk display below. (See also the Dice Generator described on the next page.)



Uniform Generator

Sample Size: 500
Minimum: 1
Maximum: 6
Number of Decimals: 0
Random Seed: 3561582
Sample Editor column for generated data: Select...

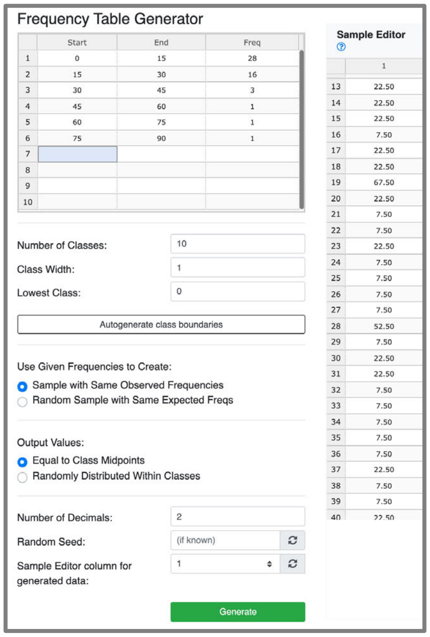
	1
1	6
2	4
3	2
4	1
5	2
6	1
7	2
8	3
9	6
10	3

- **Binomial Generator**: Generates numbers of successes for a binomial probability distribution. Enter 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. Enter the desired number of tosses (trials), and enter the number of coins to be tossed in each trial. The generated values are the numbers of heads that turn up.

- **Dice Generator:** Enter the desired sample size (trials), and enter 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 3-5 of this manual for more detail. The generated data can be copied to the Sample Editor where it can be used with other functions, such as *Explore Data - Descriptive Statistics* or *Histogram*.



Frequency Table Generator

	Start	End	Freq
1	0	15	28
2	15	30	16
3	30	45	3
4	45	60	1
5	60	75	1
6	75	90	1
7			
8			
9			
10			

Number of Classes: 10

Class Width: 1

Lowest Class: 0

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

Number of Decimals: 2

Random Seed: (if known)

Sample Editor column for generated data: 1

Sample Editor

	1
13	22.50
14	22.50
15	22.50
16	7.50
17	22.50
18	22.50
19	67.50
20	22.50
21	7.50
22	7.50
23	22.50
24	7.50
25	7.50
26	7.50
27	7.50
28	52.50
29	7.50
30	22.50
31	22.50
32	7.50
33	7.50
34	7.50
35	7.50
36	7.50
37	22.50
38	7.50
39	7.50
40	22.50

Random Seed: The preceding Statdisk tools include an option for entering a "random seed" if it is known. This entry will usually be 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.

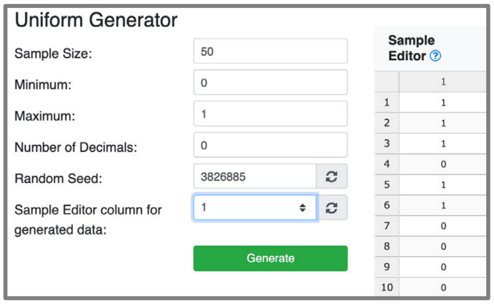
4-2 Simulation Examples

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 4-1) 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 2-5 of this manual/workbook for the procedure for sorting data.

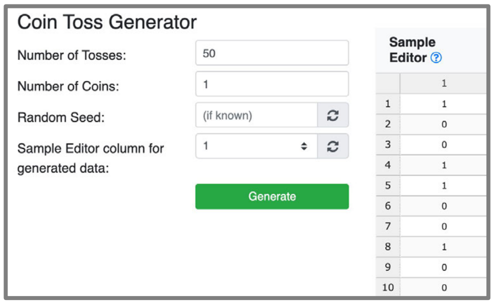


Uniform Generator

Sample Size: 50
Minimum: 0
Maximum: 1
Number of Decimals: 0
Random Seed: 3826885
Sample Editor column for generated data: 1

	1
1	1
2	1
3	1
4	0
5	1
6	1
7	0
8	0
9	0
10	0

- Use Statdisk's **Coin Toss Generator** (see Section 4-1). Enter 50 for the number of tosses 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 2-5 of this manual/workbook.



Coin Toss Generator

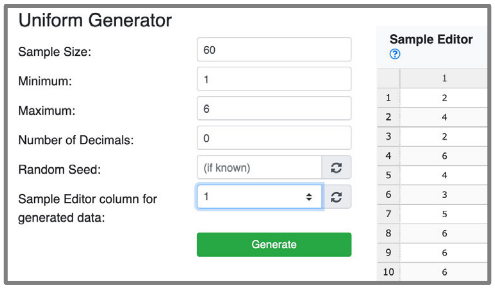
Number of Tosses: 50
Number of Coins: 1
Random Seed: (if known)
Sample Editor column for generated data: 1

	1
1	1
2	0
3	0
4	1
5	1
6	0
7	0
8	1
9	0
10	0

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 4-1) 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.

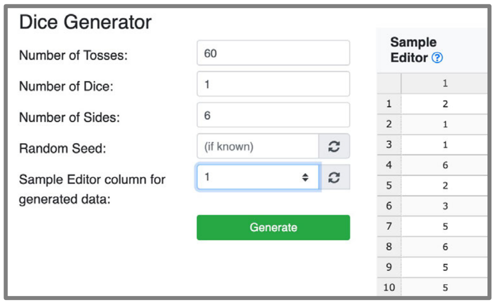


Uniform Generator

Sample Size: 60
Minimum: 1
Maximum: 6
Number of Decimals: 0
Random Seed: (if known)
Sample Editor column for generated data: 1

	1
1	2
2	4
3	2
4	6
5	4
6	3
7	5
8	6
9	6
10	6

- 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.



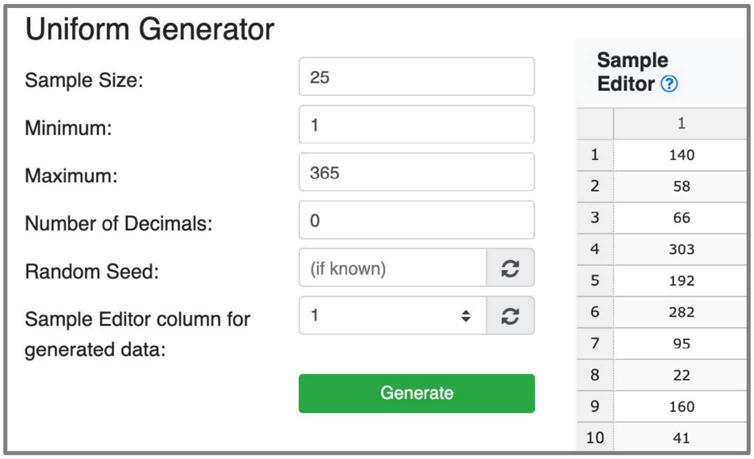
Dice Generator

Number of Tosses: 60
Number of Dice: 1
Number of Sides: 6
Random Seed: (if known)
Sample Editor column for generated data: 1

	1
1	2
2	1
3	1
4	6
5	2
6	3
7	5
8	6
9	5
10	5

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 Statdisk display shown here. The first generated value of 140 corresponds to May 20, because the 140th day in a year is May 20.



Uniform Generator

Sample Size: 25

Minimum: 1

Maximum: 365

Number of Decimals: 0

Random Seed: (if known)

Sample Editor column for generated data: 1

Sample Editor

	1
1	140
2	58
3	66
4	303
5	192
6	282
7	95
8	22
9	160
10	41

Even though the display shows only the first 10 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 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 WORKBOOK: 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, and 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 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.

____

____

4-4 **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, and 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? ___________

4-5 **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)? ___________

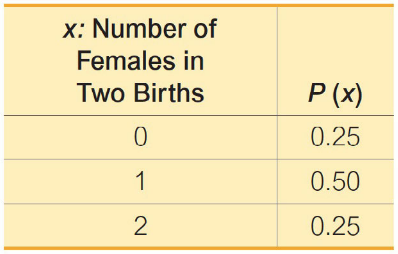
5 Discrete Probability Distributions

Statdisk

5-1 Exploring Probability Distributions

Although Statdisk is not designed to deal directly with a probability distribution, it can often be used. Consider Table 5-2 below. This table lists the probabilities for the number of females in two births.

TABLE 5-2 Probability Distribution for the Number of Females in Two Births



x: Number of Females in Two Births	*P* (*x*)
0	0.25
1	0.50
2	0.25

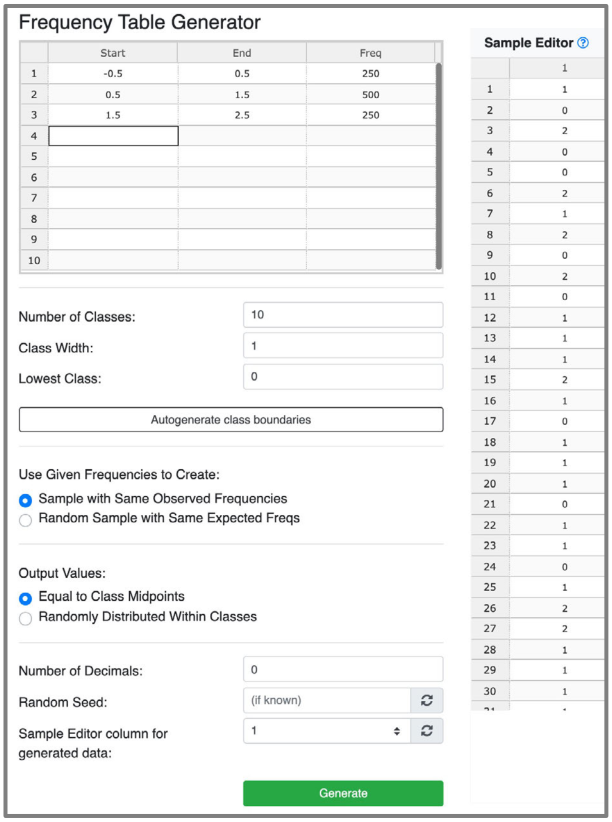
If you examine the data in the Table 5-2, you can verify that a probability distribution is defined because the three key requirements are satisfied:

1. The variable *x* is a numerical random variable and its values are associated with probabilities.
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 5-2 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 **Data** in the top menu.
2. Select **Frequency Table Generator** from the dropdown menu.
3. Enter class limits and frequencies that correspond to the probability distribution. A sample size of 1000 is used in this example. Shown on the next page are the entries corresponding to the probability distribution given in Table 5-2. 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).



Frequency Table Generator

	Start	End	Freq
1	-0.5	0.5	250
2	0.5	1.5	500
3	1.5	2.5	250
4			
5			
6			
7			
8			
9			
10			

Number of 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

Number of Decimals: 0

Random Seed: (if known)

Sample Editor column for generated data: 1

Generate

Sample Editor

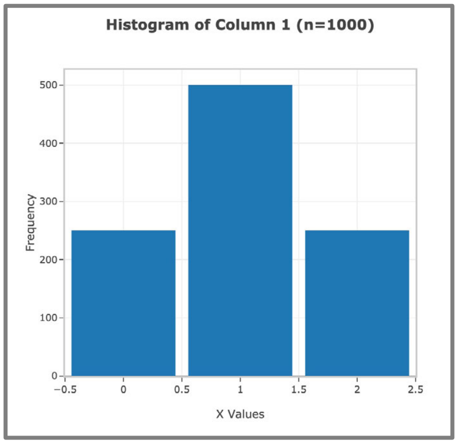
	1
1	1
2	0
3	2
4	0
5	0
6	2
7	1
8	2
9	0
10	2
11	0
12	1
13	1
14	1
15	2
16	1
17	0
18	1
19	1
20	1
21	0
22	1
23	1
24	0
25	1
26	2
27	2
28	1
29	1
30	1

Note these settings in the above display:

- The **Number of Decimals** in the generated values set to **0**.
- **Samples with Same Observed Frequencies** is selected.
- **Equal to Class Midpoints** is selected under *Output Values*.

4. Click **Generate**. Statdisk will generate a set of data corresponding to the probability distribution.

Using the generated data, you can find the mean and standard deviation or you can construct a histogram. There's one correction needed: If using the *Explore Data - 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 Table 5-2, the actual standard deviation is σ = 0.7071068, but Statdisk yields σ = 0.7074606. The value of the mean will be correct. The Statdisk histogram that shows the shape of the probability distribution is shown to the right.



5-2 Binomial Distributions

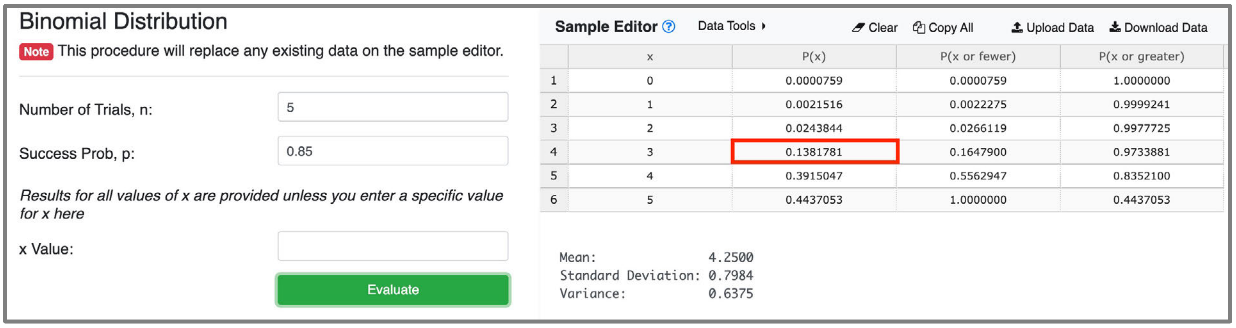
Section 5-2 in the Triola textbook describes three methods for determining probabilities in binomial experiments. Method 2 requires technology. We noted in the textbook that if software is available, this method of finding binomial probabilities is fast and easy, as shown in the Statdisk procedure that follows. We illustrate the Statdisk procedure with the following example:

> **EXAMPLE *Twitter*** Based on a Pew Research Center survey, 85% of adults know what Twitter is. Assuming that we randomly select 5 adults, find the probability that exactly 3 of the 5 adults know what Twitter is

For this example, n = 5, p = 0.85, and the possible values of x are 0, 1, 2, 3, 4, 5.

Statdisk Procedure for Finding Probabilities with a Binomial Distribution

1. Click **Analysis** in the top menu bar.
2. Select **Probability Distributions** from the dropdown menu**.**
3. Select **Binomial Distribution** from the submenu.
4. You will now see the *Binomial Probability* dialog box:
    - Enter the number of trials n (n = 5 in the example).
    - Enter the probability of success p (p = 0.85 in the example).
5. Click **Evaluate**.



Binomial Distribution

Note This procedure will replace any existing data on the sample editor.

Number of Trials, n: 5

Success Prob, p: 0.85

Results for all values of x are provided unless you enter a specific value for x here

x Value:

	x	P(x)	P(x or fewer)	P(x or greater)
1	0	0.0000759	0.0000759	1.0000000
2	1	0.0021516	0.0022275	0.9999241
3	2	0.0243844	0.0266119	0.9977725
4	3	0.1381781	0.1647900	0.9733881
5	4	0.3915047	0.5562947	0.8352100
6	5	0.4437053	1.0000000	0.4437053

```
Mean:                4.2500
Standard Deviation:  0.7984
Variance:            0.6375
```

From this Statdisk display, we can see that $P(3)$ = 0.1381781. 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 0.0266119.
- The probability of 3 or more correct responses is 0.9733881.

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

5-3 Poisson Distributions

The textbook 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 since we can easily find binomial probabilities for a wide range of values for *n* and *p*, so an approximation is not necessary.

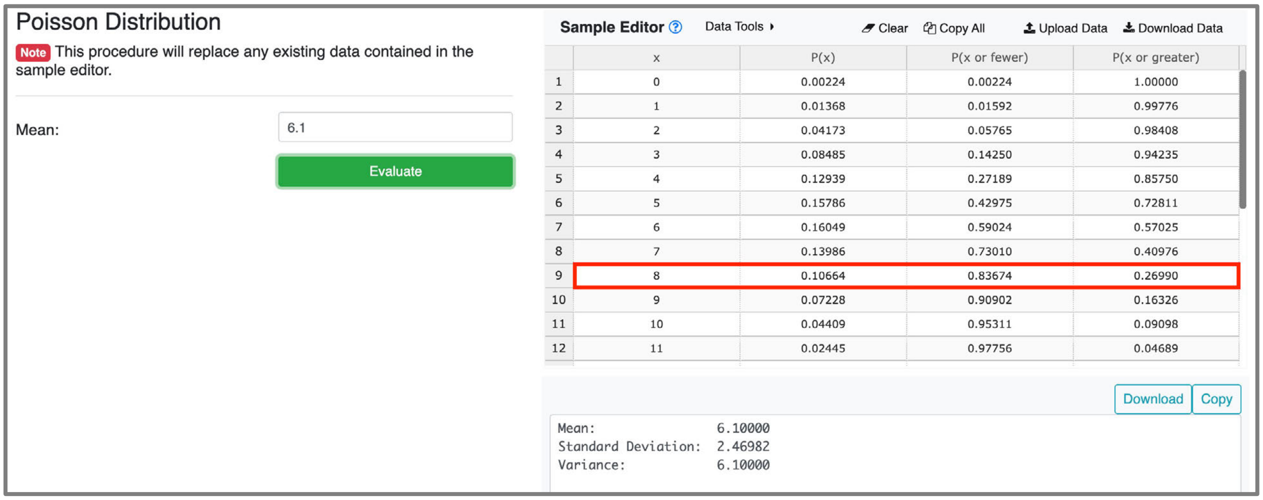
We illustrate the Statdisk procedure with the following problem:

> **EXAMPLE *Hurricanes*** For the 55-year period since 1960, there were 336 Atlantic hurricanes, so that the mean number of hurricanes is 6.1 per year. Find the probability that in a randomly selected year, there are exactly 8 Atlantic hurricanes.

For this example, $\mu = 6.1$ and $x = 8$.

Statdisk Procedure for Finding Probabilities for a Poisson Distribution

1. Determine the value of the mean μ.
2. Click **Analysis** in the top menu bar.
3. Select **Probability Distributions** in the dropdown menu.
4. Select **Poisson Distribution** from the submenu.
5. Enter the value of the mean μ, and then click **Evaluate**. In this example, $\mu = 6.1$.



Mean: 6.1

	x	P(x)	P(x or fewer)	P(x or greater)
1	0	0.00224	0.00224	1.00000
2	1	0.01368	0.01592	0.99776
3	2	0.04173	0.05765	0.98408
4	3	0.08485	0.14250	0.94235
5	4	0.12939	0.27189	0.85750
6	5	0.15786	0.42975	0.72811
7	6	0.16049	0.59024	0.57025
8	7	0.13986	0.73010	0.40976
9	8	0.10664	0.83674	0.26990
10	9	0.07228	0.90902	0.16326
11	10	0.04409	0.95311	0.09098
12	11	0.02445	0.97756	0.04689

```
Mean:                6.10000
Standard Deviation:  2.46982
Variance:            6.10000
```

Note that the display includes values for the mean, standard deviation, and variance. The probabilities and cumulative probabilities are listed in the Sample Editor. For example, *P*(8) = 0.10664, which is the probability of exactly 8 Atlantic hurricanes in a given year. The probability of 8 or fewer hurricanes in a given year is 0.83674. The probability of 8 or or more hurricanes in a given year is 0.26990.

CHAPTER 5 WORKBOOK: Probability Distributions

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

5-1 **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).

Use Statdisk with the procedure described in Section 5-1 of this manual to find the mean and standard deviation: Mean ___________ Standard Deviation _____________

5-2 **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.

$P(35)$ = _______________
$P(45)$ = _______________
$P(50)$ = _______________

5-3 **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 _______________

5-4 **Binomial: 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 10 randomly selected consumers, find the probability that exactly 9 of the 10 consumers recognize her brand name. Find the probability that the number of consumers who recognize her brand name is *not* 9.

____

____

5-5 **Poisson: 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. 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 in *Elementary Statistics* 14th edition, Data Set 39 “Chocolate Chip 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. __________

6 Normal Probability 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 A 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 Appendix Table A-2. Statdisk is much more flexible than the table, and isn't limited to the values included in Table A-2.

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

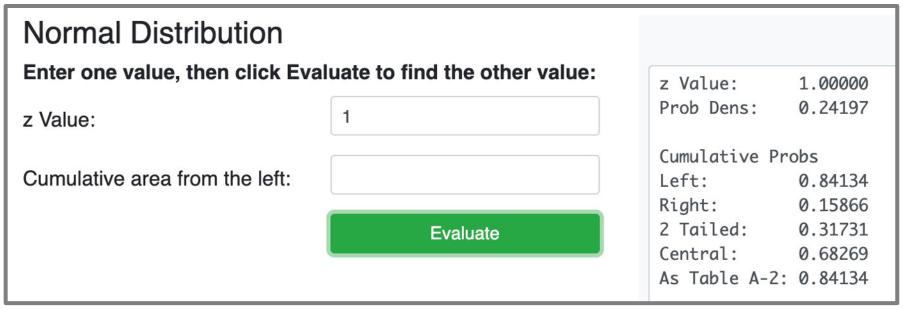
Statdisk Procedure for Finding Probabilities or *z* Scores with a Normal Distribution

1. Click **Analysis** in the top menu.
2. Select **Probability Distributions** from the dropdown menu.
3. Select **Normal Distribution** from the submenu.
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 on the next page.



Normal Distribution

Enter one value, then click Evaluate to find the other value:

z Value: 1

Cumulative area from the left:

```
z Value:      1.00000
Prob Dens:    0.24197

Cumulative Probs
Left:         0.84134
Right:        0.15866
2 Tailed:     0.31731
Central:      0.68269
As Table A-2: 0.84134
```

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

Left	Area below the normal curve and to the left of *z* = 1:	0.84134
Right	Area below the normal curve and to the right of *z* = 1:	0.15866
2 Tailed	Twice the area in the tail bounded by *z* = 1:	0.31731
Central	Twice the area below the curve and bounded by the centerline and *z* = 1:	0.68269
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 Appendix Table A-2.)	0.84134

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 0.24197. This particular value is not used in the textbook.

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 technology to obtain samples from theoretical normal distributions, and Statdisk has such a capability, as described below.

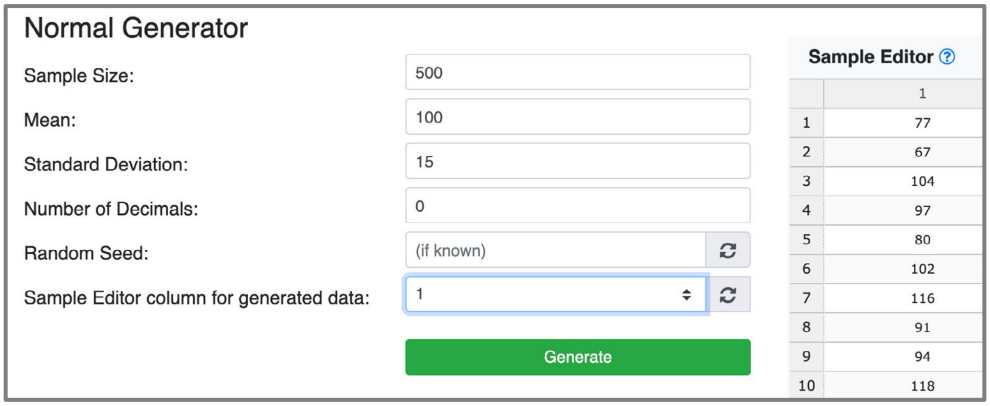
Let's consider IQ scores. IQ tests are designed to produce a mean of 100 and a standard deviation of 15, and we expect that such scores are normally distributed. Suppose we want to learn about the variation of sample means for samples of IQ scores. Instead of going out into the world and randomly selecting 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. Click **Data** in the top menu bar.
2. Select **Normal Generator** from dropdown menu
3. The Normal Generator dialog box appears. 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* input 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 **Generate**.

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. The values can be used with functions 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.



Normal Generator

Sample Size: 500

Mean: 100

Standard Deviation: 15

Number of Decimals: 0

Random Seed: (if known)

Sample Editor column for generated data: 1

Sample Editor

	1
1	77
2	67
3	104
4	97
5	80
6	102
7	116
8	91
9	94
10	118

6-3 Assessing Normality

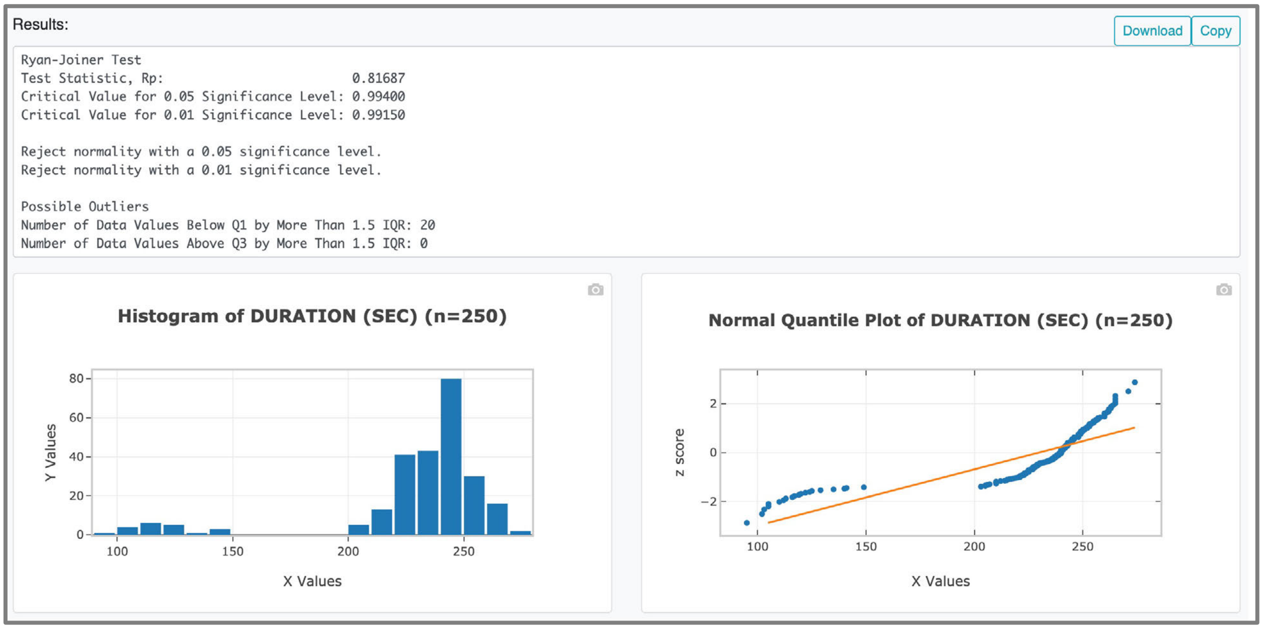
Statdisk includes the **Normality Assessment** function, which generates a histogram and normal quantile plot and also identifies potential outliers. This simple function allows you to determine whether sample data appear to come from a population having a normal distribution.

Statdisk Procedure for Normality Assessment

1. Enter the list of sample data in a single column of the Statdisk Sample Editor. Either manually enter the data or open a data set that had been previously saved, such as any of the textbook Appendix B data sets included in Statdisk.
2. Click the top menu item of **Data**.
3. Select **Normality Assessment** in the dropdown menu.
4. Select the column containing the sample data that is to be analyzed.
5. Click **Evaluate**.

EXAMPLE *Old Faithful Eruption Times* As an example, consider the 250 eruption times of the Old Faithful geyser listed in the *Elementary Statistics* 14th Edition, Data Set 26 "Old Faithful".

After opening this data set, the 250 eruption times (DURATION) are listed in column 2 of the Sample Editor. Using the Statdisk procedure for normality assessment, the Statdisk display will appear as shown below.



```
Ryan-Joiner Test
Test Statistic, Rp:                     0.81687
Critical Value for 0.05 Significance Level: 0.99400
Critical Value for 0.01 Significance Level: 0.99150

Reject normality with a 0.05 significance level.
Reject normality with a 0.01 significance level.

Possible Outliers
Number of Data Values Below Q1 by More Than 1.5 IQR: 20
Number of Data Values Above Q3 by More Than 1.5 IQR: 0
```

We now evaluate the results included in the previous display.

Histogram: The Statdisk display on the previous page includes a histogram showing that the shape of the distribution is *skewed* to the left and is far from being bell-shaped. This suggests that the sample data are *not* from a normally distributed population.

Outliers: The display includes the number of "Possible Outliers." In this case, we see that there are 20 potential outliers that are below the first quartile Q_1 by more than 1.5 times the IQR (interquartile range). If we examine the sorted list of sample *DURATION* values, we see that the lowest values in the sorted list are 95, 102, 103, 105, 105, and so on. The potential outliers are far away from the other sample values, so these potential outliers do appear to be actual outliers. Because there are twenty *potential* outliers and they are far away from the other values, we should reject normality because of outliers.

Normal Quantile Plot: The display includes a normal quantile plot. The points in the normal quantile plot show a systematic pattern that are very far from a straight-line pattern, suggesting that the eruption times are not from a normally distributed population.

Ryan-Joiner Test: The Ryan-Joiner test is one of several formal tests of normality, each having their own advantages and disadvantages. The Statdisk display on the previous page shows that the Ryan-Joiner test results in a conclusion of "reject normality." That is, it appears that the 250 eruption times do not appear to be from a population with a normal distribution.

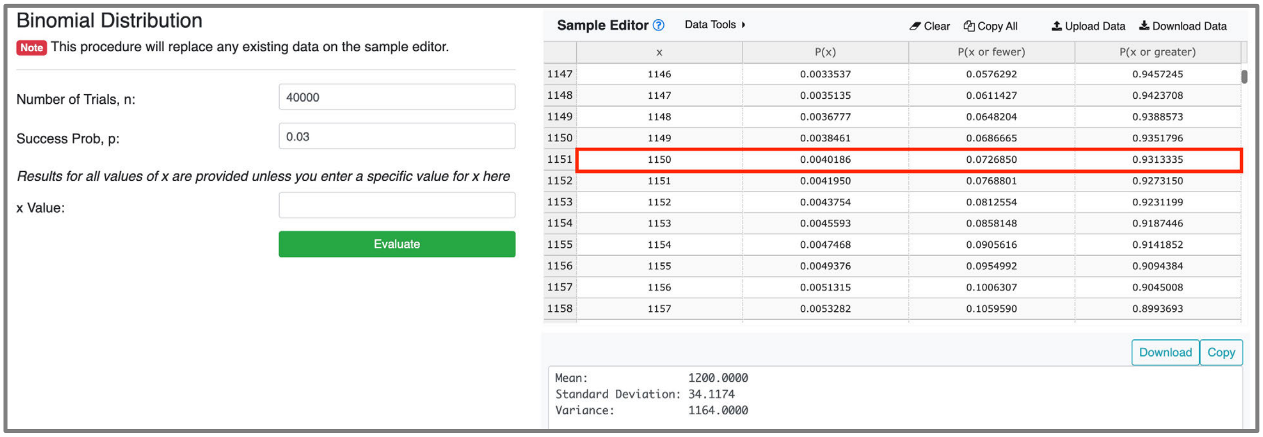
Conclusion: After considering all of the above elements from the single Statdisk display on the previous page, we conclude that 250 eruption times do *not* appear to be from a population with a normal distribution.

6-4 Normal Approximation to Binomial

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 the following example:

> **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 at least 1150 e-mail addresses, find the probability of getting at least 1150 responses with e-mail addresses.

Based on the above statements, 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.



This procedure will replace any existing data on the sample editor.

Number of Trials, n: 40000

Success Prob, p: 0.03

Results for all values of x are provided unless you enter a specific value for x here

x Value:

	x	P(x)	P(x or fewer)	P(x or greater)
1147	1146	0.0033537	0.0576292	0.9457245
1148	1147	0.0035135	0.0611427	0.9423708
1149	1148	0.0036777	0.0648204	0.9388573
1150	1149	0.0038461	0.0686665	0.9351796
1151	1150	0.0040186	0.0726850	0.9313335
1152	1151	0.0041950	0.0768801	0.9273150
1153	1152	0.0043754	0.0812554	0.9231199
1154	1153	0.0045593	0.0858148	0.9187446
1155	1154	0.0047468	0.0905616	0.9141852
1156	1155	0.0049376	0.0954992	0.9094384
1157	1156	0.0051315	0.1006307	0.9045008
1158	1157	0.0053282	0.1059590	0.8993693

```
Mean:               1200.0000
Standard Deviation: 34.1174
Variance:           1164.0000
```

CHAPTER 6 WORKBOOK: Normal Distributions

6-1 **Finding Probabilities for a Normal Distribution** Use Statdisk's *Normal Distribution* function to find the indicated probabilities. First select **Analysis** in 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 98.6, and a standard deviation of 0.62, find the probability of a value between 97.0 and 99.0._________

c. 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?_________________

d. 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?___________________

Assessing Normality *In exercises 6-2 through 6-5, 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.*

6-2 **Earthquake Depth**. The depth (km) measurements of earthquakes recorded in one year from a location in Southern California, as listed in the *Elementary Statistics* 14th Edition, Data Set 24 "Earthquakes."

6-3 **President Heights** The heights of U.S. Presidents, as listed in the *Elementary Statistics* 14th Edition, Data Set 22 "Presidents."

6-4 **Oscar Winning Actor Ages** The ages of actors when they won an Oscar, as listed in the *Elementary Statistics* 14th Edition, Data Set 21 "Oscar Winner Age."

6-5 **Sprint Data Speeds** The measured Sprint airport data speeds (Mbps), as listed in the *Elementary Statistics* 14th Edition, Data Set 34 "Airport Data Speeds."

7 Estimating Parameters and Determining Sample Sizes

Statdisk

7-1 Confidence Intervals for Estimating *p*

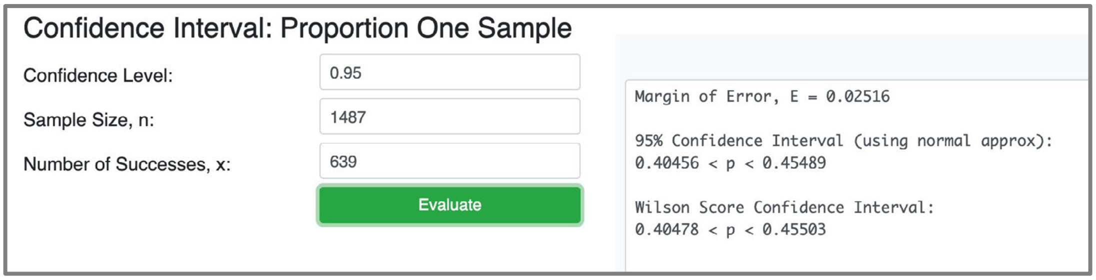
When finding a confidence interval estimate of a population proportion *p*, Statdisk requires the sample size *n* and the number of successes *x*. In some cases, the values of *x* and *n* are both known, but in other cases the given information may consist of the sample size *n* and a sample percentage.

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** in the top menu bar.
2. Select **Confidence Intervals** from the dropdown menu.
3. Select **Proportion One Sample** from the submenu.
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 **Evaluate**, and the Statdisk results are as shown below.



Confidence Interval: Proportion One Sample

Confidence Level: 0.95

Sample Size, n: 1487

Number of Successes, x: 639

```
Margin of Error, E = 0.02516

95% Confidence Interval (using normal approx):
0.40456 < p < 0.45489

Wilson Score Confidence Interval:
0.40478 < p < 0.45503
```

Based on the above Statdisk display, we can express the 95% confidence interval estimate of *p* as 0.40456 < *p* < 0.45489. After rounding, the confidence interval becomes 0.405 < *p* < 0.455. This can also be expressed as 40.5% < *p* < 45.5% or as 43.0% ± 2.5%. (The Wilson Score confidence interval is discussed briefly in the textbook near the end of Section 7-1.)

7-2 Confidence Intervals for Estimating μ

Section 7-2 in the textbook introduces a method for using a sample mean $\bar{x}$ to estimate the value of a population mean μ. Statdisk is very easy to use for constructing confidence interval estimates for a population mean μ.

Statdisk Procedure for Finding Confidence Intervals for μ

1. Select **Analysis** in the menu at the top of the screen.
2. Select **Confidence Intervals** from the subdirectory.
3. Select **Mean One Sample** from the submenu. The *Mean One Sample* dialog box appears.
4. Choose the **Use Summary Statistics** tab or **Use Data** tab.

   Using Summary Statistics: Select the **Use Summary Statistics** tab and enter the confidence level, sample size (n), sample mean ($\bar{x}$), and sample standard deviation (s).

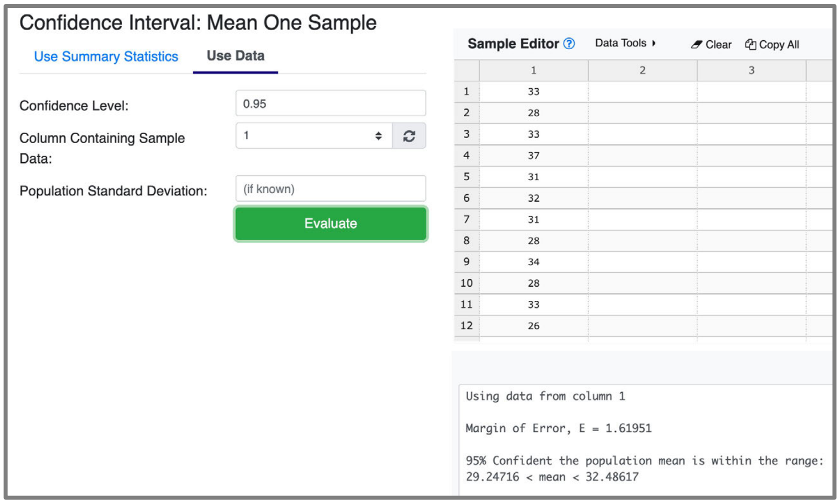
   Using Sample Data: Select the **Use Data** tab and enter the confidence level and select the desired data column.
5. Click the **Evaluate** button.

Consider the following example:

EXAMPLE Listed below are weights (hectograms or hg) of randomly selected girls at birth, based on data from the National Center for Health Statistics.

33 28 33 37 31 32 31 28 34 28 33 26 30 31 28

The following screen shows the results obtained using the 15 birth weights listed above.



Confidence Interval: Mean One Sample

Confidence Level: 0.95

Column Containing Sample Data: 1

Population Standard Deviation: (if known)

	1	2	3
1	33		
2	28		
3	33		
4	37		
5	31		
6	32		
7	31		
8	28		
9	34		
10	28		
11	33		
12	26		

```
Using data from column 1

Margin of Error, E = 1.61951

95% Confident the population mean is within the range:
29.24716 < mean < 32.48617
```

Based on the above Statdisk display, the 95% confidence interval is 29.2 < μ < 32.5 (rounded). The solution includes this statement: We are 95% confident that the limits of 29.2 hg and 32.5 hg actually do contain the value of the population mean μ.

7-3 Confidence Intervals for Estimating σ

After selecting a confidence level and entering the sample data or summary statistics, 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.* 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** in the top menu bar.
2. Select **Confidence Intervals** from the dropdown menu**.**
3. Select **Standard Deviation One Sample** from the submenu. The *Standard Deviation One Sample* dialog box appears.
4. Choose the **Use Summary Statistics** tab or **Use Data** tab.

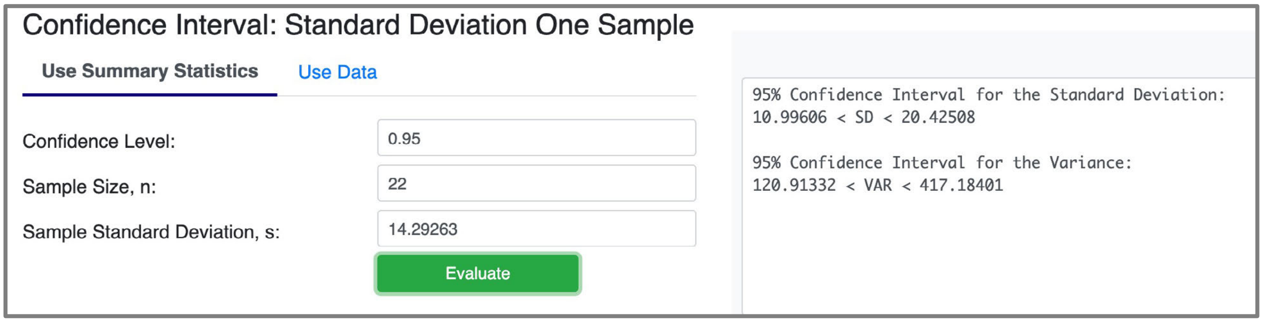
   Using Summary Statistics: Select the **Use Summary Statistics** tab and enter the confidence level, sample size (*n*), and sample standard deviation (*s*).

   Using Sample Data: Select the **Use Data** tab, enter the confidence level and select the desired data column.
5. Click the **Evaluate** button.

Consider the following example.

EXAMPLE A sample of size $n = 22$ has standard deviation $s = 14.29263$, and we want to construct a 95% confidence interval estimate of σ.

Because the values of $n = 22$ and $s = 14.29263$ are known, we use Statdisk to obtain the 95% confidence interval estimate of the population standard deviation: $11.0 < \sigma < 20.4$ (rounded).



Confidence Interval: Standard Deviation One Sample

Confidence Level: 0.95

Sample Size, n: 22

Sample Standard Deviation, s: 14.29263

```
95% Confidence Interval for the Standard Deviation:
10.99606 < SD < 20.42508

95% Confidence Interval for the Variance:
120.91332 < VAR < 417.18401
```

7-4 Sample Sizes for Estimating *p*

Section 7-1 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.

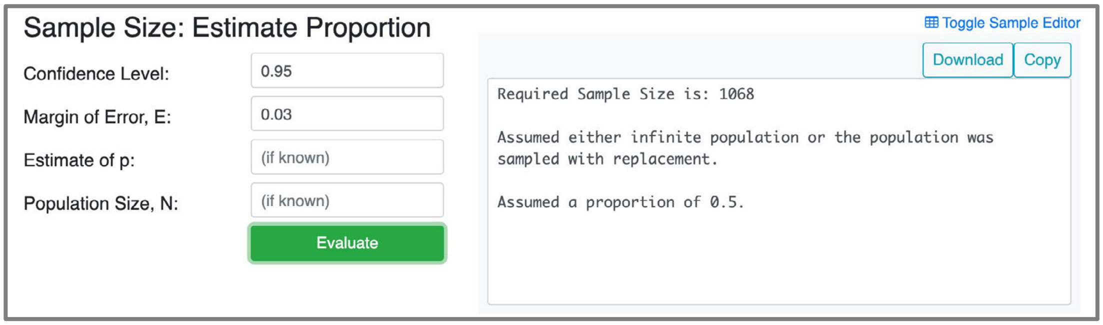
Statdisk Procedure for Finding Sample Sizes Required to Estimate *p*

1. Select **Analysis** in the top menu bar.
2. Select **Sample Size Determination** in the dropdown menu.
3. Select **Estimate Proportion** from the submenu.
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 the following example.

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


Sample Size: Estimate Proportion

Confidence Level: 0.95

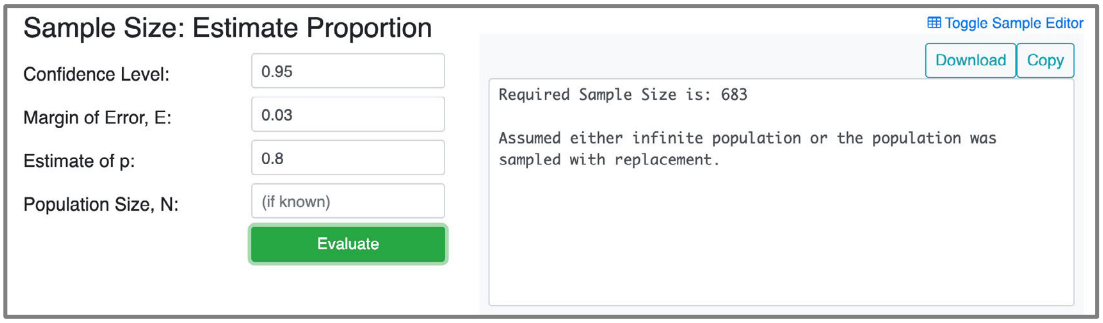
Margin of Error, E: 0.03

Estimate of p: (if known)

Population Size, N: (if known)

```
Required Sample Size is: 1068

Assumed either infinite population or the population was
sampled with replacement.

Assumed a proportion of 0.5.
```

We can also determine 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 estimated proportion is 80%. Using the Statdisk procedure on the previous page, we find that a simple random sample of 683 adults is needed, as shown in the display below.


Sample Size: Estimate Proportion

Confidence Level: 0.95

Margin of Error, E: 0.03

Estimate of p: 0.8

Population Size, N: (if known)

```
Required Sample Size is: 683

Assumed either infinite population or the population was
sampled with replacement.
```

7-5 Sample Sizes for Estimating μ

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.

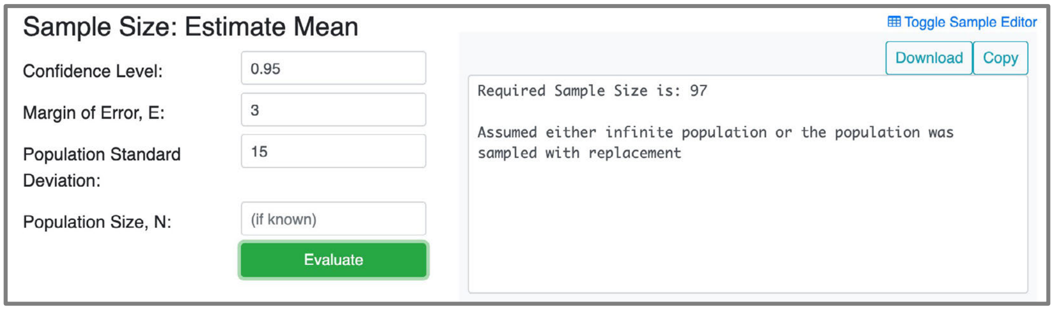
Statdisk Procedure for Finding Sample Sizes Required to Estimate μ

1. Select **Analysis** from the top menu bar.
2. Select **Sample Size Determination** from the dropdown menu.
3. Select the option of **Estimate Mean** from the submenu.
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 the **Evaluate** button.

Consider the following example.

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



Sample Size: Estimate Mean

Confidence Level: 0.95

Margin of Error, E: 3

Population Standard Deviation: 15

Population Size, N: (if known)

```
Required Sample Size is: 97

Assumed either infinite population or the population was
sampled with replacement
```

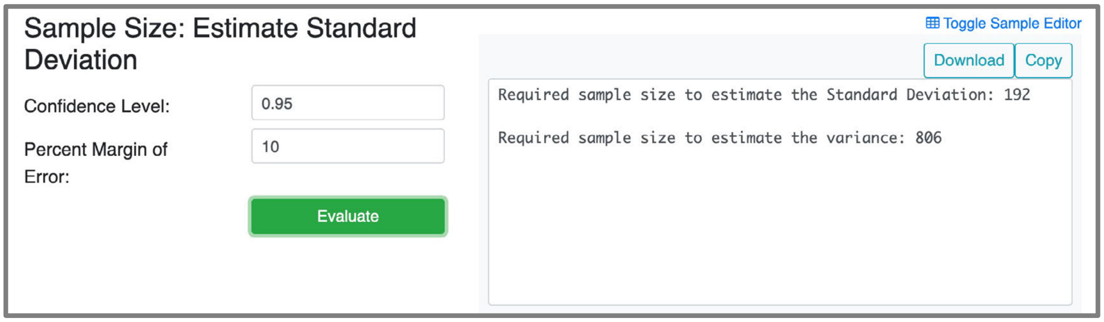
7-6 Sample Sizes for Estimating σ

The textbook describes a procedure for determining the sample size required to estimate a population standard deviation σ or population variance σ^2. The Triola 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** from the dropdown menu.
3. Select **Estimate Standard Deviation** from the submenu.
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.



Sample Size: Estimate Standard Deviation

Confidence Level: 0.95

Percent Margin of Error: 10

```
Required sample size to estimate the Standard Deviation: 192

Required sample size to estimate the variance: 806
```

7-7 Bootstrap Resampling

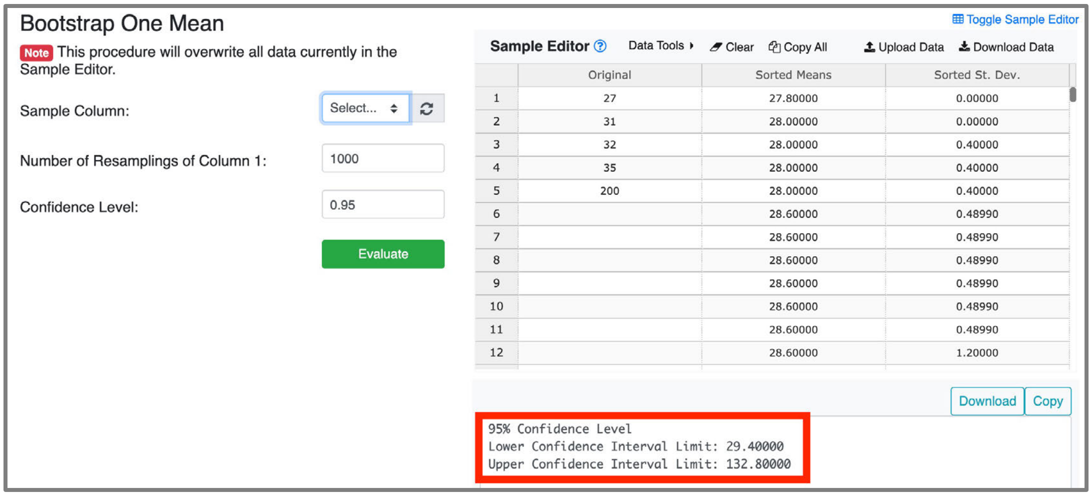
When the sample is small, the σ is unknown, and the distribution is not normal, the methods of Chapter 7 in the textbook cannot be used to construct a confidence interval estimate of the population mean. One alternative is to use the method of **bootstrap resampling**, which does not require normally distributed data.

Statdisk Procedure for Bootstrap Resampling

1. Select **Resampling** from the top menu bar.
2. Select the desired type of resampling from the dropdown menu. Options include:
   - **Bootstrap One Proportion**
   - **Bootstrap Two Proportions**
   - **Bootstrap One Mean**
   - **Bootstrap Two Means**
   - **Bootstrap Matched Pairs**
3. Enter the required inputs which include the confidence level (such as 0.95 or 0.99) and desired number of resamplings (such as 1000).
4. Click the **Evaluate** button. The sorted results are listed in the Sample Editor.

EXAMPLE Use bootstrap resampling to construct a 95% confidence interval estimate of the population mean μ using the sample values of 27, 31, 32, 35, and 200.

Using Statdisk, 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.)



Bootstrap One Mean

Note This procedure will overwrite all data currently in the Sample Editor.

Sample Column:

Number of Resamplings of Column 1: 1000

Confidence Level: 0.95

	Original	Sorted Means	Sorted St. Dev.
1	27	27.80000	0.00000
2	31	28.00000	0.00000
3	32	28.00000	0.40000
4	35	28.00000	0.40000
5	200	28.00000	0.40000
6		28.60000	0.48990
7		28.60000	0.48990
8		28.60000	0.48990
9		28.60000	0.48990
10		28.60000	0.48990
11		28.60000	0.48990
12		28.60000	1.20000

```
95% Confidence Level
Lower Confidence Interval Limit: 29.40000
Upper Confidence Interval Limit: 132.80000
```

CHAPTER 7 WORKBOOK: Confidence Intervals and Sample Sizes

In 7–1 and 7–2, 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 identity 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 **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 *Elementary Statistics* 14th edition, Data Set 1 “Body Data” (GENDER: 1 = male, 0 = female).

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?

____

7-4 **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.________________

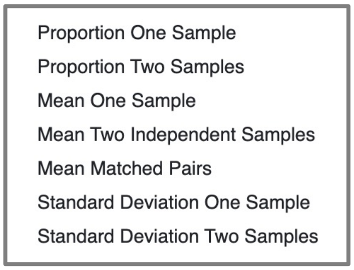
7-5 **Bootstrap Resampling** The sample values 2.9, 564.2, 1.4, 4.7, 67.6, 4.8, 51.3, 3.6, 18.0, and 3.6 are randomly selected from a population with a distribution that is far from normal. Use bootstrap resampling to construct a 95% confidence interval estimate of μ and use bootstrap resampling to construct a 95% confidence interval estimate of σ. Enter the results here.

________________________________ ________________________________

8 Hypothesis Testing

Statdisk

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


Proportion One Sample
Proportion Two Samples
Mean One Sample
Mean Two Independent Samples
Mean Matched Pairs
Standard Deviation One Sample
Standard Deviation Two Samples

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.

8-1 Testing Hypotheses About a Proportion *p*

The Statdisk procedure for testing claims about a population proportion is quite easy. Given a claim about a proportion and knowing *n* and *x*, we can use Statdisk as follows.

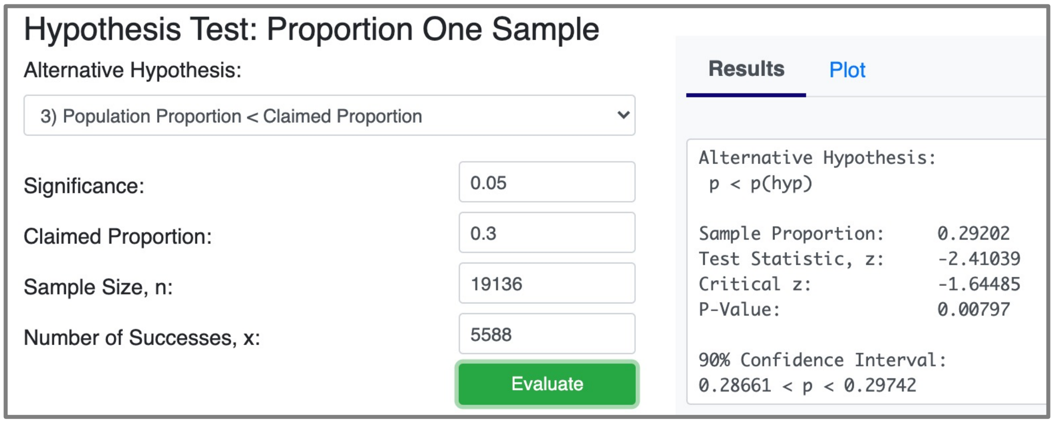
Statdisk Procedure for Testing Claims about *p*

1. Select **Analysis** from the top menu bar.
2. Select **Hypothesis Testing** from the dropdown menu.
3. Select **Proportion One Sample** from the submenu.
4. Make these entries in the dialog box.
   - In the *Alternative Hypothesis* box, select the format of the alternative hypothesis.
   - 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 the following example.

> **Example** A study of sleepwalking or "nocturnal wandering" was described in *Neurology* magazine, and it included information that 29.2% of 19,136 American adults have sleepwalked. Use a 0.05 significance level to test the claim that fewer than 30% of adults have sleepwalked.

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



Hypothesis Test: Proportion One Sample

Alternative Hypothesis: 3) Population Proportion < Claimed Proportion

Significance: 0.05

Claimed Proportion: 0.3

Sample Size, n: 19136

Number of Successes, x: 5588

```
Alternative Hypothesis:
 p < p(hyp)

Sample Proportion:      0.29202
Test Statistic, z:      -2.41039
Critical z:             -1.64485
P-Value:                0.00797

90% Confidence Interval:
0.28661 < p < 0.29742
```

Important elements of the Statdisk results include the *P*-value of 0.00797, the test statistic of *z* = -2.41039, and a critical value of -1.64485. Having the *P*-value and critical value 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 0.00797 is less than the significance level of 0.05, so we reject the null hypothesis. If we were to use the critical value approach, we see that the test statistic of *z* = -2.41039 does fall within the critical region, so we reject the null hypothesis. Conclusion: There is sufficient sample evidence to support the claim that fewer than 30% of adults have sleepwalked.

Note also that the Statdisk display includes this 90% confidence interval:

$$0.28661 < p < 0.29742$$

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 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

When testing claims about a population mean μ, 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.

Statdisk greatly simplifies the process because it is programmed to use the correct procedure 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 **Analysis** in the top menu.
2. Select **Hypothesis Testing** from the dropdown menu.
3. Select **Mean One Sample** from the submenu.
4. Choose the **Use Summary Statistics** tab or **Use Data** tab.

   Using Summary Statistics: Select the **Use Summary Statistics** tab. This option requires that you enter the sample size (n), sample mean ($\bar{x}$), and sample standard deviation (s).

   Using Sample Data: Select the **Use Data** tab and select the desired data column.
5. You will now see the required inputs.
   - In the **Alternative Hypothesis** box, select the format of the alternative hypothesis.
   - 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.)
6. Click the **Evaluate** button to get the test results.
7. Click the **Plot** button to get a graph that shows the test statistic and critical values.

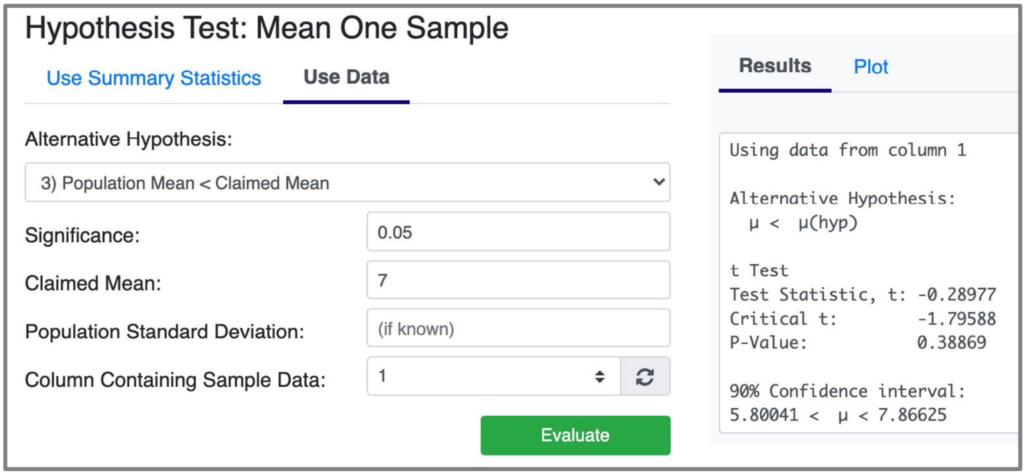
As an illustration, consider the following example.

EXAMPLE ***Adult Sleep*** Listed below are the sleep times (in hours) for adult subjects included in the National Health and Nutrition Examination Study. Use a 0.05 significance level to test the claim that the mean amount of sleep for adults is less than 7 hours.

4 8 4 4 8 6 9 7 7 10 7 8

Because $n \leq 30$, we must verify that the sample data appear to be from a normally distributed population. Using Statdisk's **Data - Normality Assessment** function (see section 6-3 in this manual), we find that this requirement is met.

Follow the procedure on the previous page and make the entries as shown in the Statdisk display.



Hypothesis Test: Mean One Sample

Alternative Hypothesis: 3) Population Mean < Claimed Mean

Significance: 0.05

Claimed Mean: 7

Population Standard Deviation: (if known)

Column Containing Sample Data: 1

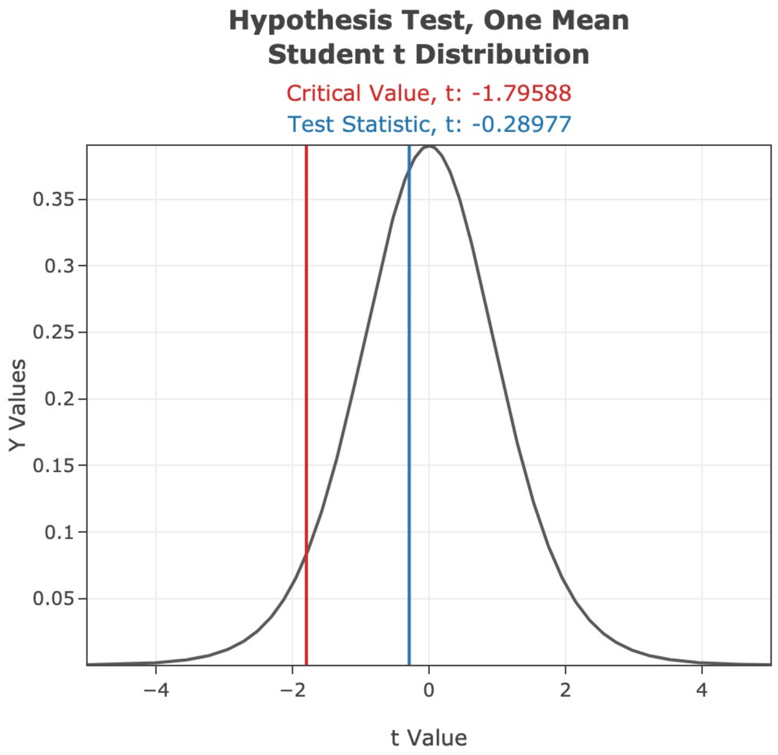
```
Using data from column 1

Alternative Hypothesis:
  μ <  μ(hyp)

t Test
Test Statistic, t: -0.28977
Critical t:        -1.79588
P-Value:           0.38869

90% Confidence interval:
5.80041 <  μ < 7.86625
```

From the above Statdisk display, we see that the test statistic is t = -0.28977. The P-value is 0.38869. 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 mean amount of sleep is less than 7 hours.

Note that the display also includes the 90% confidence interval of $5.8 < \mu < 7.9$ (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 $5.8 < \mu < 7.9$ (rounded) suggests that the population mean μ can be any value between 5.8 and 7.9, and the assumed mean of 7.0 does fall within those limits, so there is not sufficient evidence to support the claim that $\mu < 7.0$.

Clicking the **Plot** button will generate graph shown below. The vertical line representing the test statistic will be in blue, and vertical line(s) representing the critical value(s) will be in red.



8-3 Testing Hypotheses About σ or σ^2

Although Statdisk is designed to work only with standard deviations, claims about a population variance can be handled as well. 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.

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*. Given sample data, use Statdisk's **Normality Assessment** function (see section 6-3 in this manual) 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.

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 **Standard Deviation One Sample**. (Select this option for claims about standard deviations or variances.)
4. Next, choose the **Use Summary Statistics** tab or **Use Data** tab.

   Using Summary Statistics: Select the **Use Summary Statistics** tab. You will need to enter the sample size (*n*), and sample standard deviation (*s*).

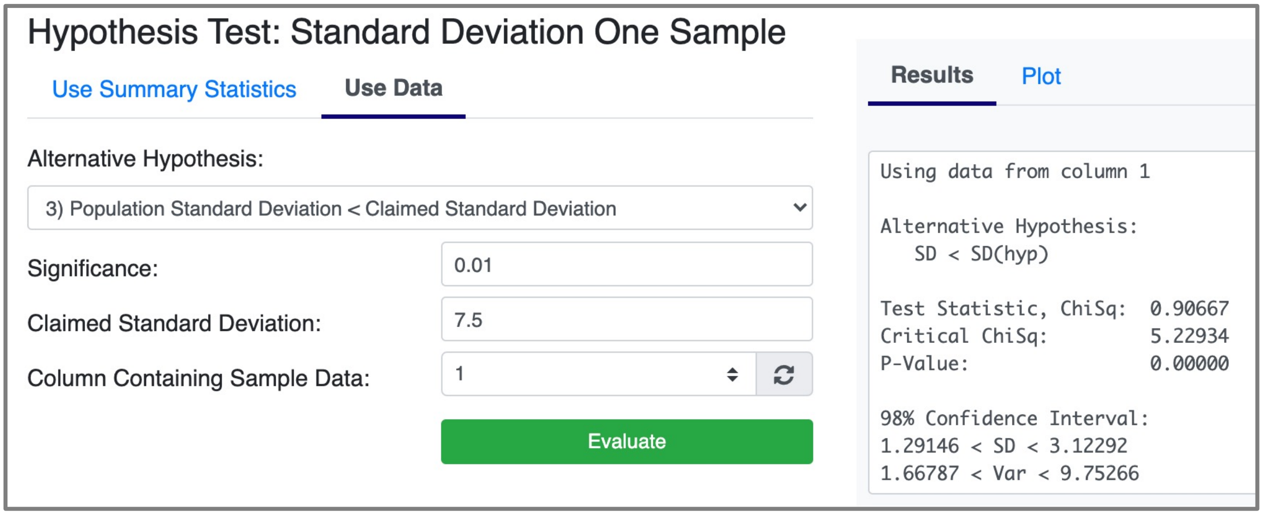
   Using Sample Data: Select the **Use Data** tab and select the desired data column.
5. Make the following entries in the dialog box.
   - In the **Alternative Hypothesis** box, select the format of the claim being tested.
   - Enter a significance level, such as 0.05 or 0.01.
   - Enter the *claimed* value of the standard deviation. (This is the value used in the statement of the null hypothesis.)
6. Click **Evaluate**.

Consider the following example.

EXAMPLE ***Supermodel Heights*** Listed below are the heights (cm) for the simple random sample of supermodels. We will use a 0.01 significance level to test the claim that supermodels have heights with a standard deviation that is less than 7.5 cm for the population of women in general.

178 177 176 174 175 178 175 178 178 177 180 176 180 178 180 176

From the example, we see that we want to test the claim that $\sigma < 7.5$ cm, and we want to use a 0.01 significance level. 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 results is shown below.



Hypothesis Test: Standard Deviation One Sample

Alternative Hypothesis: 3) Population Standard Deviation < Claimed Standard Deviation

Significance: 0.01

Claimed Standard Deviation: 7.5

Column Containing Sample Data: 1

```
Using data from column 1

Alternative Hypothesis:
   SD < SD(hyp)

Test Statistic, ChiSq:  0.90667
Critical ChiSq:         5.22934
P-Value:                0.00000

98% Confidence Interval:
1.29146 < SD < 3.12292
1.66787 < Var < 9.75266
```

The Statdisk results include the test statistic of $\chi^2 = 0.90667$ and a *P*-value of 0.00000. 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 7.5 cm.

Because critical values are included in the Statdisk display, the critical value method 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).

8-4 Hypothesis Testing with Randomization (One Proportion or One Mean)

Sections 8-1, 8-2, and 8-3 of this manual 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 resampling. Resampling methods include bootstrap resampling and randomization. Here we review the Statdisk procedure for randomization using one sample. See Section 7-7 "Bootstrap Resampling" in this manual for a review of the Statdisk procedure for bootstrap resampling. See Section 9-5 for a review of the Statdisk procedure for randomization using two samples.

Statdisk Procedure for Testing Hypotheses using Randomization

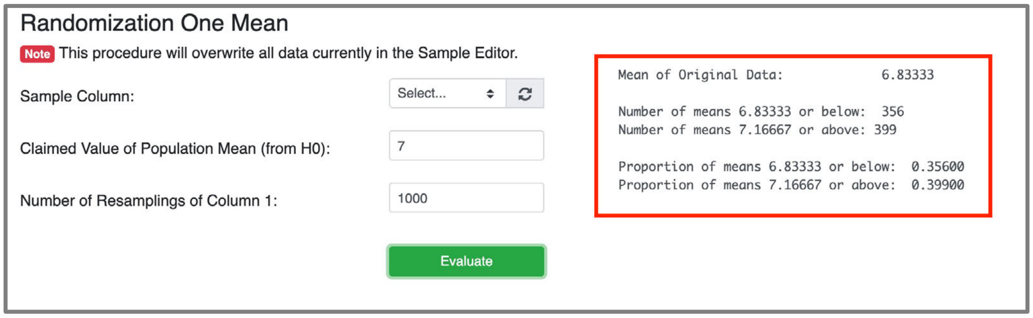
1. Select **Resampling** from the top menu bar.
2. Select the desired type of resampling from the dropdown menu. Options include:
   - **Randomization One Proportion**
   - **Randomization One Mean**
3. Enter the required inputs which include the claimed value of the population mean or proportion and desired number of resamplings.
4. Click the **Evaluate** button.

Consider the following example.

EXAMPLE ***Adult Sleep*** Listed below are hours slept in one night for randomly selected adults. Use randomization to test the claim that the mean amount of sleep for adults is less than 7 hours. Use a 0.05 significance level.

4 8 4 4 4 8 6 9 7 7 10 7 8

Using Statdisk to perform 1000 resamples yields the results shown below, which are typical. Among 1000 resamples there are 356 sample means that are 6.83333 or lower, so there appears to be about a 0.356 chance of getting a sample mean of 6.8333 or lower. The sample mean does not appear to be significantly low, so there is not sufficient evidence to support the claim that the mean amount of sleep for adults is less than 7 hours.



Randomization One Mean

Note This procedure will overwrite all data currently in the Sample Editor.

Sample Column: Select...

Claimed Value of Population Mean (from H0): 7

Number of Resamplings of Column 1: 1000

```
Mean of Original Data:            6.83333

Number of means 6.83333 or below:  356
Number of means 7.16667 or above: 399

Proportion of means 6.83333 or below:  0.35600
Proportion of means 7.16667 or above:  0.39900
```

CHAPTER 8 WORKBOOK: Hypothesis Testing

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 **Earthquake Magnitudes** Use the earthquake magnitudes listed in *Elementary Statistics* 14th edition Data Set 24 “Earthquakes” 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: ____

____

8-3 **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: ____

____

8-4 **Body Temperatures** *Elementary Statistics* 14th edition Data Set 5 “Body Temperatures” includes body temperatures measured at different times on different days. Use randomization (1000 resamples) and the column “DAY 1 – 12AM” data to test the claim that the population has a mean body temperature equal to 98.6°F. How many resample means are at least as extreme as the original sample mean of 98.12 °F? ____________

What should be concluded base on the results?___________________________________

____

9 Inferences from Two Samples

9-1 Two Proportions

9-2 Two Means: Independent Samples

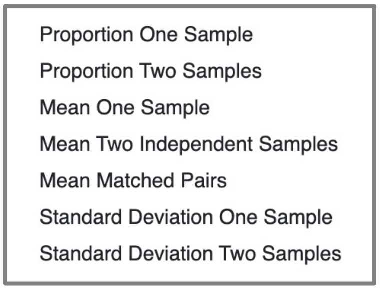
9-3 Matched Pairs

9-4 Two Variances or Standard Deviations

9-5 Inferences with Randomization (Two Samples)

Statdisk

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


Proportion One Sample
Proportion Two Samples
Mean One Sample
Mean Two Independent Samples
Mean Matched Pairs
Standard Deviation One Sample
Standard Deviation Two Samples

Among the items in the above list, four involve *two samples*, as described in Chapter 9 of the textbook. This chapter focuses on these four Statdisk functions, including **Proportion Two Samples**, **Mean Two Independent Samples**, **Mean Matched Pairs** and **Standard Deviation Two Samples**.

9-1 Two Proportions

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.

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** in the top menu bar.
3. Select **Hypothesis Testing** in the dropdown menu.
4. Select **Proportion Two Samples** in the submenu.
5. Make the following entries in the dialog box.
   - In the **Alternative Hypothesis** box, select the format corresponding to the alternative hypothesis
   - 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 the **Evaluate** button to obtain the test results.
7. Click the **Plot** button to obtain a graph that includes the test statistic and critical value(s).

As an illustration, consider the following example.

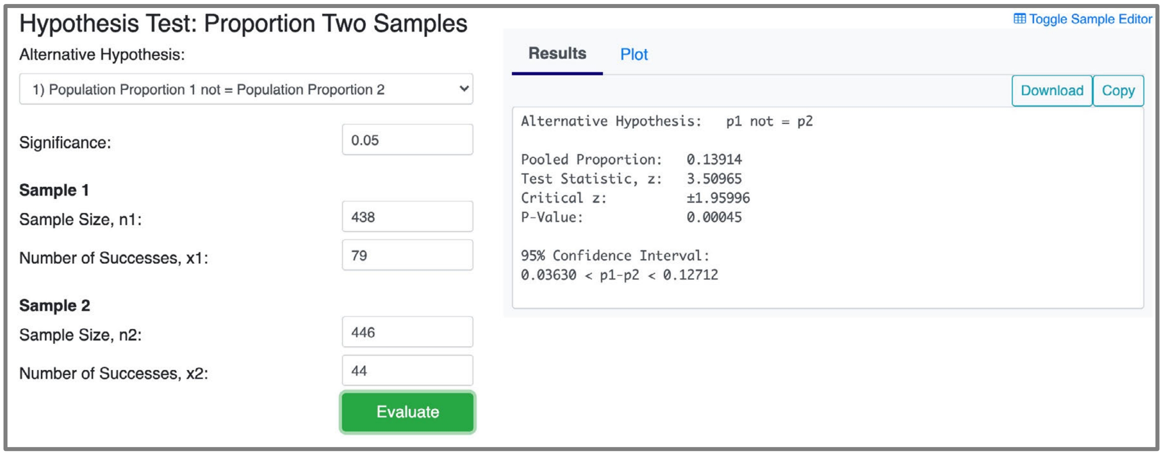
EXAMPLE ***Results of Smoking Cessation Trials*** The table below contains sample results from a smoking cessation trial comparing the effectiveness of e-cigarette and nicotine replacement treatments. Use a 0.05 significance level to test the claim that there is no difference between these two treatment groups.

TABLE 9-1 Results of Smoking Cessation Trials

	E-Cigarettes	Nicotine Replacement
Not smoking after 52 weeks	79	44
Number of subjects	438	446

We can now proceed to use the Statdisk procedure on the previous page.

The Statdisk results below include the pooled proportion, test statistic, critical value, and *P*-value. Because the *P*-value of 0.00045 is less than the significance level of 0.05, we reject the null hypothesis and we conclude that there is sufficient evidence to warrant rejection of the claim that $p_1 = p_2$. It appears that there is a significant difference in success rates in the two treatment groups.



Hypothesis Test: Proportion Two Samples

Alternative Hypothesis: 1) Population Proportion 1 not = Population Proportion 2

Significance: 0.05

Sample 1

Sample Size, n1: 438

Number of Successes, x1: 79

Sample 2

Sample Size, n2: 446

Number of Successes, x2: 44

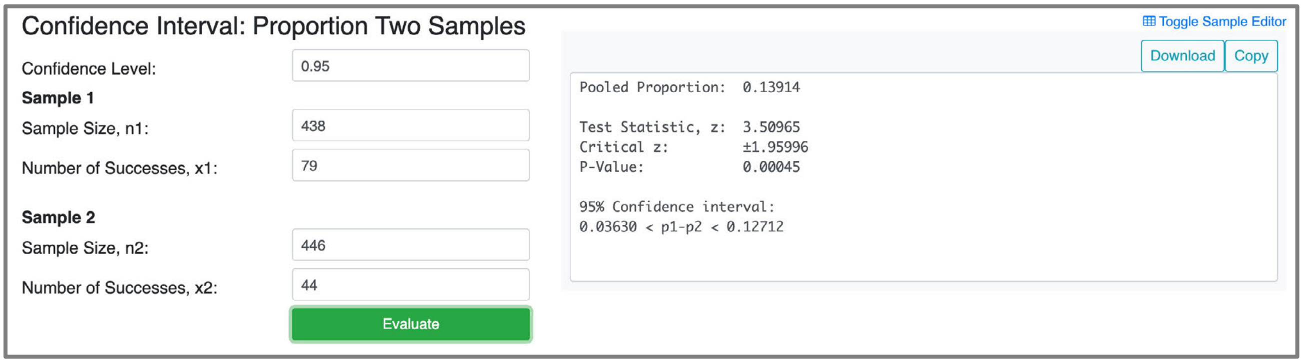
```
Alternative Hypothesis:   p1 not = p2

Pooled Proportion:   0.13914
Test Statistic, z:   3.50965
Critical z:          ±1.95996
P-Value:             0.00045

95% Confidence Interval:
0.03630 < p1-p2 < 0.12712
```

Note that the above results also provide the 95% 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 using the procedure below. Note that this procedure provides the same confidence interval results as the previous procedure.

Statdisk Procedure for a Confidence Interval Estimate of the Difference Between Two Proportions

1. Select **Analysis** in the top menu bar.
2. Select **Confidence Intervals** in the dropdown menu.
3. Select **Proportion Two Samples** in the submenu.
4. Enter the desired confidence level.
5. Enter the sample size and number of successes for each of the two samples.
6. Click the **Evaluate** button.


Confidence Interval: Proportion Two Samples

Confidence Level: 0.95

Sample 1

Sample Size, n1: 438

Number of Successes, x1: 79

Sample 2

Sample Size, n2: 446

Number of Successes, x2: 44

```
Pooled Proportion:  0.13914

Test Statistic, z:  3.50965
Critical z:         ±1.95996
P-Value:            0.00045

95% Confidence interval:
0.03630 < p1-p2 < 0.12712
```

9-2 Two Means: Independent Samples

Statdisk Procedure for Tests of Hypotheses about Two Means: Independent Samples

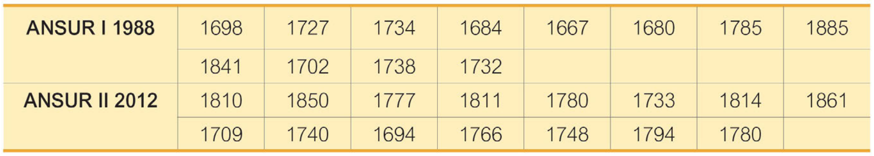
1. Select **Analysis** in the top menu bar.
2. Select **Hypothesis Testing** in the dropdown menu.
3. Select **Mean Two Independent Samples** from the submenu.
4. Next, choose the **Use Summary Statistics** tab or **Use Data** tab.
   Using Summary Statistics: Select the **Use Summary Statistics** tab. and enter the sample size (n), sample mean ($\overline{x}$), and sample standard deviation (s) for both samples.

   Using Sample Data: Select the **Use Data** tab and select the desired data column for both samples.

   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.
5. Select the desired format for **Alternative Hypothesis** and enter the significance level.
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 **Unequal variances: No Pool** (which means that the sample variances will not be pooled as described in the textbook).
   - The option of "Equal variances: 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 "Preliminary 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.**

Consider the following example.

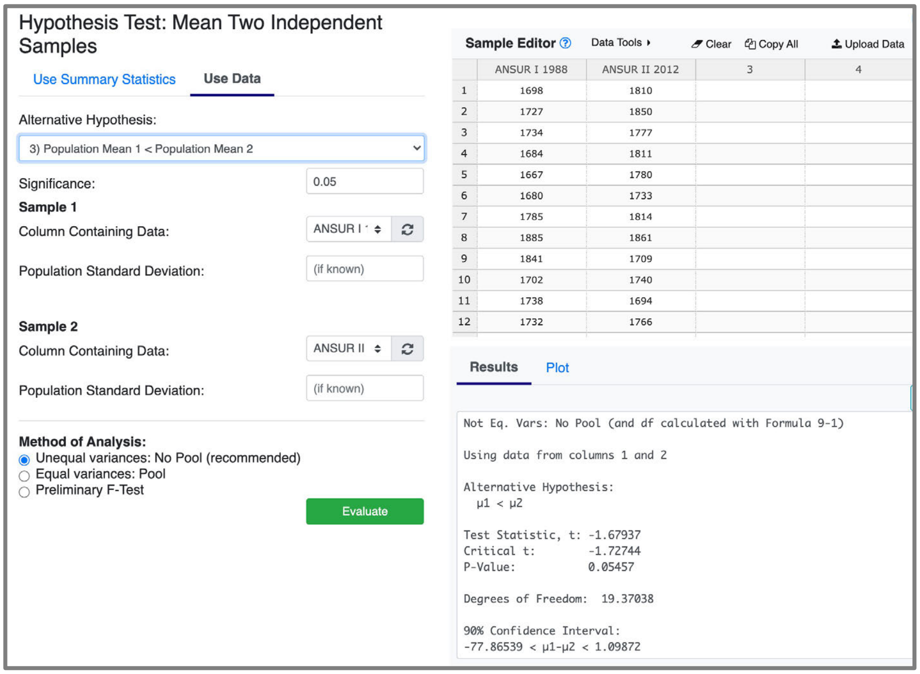
EXAMPLE ***Are People Getting Taller?*** Listed below are heights (mm) of randomly selected U.S. Army personnel measured in 1988 and different heights (mm) measured in 2012. Use a 0.05 significance level to test the claim that the mean height of the 1988 population is less than the mean height of the 2012 population.



ANSUR I 1988	1698	1727	1734	1684	1667	1680	1785	1885
	1841	1702	1738	1732				
ANSUR II 2012	1810	1850	1777	1811	1780	1733	1814	1861
	1709	1740	1694	1766	1748	1794	1780	

Both samples are small (30 or fewer), so we need to determine whether both samples come from populations having normal distributions. Statdisk's **Normality Assessment** function (see section 6-3 in this manual) indicates that the samples are from populations having distributions that are not far from normal.

The Statdisk results are shown below. Because the *P*-value of 0.05457 is greater than the significance level of 0.05, we do not support $\mu_1 < \mu_2$. We conclude that there is not sufficient evidence to support the claim that the mean height of the 1988 male population is less than the mean height of the 2012 male population.


Hypothesis Test: Mean Two Independent Samples

Alternative Hypothesis: 3) Population Mean 1 < Population Mean 2

Significance: 0.05

Sample 1

Column Containing Data: ANSUR I

Population Standard Deviation: (if known)

Sample 2

Column Containing Data: ANSUR II

Population Standard Deviation: (if known)

Method of Analysis:

● Unequal variances: No Pool (recommended)
○ Equal variances: Pool
○ Preliminary F-Test

	ANSUR I 1988	ANSUR II 2012	3	4
1	1698	1810		
2	1727	1850		
3	1734	1777		
4	1684	1811		
5	1667	1780		
6	1680	1733		
7	1785	1814		
8	1885	1861		
9	1841	1709		
10	1702	1740		
11	1738	1694		
12	1732	1766		

```
Not Eq. Vars: No Pool (and df calculated with Formula 9-1)

Using data from columns 1 and 2

Alternative Hypothesis:
  μ1 < μ2

Test Statistic, t: -1.67937
Critical t:        -1.72744
P-Value:           0.05457

Degrees of Freedom:  19.37038

90% Confidence Interval:
-77.86539 < μ1-μ2 < 1.09872
```

The above results also provide the 90% confidence interval estimate of the difference between means ($\mu_1 - \mu_2$). 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 procedure on the next page. This procedure provides the same confidence interval results as the previous procedure.

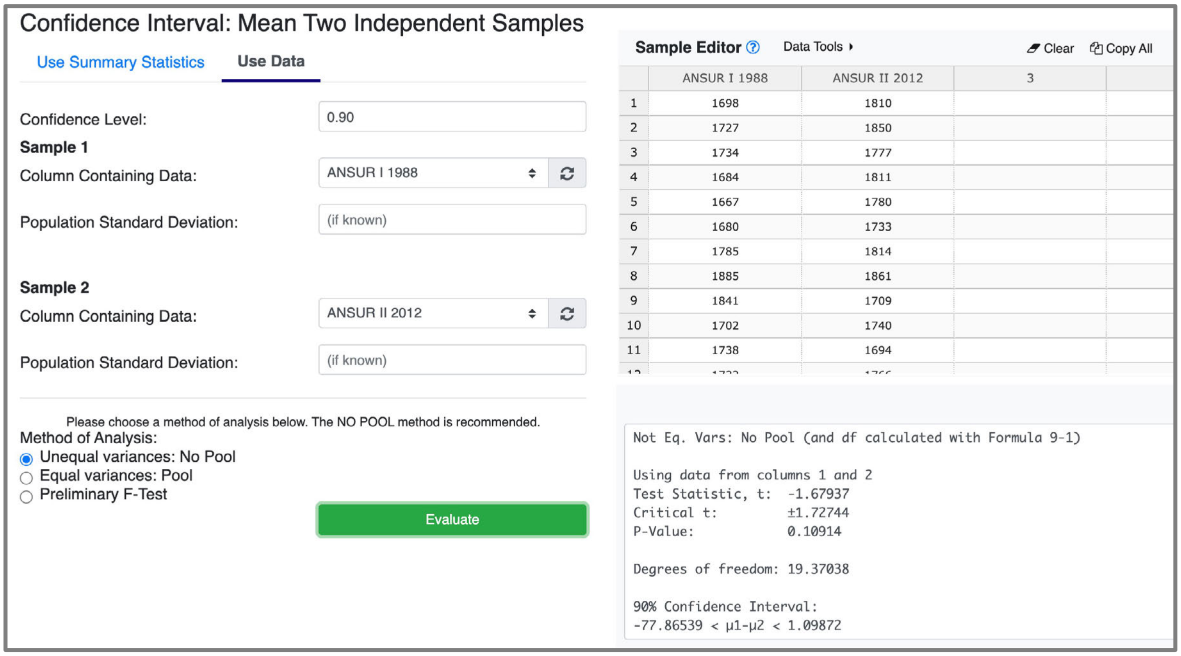
Statdisk Procedure for Confidence Interval Estimates of the Difference Between Two Means: Independent Samples

1. Select **Analysis** in the top menu bar.
2. Select **Confidence Intervals** in the dropdown menu.
3. Choose **Mean Two Independent Samples** in the submenu.
4. Next, choose the **Use Summary Statistics** tab or **Use Data** tab.
   Using Summary Statistics: Select the **Use Summary Statistics** tab and enter the confidence level, sample size (n), sample mean ($\overline{x}$), and sample standard deviation (s) for both samples.

   Using Sample Data: Select the **Use Data** tab, enter the confidence level, and select the desired data column for both samples.

   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.
5. Select the **Method of Analysis** (*Unequal variances: No Pool* is recommended. See the beginning of Section 9-2 in this manual for more detail on these options.)
6. Click **Evaluate.**

If this procedure is used with the sample data from the preceding example, the 90% 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.



Confidence Interval: Mean Two Independent Samples

Confidence Level: 0.90

Sample 1

Column Containing Data: ANSUR I 1988

Population Standard Deviation: (if known)

Sample 2

Column Containing Data: ANSUR II 2012

Population Standard Deviation: (if known)

Please choose a method of analysis below. The NO POOL method is recommended.

Method of Analysis:

● Unequal variances: No Pool
○ Equal variances: Pool
○ Preliminary F-Test

	ANSUR I 1988	ANSUR II 2012	3
1	1698	1810	
2	1727	1850	
3	1734	1777	
4	1684	1811	
5	1667	1780	
6	1680	1733	
7	1785	1814	
8	1885	1861	
9	1841	1709	
10	1702	1740	
11	1738	1694	

```
Not Eq. Vars: No Pool (and df calculated with Formula 9-1)

Using data from columns 1 and 2
Test Statistic, t:  -1.67937
Critical t:         ±1.72744
P-Value:            0.10914

Degrees of freedom: 19.37038

90% Confidence Interval:
-77.86539 < μ1-μ2 < 1.09872
```

9-3 Matched Pairs

Statdisk Procedure for Testing Hypothesis About the Mean of the Differences from Matched Pairs

1. Select **Analysis** in the top menu bar.
2. Select **Hypothesis Testing** in the dropdown menu.
3. Select **Mean Matched Pairs** in the submenu.
4. Make the following entries and selections in the dialog box:
   - Select the desired format in the **Alternative Hypothesis** box.
   - Enter a significance level, such as 0.05 or 0.01.
   - Select the columns of the matched pairs data to be used for the calculations.
5. Click the **Evaluate** button.

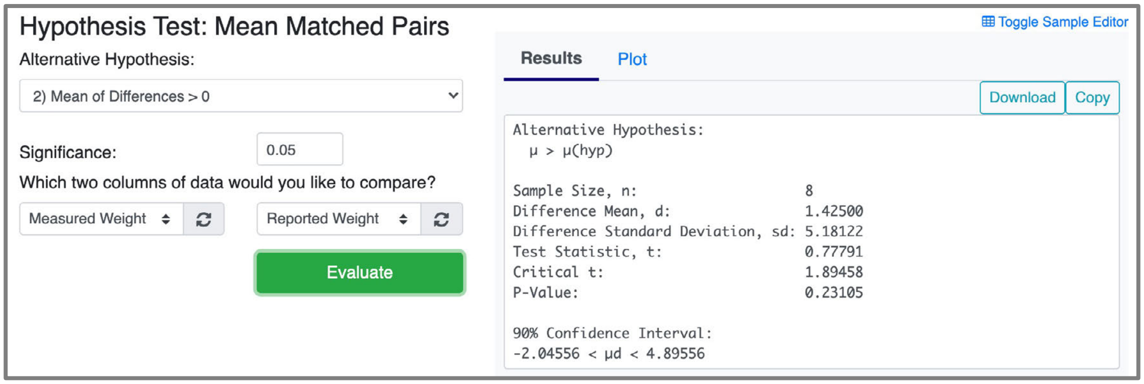
EXAMPLE ***Are People Honest About their Weight?*** Listed below are measured and reported weights (lb) of random male subjects. Use a 0.05 significance level to test the claim that for males, measured weights tend to be higher than the reported weights.

TABLE 9-2 Measured and Reported Weights (lb)

Subject	1	2	3	4	5	6	7	8
Measured Weight (lb)	152.6	149.3	174.8	119.5	194.9	180.3	215.4	239.6
Reported Weight (lb)	150	148	170	119	185	180	224	239

Because the sample is small, we should verify that the sample differences appear to come from a population with a normal distribution. Statdisk's **Normality Assessment** function does suggest that the differences appear to be from a normally distributed population.

Using the Statdisk procedure for matched pairs, we get the following results. The *P*–value of 0.23105 is greater than the significance level of 0.05, so we fail to reject the null hypothesis H_0: $\mu_d = 0$ lb. There is not sufficient evidence to support the claim that for males, the measured weights tend to be higher than reported weights.



Hypothesis Test: Mean Matched Pairs

Alternative Hypothesis: 2) Mean of Differences > 0

Significance: 0.05

Which two columns of data would you like to compare? Measured Weight, Reported Weight

```
Alternative Hypothesis:
  μ > μ(hyp)

Sample Size, n:                    8
Difference Mean, d:                1.42500
Difference Standard Deviation, sd: 5.18122
Test Statistic, t:                 0.77791
Critical t:                        1.89458
P-Value:                           0.23105

90% Confidence Interval:
-2.04556 < μd < 4.89556
```

Note that the previous results also provide the 90% 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. There is not sufficient evidence to conclude that the measured weights tend to be higher than reported weights.

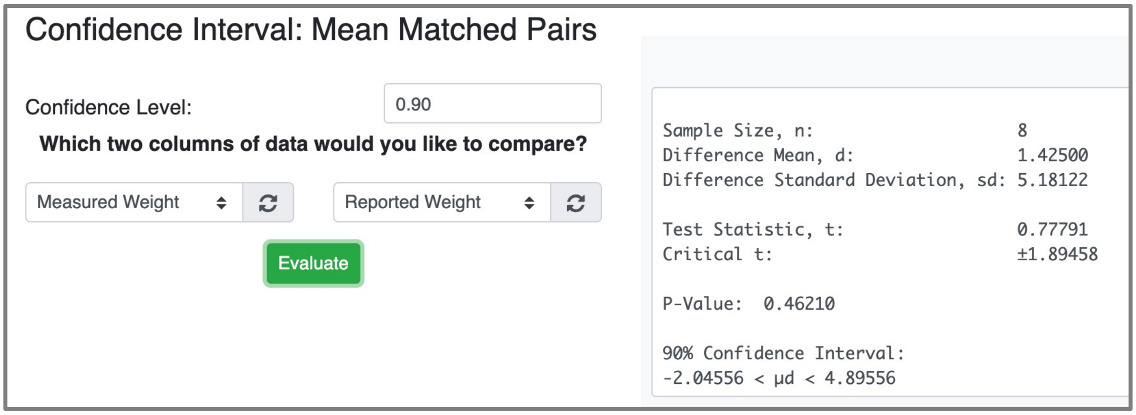
Statdisk can also determine the confidence interval estimate of the mean of the difference from matched pairs using the following procedure. This procedure provides the same confidence interval results as the previous procedure.

Statdisk Procedure for a Confidence Interval Estimate of the Mean of the Differences from Matched Pairs

1. Select **Analysis** in the top menu.
2. Select **Confidence Intervals** in the dropdown menu
3. Select **Mean Matched Pairs** in the submenu.
4. Enter a confidence level and select the columns containing the sample data.
5. Click the **Evaluate** button.

Shown below is the Statdisk display that includes a 90% confidence interval based on the data in Table 9-2 on the previous page.

Note that the confidence interval limits do contain 0, indicating that the true value of μ_d is not significantly different from 0. There is not sufficient evidence to support the claim that for males, the measured weights tend to be higher than reported weights.



Confidence Interval: Mean Matched Pairs

Confidence Level: 0.90

Which two columns of data would you like to compare?

Measured Weight

Reported Weight

```
Sample Size, n:                        8
Difference Mean, d:                    1.42500
Difference Standard Deviation, sd:     5.18122

Test Statistic, t:                     0.77791
Critical t:                            ±1.89458

P-Value:  0.46210

90% Confidence Interval:
-2.04556 < μd < 4.89556
```

9-4 Two Variances or Standard Deviations

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 when using the *P*-value method. (If using the critical value method, the sample with the larger variance must be designated as Sample 1.) When calculating the *P*-value, Statdisk automatically does the required calculations and it correctly handles cases in which the first sample has a variance smaller than the second sample.

Statdisk Procedure for Hypothesis Tests About Two Variances or Two Standard Deviations

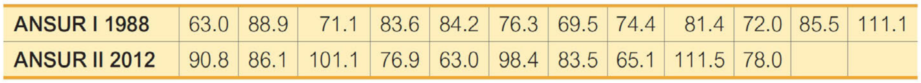
1. Select **Analysis** in the top menu bar.
2. Select **Hypothesis Testing** in the dropdown menu.
3. Select **Standard Deviation Two Samples** in the submenu.
4. Choose the **Use Summary Statistics** tab or **Use Data** tab.

   Using Summary Statistics: Select the **Use Summary Statistics** tab, select the desired format for the alternative hypothesis and enter the significance level. Also enter the sample size (*n*), and sample standard deviation (*s*) for both samples.

   Using Sample Data: Select the **Use Data** tab, select the desired format for the alternative hypothesis and enter the significance level. Also select the desired data column for both samples.
5. Click the **Evaluate** button to obtain the test results.

Consider the following example.

EXAMPLE Listed below are weights (kg) of randomly selected male U.S. Army personnel. Use these data with a 0.05 significance level to test the claim that the variation among weights did not change from the ANSUR I study in 1998 to the ANSUR II study in 2012.

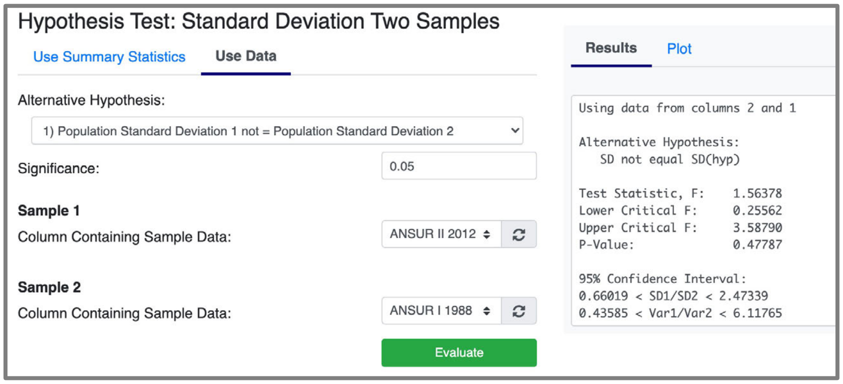


ANSUR I 1988	63.0	88.9	71.1	83.6	84.2	76.3	69.5	74.4	81.4	72.0	85.5	111.1
ANSUR II 2012	90.8	86.1	101.1	76.9	63.0	98.4	83.5	65.1	111.5	78.0		

The Statdisk display on the next page shows a *P*–value of 0.47787. Because that *P*–value is more than the significance level of 0.05, we fail to reject the null hypothesis. There is not sufficient evidence to warrant rejection of the claim that the two standard deviations are equal. It appears the variation among weights of U.S. Army personnel did not change from 1988 to 2012.

The results also provide the test statistic *F* = 1.56378 and the upper and lower critical *F* values. The test statistic *F* = 1.56378 does not fall within the critical region, so we fail to reject the null hypothesis of equal standard deviations.

NOTE: If using the critical value method, the sample with the larger variance must be designated as Sample 1.



Alternative Hypothesis: 1) Population Standard Deviation 1 not = Population Standard Deviation 2

Significance: 0.05

Sample 1
Column Containing Sample Data: ANSUR II 2012

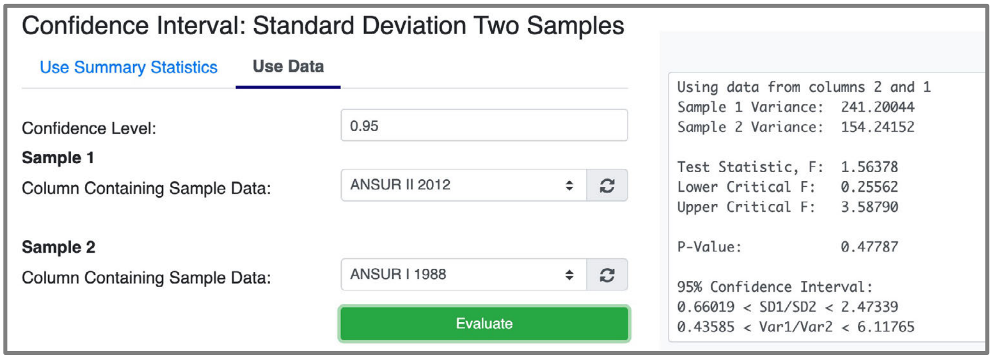
Sample 2
Column Containing Sample Data: ANSUR I 1988

```
Using data from columns 2 and 1

Alternative Hypothesis:
   SD not equal SD(hyp)

Test Statistic, F:    1.56378
Lower Critical F:     0.25562
Upper Critical F:     3.58790
P-Value:              0.47787

95% Confidence Interval:
0.66019 < SD1/SD2 < 2.47339
0.43585 < Var1/Var2 < 6.11765
```

Statdisk can also determine the confidence interval estimate of the ratio of variation for the two samples. This procedure requires that the sample with the larger variance be designated as Sample 1. This procedure provides the same results as the Statdisk procedure for hypothesis tests about two standard deviations or two variances.

Statdisk Procedure for Confidence Interval Estimates of the Ratio of Two Standard Deviations or Variances

1. Select **Analysis** in the top menu bar.
2. Select **Confidence Intervals** in the dropdown menu.
3. Select **Standard Deviation Two Samples** in the submenu.
4. Choose the **Use Summary Statistics** tab or **Use Data** tab and enter the required inputs. (The sample with the larger variance be designated as Sample 1.)
5. Click the **Evaluate** button to obtain the results.



Confidence Level: 0.95

Sample 1
Column Containing Sample Data: ANSUR II 2012

Sample 2
Column Containing Sample Data: ANSUR I 1988

```
Using data from columns 2 and 1
Sample 1 Variance:  241.20044
Sample 2 Variance:  154.24152

Test Statistic, F:  1.56378
Lower Critical F:   0.25562
Upper Critical F:   3.58790

P-Value:            0.47787

95% Confidence Interval:
0.66019 < SD1/SD2 < 2.47339
0.43585 < Var1/Var2 < 6.11765
```

9-5 Inferences with Randomization (Two Samples)

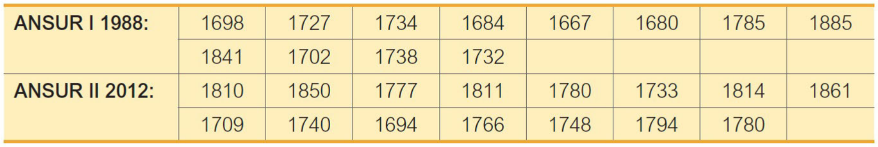
The previous sections of this chapter have all described the use of Statdisk for testing claims about two proportions, two independent means, means of differences from matched pairs and two standard deviations. Another very different approach is to use resampling. Resampling methods include bootstrap resampling and randomization. Here we review the Statdisk resampling procedure for using two samples. See Section 7-7 “Bootstrap Resampling” in this manual for a review of the Statdisk procedure for bootstrap resampling using one sample. See Section 8-4 for a review of the Statdisk procedure for randomization using one sample.

Statdisk Procedure for Bootstrapping and Randomization using Two Samples

1. Select **Resampling** from the top menu bar.
2. Select the desired type of resampling from the dropdown menu. Options include:
   - **Bootstrap Two Proportions**
   - **Bootstrap Two Means**
   - **Randomization Two Proportions**
   - **Randomization Two Means**
   - **Randomization Matched Pairs**
3. Enter the required inputs which include the data to be used and desired number of resamplings.
4. Click the **Evaluate** button.

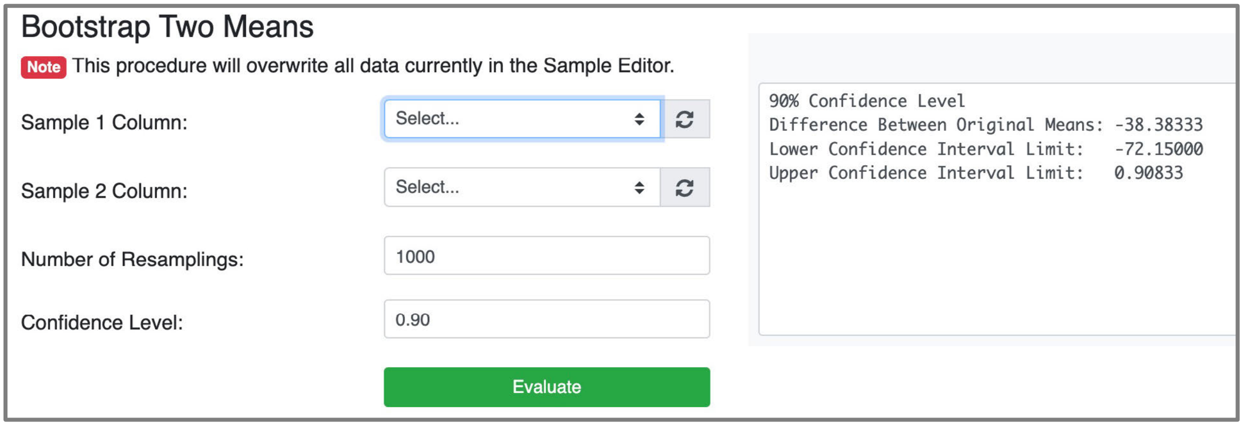
As an illustration, consider the following example.

EXAMPLE ***U.S. Army Heights*** Listed below are heights (mm) of randomly selected U.S. Army male personnel measured in 1988 and different heights (mm) of U.S. Army male personnel in 2012. Use resampling to test the claim that the mean height of the 1988 population is less than the mean height of the 2012 population. Use a 0.05 significance level.



ANSUR I 1988:	1698	1727	1734	1684	1667	1680	1785	1885
	1841	1702	1738	1732				
ANSUR II 2012:	1810	1850	1777	1811	1780	1733	1814	1861
	1709	1740	1694	1766	1748	1794	1780	

Bootstrapping Using the Statdisk function **Bootstrap Two Means** and the two data samples, the below results are obtained. (A 90% confidence interval is required for this left-tailed case.) The resulting confidence interval of -72.15 mm < $\mu_1 - \mu_2$ < 0.09 mm is typical and does include 0, so it appears there is not a significant difference between the mean height in 1998 and the mean height in 2012. We conclude there is not sufficient evidence to support that claim that the mean height in 1988 is less than the mean height in 2012.



Bootstrap Two Means

Note This procedure will overwrite all data currently in the Sample Editor.

Sample 1 Column: Select...

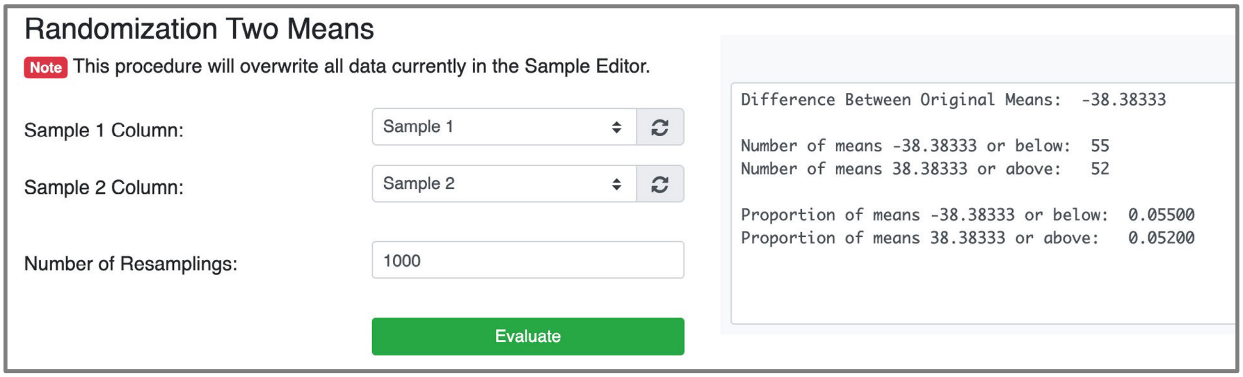
Sample 2 Column: Select...

Number of Resamplings: 1000

Confidence Level: 0.90

```
90% Confidence Level
Difference Between Original Means: -38.38333
Lower Confidence Interval Limit:   -72.15000
Upper Confidence Interval Limit:   0.90833
```

Randomization Using the Statdisk function **Randomization Two Means** and the two data samples, the below results are obtained. The difference in means between the two samples is -38.4 mm (rounded) and randomization finds that the difference of -38.4 mm or below occurs 55 times with 1000 resamplings. This is a typical result. This represents a proportion of 0.055, which is greater than the significance level of 0.05. We again conclude there is not sufficient evidence to support that claim that the mean height in 1988 is less than the mean height in 2012.



Randomization Two Means

Note This procedure will overwrite all data currently in the Sample Editor.

Sample 1 Column: Sample 1

Sample 2 Column: Sample 2

Number of Resamplings: 1000

```
Difference Between Original Means:  -38.38333

Number of means -38.38333 or below:  55
Number of means 38.38333 or above:   52

Proportion of means -38.38333 or below:  0.05500
Proportion of means 38.38333 or above:   0.05200
```

NOTE: The bootstrapping and randomization results above are both close to the borderline between supporting and failing to support the claim that the mean height in 1988 is less than the mean height in 2012. The sample data do not provide *compelling* evidence in favor or against that claim.

CHAPTER 9 WORKBOOK: Inferences from Two Samples

9-1 **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?

____

____

9-2 **Fast Food** Refer to *Elementary Statistics* 14th Edition, Data Set 36 "Fast Food." Use a 0.05 significance level to test the claim that the sample of drive through service times for McDonald's dinner and Burger King dinner are from populations with the same mean. If there is a statistically significant difference, does that difference have practical significance? (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.)

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

Conclusion in your own words: ____

____

9-3 **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

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

Conclusion in your own words: ____

____

9-4 **Freshman 15 Study** Use the sample weights (kg) of male and female college students measured in April of their freshman year, as listed in *Elementary Statistics* 14th Edition, Data Set 13 "Freshman 15." 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: ____

____

10 Correlation and Regression

Statdisk

10-1 Correlation and Regression

Sections 10-1, 10-2, and 10-3 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.

Given a collection of paired data, Statdisk can find the linear correlation coefficient *r* and 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, open a Statdisk data set, or upload a data set to Statdisk.
2. Select **Analysis** in the top menu bar.
3. Select **Correlation and Regression** in the dropdown menu.
4. Select a significance level, such as 0.05 or 0.01.
5. 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 generate a graph of the scatterplot.
8. Click the **Residual Plot** button to generate a graph of residuals versus *x*.

Consider the following example.

EXAMPLE: ***Eat More Chocolate and Win a Nobel Prize?*** *Elementary Statistics* 14th Edition, Data Set 23 "Nobel Laureates and Chocolate" includes chocolate consumption (kg per capita) and the numbers of Nobel Laureates (per 10 million people) for twenty-three different countries. Is there is a relationship between chocolate consumption and number of Nobel Laureates? If such a relationship exists, identify an equation so that we can predict the number of Nobel Laureates based on chocolate consumption.

Correlation

If you follow the above steps using *Elementary Statistics* 14th Edition, Data Set 23 "Nobel Laureates and Chocolate," the Statdisk results will be as shown on the next page. The column "CHOCOLATE" is selected for the *x* variable and the column "NOBEL" is selected for the *y* variable. Also shown is the scatterplot that is obtained by clicking the **Scatterplot** button and the residual plot obtained by clicking the **Residual Plot** button. The results include the linear correlation coefficient of $r = 0.80061$, the critical values of $r = \pm 0.41325$, and the *P*-value of 0.00000. Based on these results, there does appear to be a linear correlation between chocolate consumption and numbers of Nobel Laureates. (*Remember:* correlation is not causation and it is always important to use common sense in interpreting correlation results. It does not make logical sense that there is a relationship between chocolate consumption and number of Nobel laureates.)

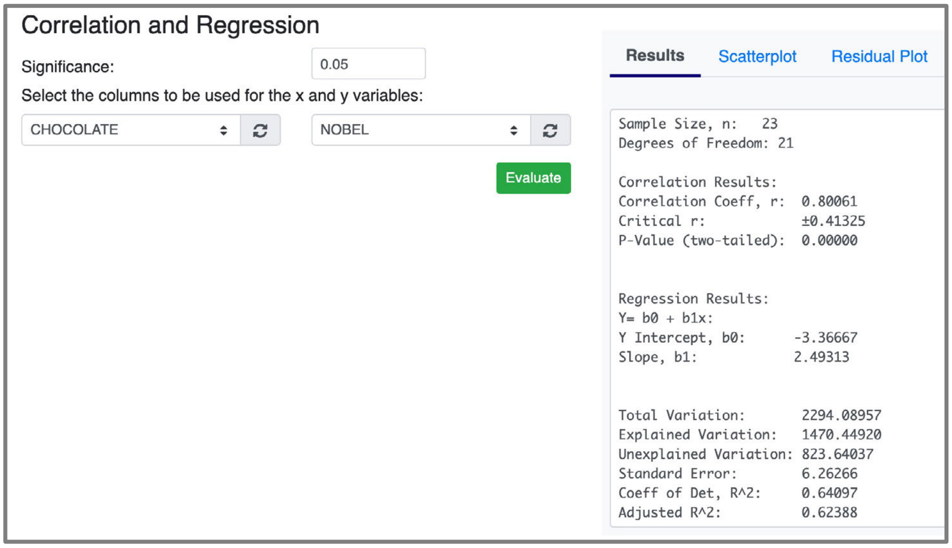
Regression

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

$$\hat{y} = -3.37 + 2.49x$$

The graph of the scatterplot 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.



Correlation and Regression

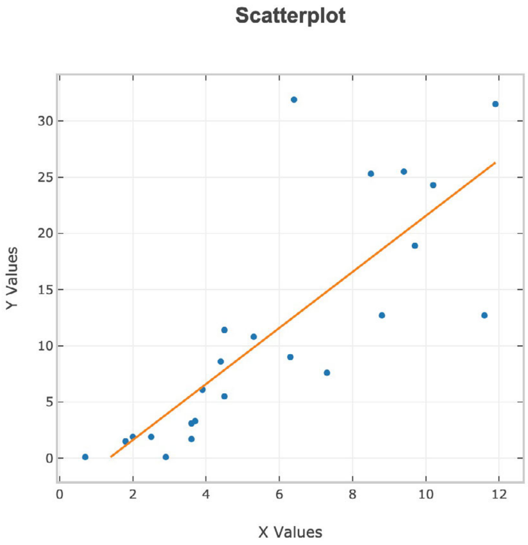
Significance: 0.05

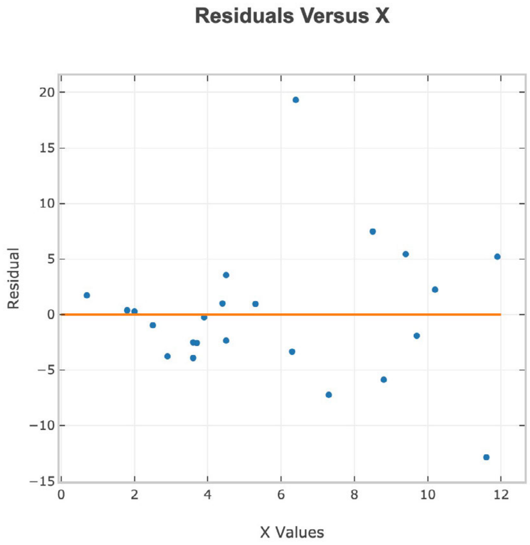
Select the columns to be used for the x and y variables:

CHOCOLATE

NOBEL

```
Sample Size, n:   23
Degrees of Freedom: 21

Correlation Results:
Correlation Coeff, r:  0.80061
Critical r:            ±0.41325
P-Value (two-tailed):  0.00000


Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      -3.36667
Slope, b1:            2.49313


Total Variation:       2294.08957
Explained Variation:   1470.44920
Unexplained Variation: 823.64037
Standard Error:        6.26266
Coeff of Det, R^2:     0.64097
Adjusted R^2:          0.62388
```





10-2 Multiple Regression

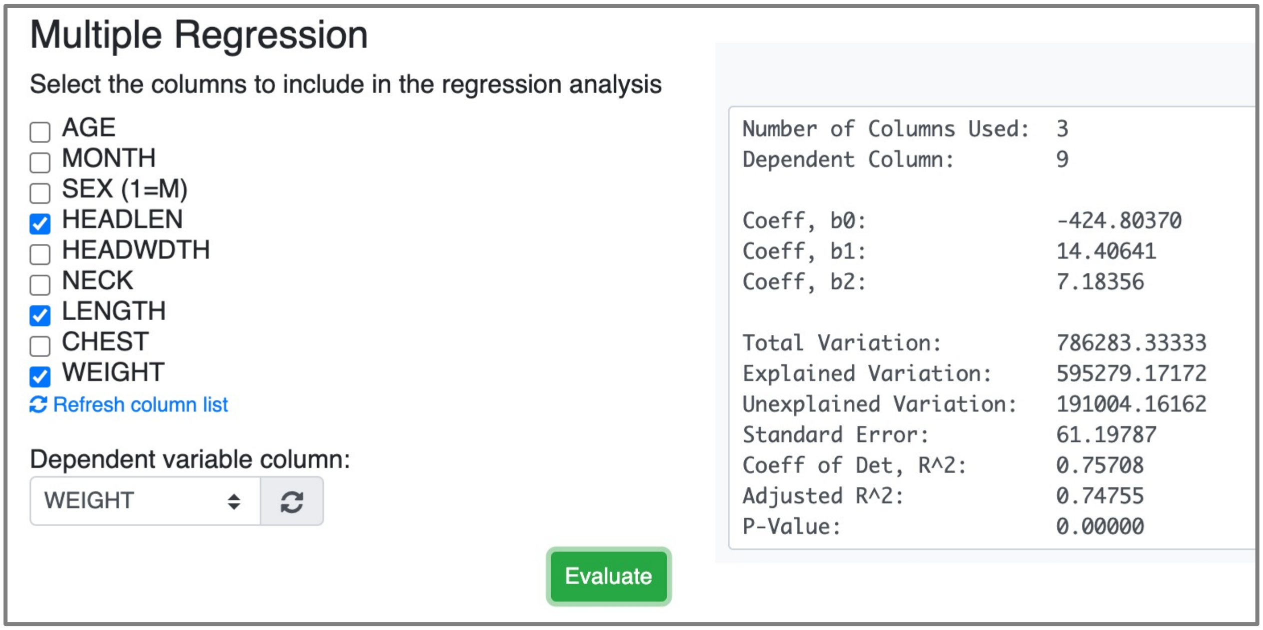
Section 10-4 of the textbook discusses multiple regression, and Statdisk does provide multiple regression results. Once a collection of sample data has been entered, you can easily experiment with different combinations of data columns to find the combination that is best.

Statdisk Procedure for Multiple Regression

1. Either manually enter the data (if the lists are not very long), open a Statdisk data set, or upload a data set to Statdisk
2. Select **Analysis** in the top menu bar.
3. Select **Multiple Regression** in the dropdown menu.
4. Select the columns to be included, and then identify the column to be used for the dependent variable. (If a column has a checkmark (✓) but you don't want it included, click the checkmark so that it is removed and the column is excluded.) See the display on the next page, where columns named HEADLEN, LENGTH, and WEIGHT are included, with WEIGHT selected for the dependent variable.
5. Click **Evaluate**.
6. To use a different combination of variables, simply click different combinations of columns.

EXAMPLE Consider the Statdisk *Elementary Statistics* 14th Edition, Data Set 18 "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 on the next page, note the selection of columns on the left of the display (HEADLEN, LENGTH, WEIGHT), and note the selection of the column WEIGHT as the dependent variable.



Multiple Regression

Select the columns to include in the regression analysis

☐ AGE
☐ MONTH
☐ SEX (1=M)
☑ HEADLEN
☐ HEADWDTH
☐ NECK
☑ LENGTH
☐ CHEST
☑ WEIGHT

Dependent variable column: WEIGHT

```
Number of Columns Used:   3
Dependent Column:         9

Coeff, b0:                -424.80370
Coeff, b1:                14.40641
Coeff, b2:                7.18356

Total Variation:          786283.33333
Explained Variation:      595279.17172
Unexplained Variation:    191004.16162
Standard Error:           61.19787
Coeff of Det, R^2:        0.75708
Adjusted R^2:             0.74755
P-Value:                  0.00000
```

The results in the Statdisk display include the intercept b_0 = -425 (rounded), the coefficient b_1 = 14.4 (rounded), and the coefficient b_2 = 7.18 (rounded). These values are included in the multiple regression equation as shown here:

$$\hat{y} = -425 + 14.4x_1 + 7.18x_2$$

or WEIGHT = -425 + 14.4 HEADLEN + 7.18 LENGTH

The results also include the adjusted coefficient of determination (Adjusted R^2 = 0.74755), as well as the P-value of 0.00000.

10-3 Nonlinear Regression

Nonlinear regression is discussed in the Triola statistics textbooks (except *Essentials of Statistics* and *Biostatistics for the Biological and Health Sciences*). 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 below. We use the coded year values for *x*, so *x* = 1, 2, 3, . . . , 12. The *y* values are the populations (in millions) of 5, 10, 17, . . . , 335.

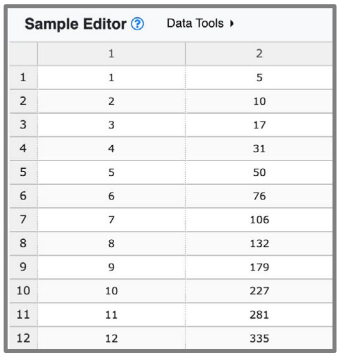
TABLE 10-7 Population (in millions) of the United States

Year	1800	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000	2020
Coded Year	1	2	3	4	5	6	7	8	9	10	11	12
Population	5	10	17	31	50	76	106	132	179	227	281	335

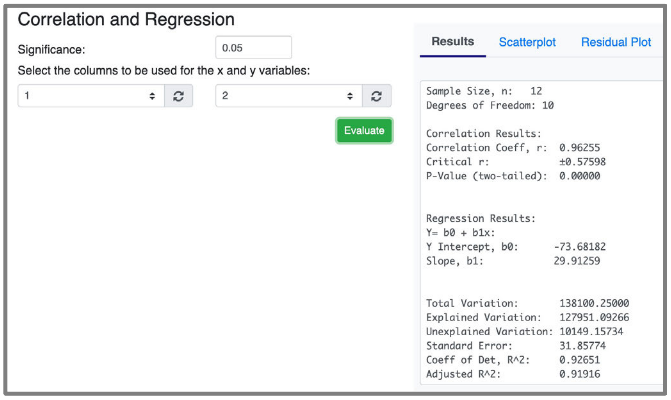
Linear Model: *y* = *a* + *bx*

The linear model can be obtained by using Statdisk's correlation and regression function.

1. For the data in Table 10-7, enter the coded year values of 1, 2, 3, . . . , 12 in the first column of the Statdisk Sample Editor, and enter the population values of 5, 10, 17, . . . , 335 in the second column.
2. Use the procedure described in Section 10-1 of this manual. Select **Analysis** from the Statdisk menu, then **Correlation and Regression**. The result will be as shown below.



	1	2
1	1	5
2	2	10
3	3	17
4	4	31
5	5	50
6	6	76
7	7	106
8	8	132
9	9	179
10	10	227
11	11	281
12	12	335



Correlation and Regression

Significance: 0.05

Select the columns to be used for the x and y variables: 1 2

```
Sample Size, n:   12
Degrees of Freedom: 10

Correlation Results:
Correlation Coeff, r:  0.96255
Critical r:            ±0.57598
P-Value (two-tailed):  0.00000


Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      -73.68182
Slope, b1:            29.91259


Total Variation:       138100.25000
Explained Variation:   127951.09266
Unexplained Variation: 10149.15734
Standard Error:        31.85774
Coeff of Det, R^2:     0.92651
Adjusted R^2:          0.91916
```

The above Statdisk display describes key results for the linear model. The resulting function is

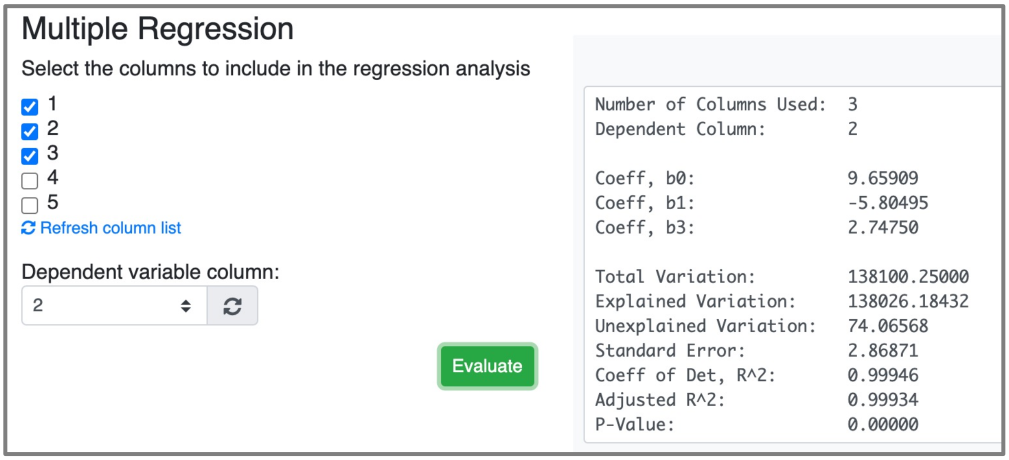
$$y = -73.7 + 29.9x$$

The coefficient of determination is displayed as r^2 = 0.92651. The high value of r^2 suggests that the linear model is a reasonably good fit.

Quadratic Model: $y = ax^2 + bx + c$

The quadratic model can be obtained by using Statdisk's multiple regression function.

1. For the data in Table 10-7, enter the coded year values (x) of 1, 2, 3, . . . , 12 in the first column of the Statdisk Sample Editor and enter the population values (y) of 5, 10, 17, . . , 335 in the second column. In the third column, enter the x^2 values of 1, 4, 9,…,144.
   - 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** in the top menu, and then select **Sample Transformations**).
2. Select **Analysis** from the Statdisk menu, 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.



Multiple Regression

Select the columns to include in the regression analysis

☑ 1
☑ 2
☑ 3
☐ 4
☐ 5

Dependent variable column: 2

```
Number of Columns Used:  3
Dependent Column:        2

Coeff, b0:               9.65909
Coeff, b1:               -5.80495
Coeff, b3:               2.74750

Total Variation:         138100.25000
Explained Variation:     138026.18432
Unexplained Variation:   74.06568
Standard Error:          2.86871
Coeff of Det, R^2:       0.99946
Adjusted R^2:            0.99934
P-Value:                 0.00000
```

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

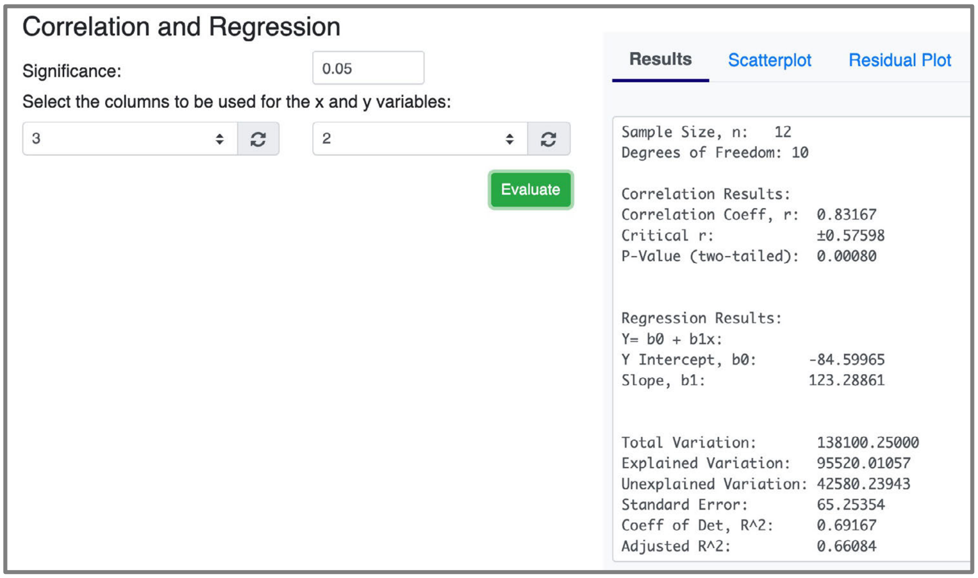
$$y = 9.66 - 5.80x + 2.75x^2$$

(The coefficient b_1 corresponds to x and b_3 corresponds to x^2.) The coefficient of determination is given by R^2 = 0.99946, suggesting a better fit than the linear model (which has R^2 =0.92651).

Logarithmic Model: *y* = *a* + *b* ln *x*

The logarithmic model can be obtained by using Statdisk's correlation and regression function

1. For the data in Table 10-7, enter the coded year values (*x*) of 1, 2, 3, . . . , 12 in the first column of the Statdisk Sample Editor and enter the population values (*y*) of 5, 10, 17, . . , 335 in the second column. In the third column, enter the ln *x* values of 0, 0.6931472, 1.098612,..,2.4849066.
    - 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** in the top menu, and then select **Sample Transformations**).
2. Select **Analysis** from the Statdisk menu, 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.


Correlation and Regression
Significance:
0.05
Select the columns to be used for the x and y variables:
3
2
Evaluate
Results
Scatterplot
Residual Plot

```
Sample Size, n:   12
Degrees of Freedom: 10

Correlation Results:
Correlation Coeff, r:  0.83167
Critical r:            ±0.57598
P-Value (two-tailed):  0.00080


Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      -84.59965
Slope, b1:            123.28861


Total Variation:       138100.25000
Explained Variation:   95520.01057
Unexplained Variation: 42580.23943
Standard Error:        65.25354
Coeff of Det, R^2:     0.69167
Adjusted R^2:          0.66084
```

The display shown above results from the logarithmic model used with the sample data in Table 10-7. The function is given by

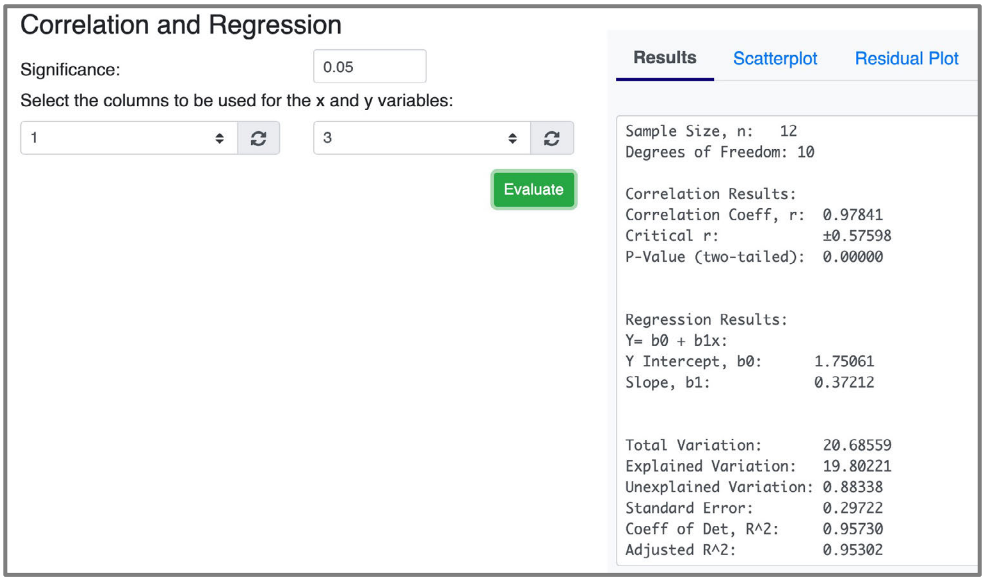
$$\boldsymbol{y = -84.6 + 123 \ln x}$$

Also, R^2 = 0.69167, 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).

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, . . . , 12 in the first column of the Statdisk Sample Editor and enter the population values (*y*) of 5, 10, 17, . . , 335 in the second column. In the third column, enter the ln *y* values of 1.609438, 2.302585, 2.833213,..,5.81413.
   - 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** in the top menu, and 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.


Correlation and Regression
Significance:
Select the columns to be used for the x and y variables:
Evaluate
Results
Scatterplot
Residual Plot

Significance: 0.05

x column: 1

y column: 3

```
Sample Size, n:   12
Degrees of Freedom: 10

Correlation Results:
Correlation Coeff, r:  0.97841
Critical r:            ±0.57598
P-Value (two-tailed):  0.00000


Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      1.75061
Slope, b1:            0.37212


Total Variation:       20.68559
Explained Variation:   19.80221
Unexplained Variation: 0.88338
Standard Error:        0.29722
Coeff of Det, R^2:     0.95730
Adjusted R^2:          0.95302
```

The value of R^2 = 0.95730 is OK as is, but the values of *a* and *b* must be computed from the Statdisk results as shown below:

$$a = e^{b0} = e^{1.75061} = 5.7581$$
$$b = e^{b1} = e^{0.37212} = 1.4508$$

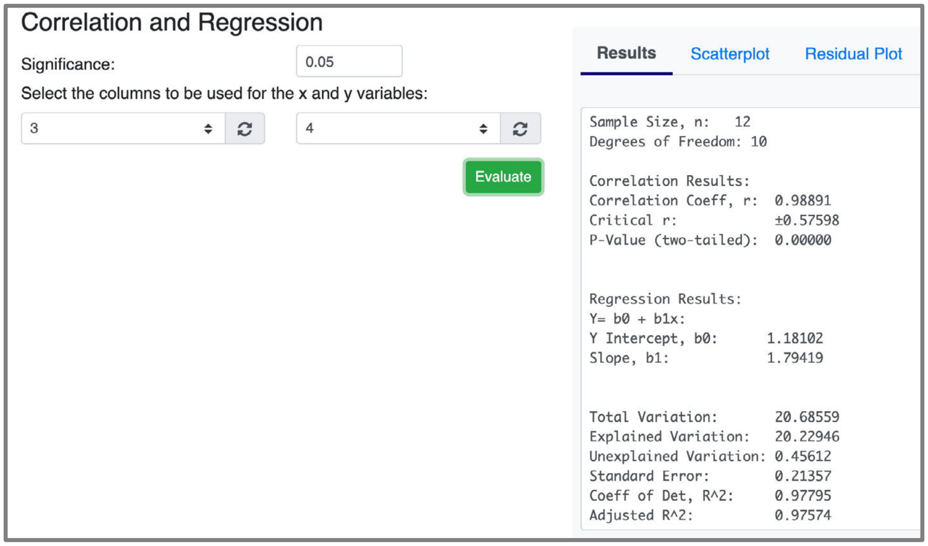
Using these values of *a* and *b*, we express the exponential model as

$$\mathbf{y = 5.76(1.45^x)}$$

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, . . . , 12 in the first column of the Statdisk Sample Editor and enter the population values (*y*) of 5, 10, 17, . . , 335 in the second column. In the third column, enter the ln *x* values of 0, 0.693147, 1.098612,..,2.484907. 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** in the top menu, and 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.


Correlation and Regression
Significance:
0.05
Select the columns to be used for the x and y variables:
3
4
Evaluate
Results
Scatterplot
Residual Plot

```
Sample Size, n:   12
Degrees of Freedom: 10

Correlation Results:
Correlation Coeff, r:  0.98891
Critical r:            ±0.57598
P-Value (two-tailed):  0.00000


Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      1.18102
Slope, b1:            1.79419


Total Variation:       20.68559
Explained Variation:   20.22946
Unexplained Variation: 0.45612
Standard Error:        0.21357
Coeff of Det, R^2:     0.97795
Adjusted R^2:          0.97574
```

The value of $R^2 = 0.97795$ 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.

$$a = e^{b0} = e^{1.18102} = 3.2550$$
$$b = b_1 \text{ from Statdisk} = 1.79419$$

Using these values of *a* and *b*, we express the power model as

$$\boldsymbol{y = 3.26(x^{1.79})}$$

CHAPTER 10 WORKBOOK: Correlation and Regression

10-1 **Bear Weights and Chest Sizes** Refer to *Elementary Statistics*, 14th Edition, Data Set 18 "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?________

10-2 **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. ____________

10-3 **Statdisk Data Set: Predicting IQ Score** Refer to *Elementary Statistics* 14th Edition, Data Set 12 "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?

____

____

____

11 Goodness-of-Fit and Contingency Tables

Statdisk

11-1 Goodness-of-Fit

In Section 11-1 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** in the top menu.
3. Select **Goodness-of-Fit** in the dropdown menu.
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 **Plot** to obtain a graph of the χ^2 distribution that includes the test statistic and critical value.

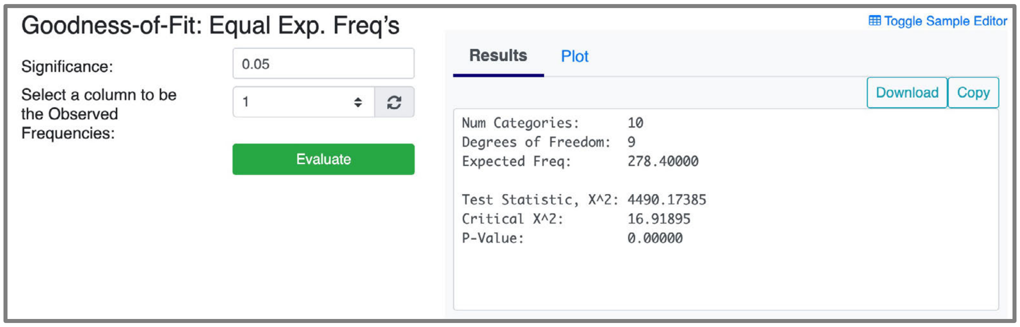
Consider the following example.

> **EXAMPLE** The Statdisk *Elementary Statistics* 14th Edition, Data Set 4 “Measured and Reported” includes weights of subjects that were both measured and reported. Table 11-2 below includes the reported weights of 2784 males. 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.

As in Section 11–1 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.

TABLE 11-2 Last Digits of Weights of Males

Last Digit	0	1	2	3	4	5	6	7	8	9
Frequency	1175	44	169	111	112	731	96	110	171	65


Goodness-of-Fit: Equal Exp. Freq's

Significance: 0.05

Select a column to be the Observed Frequencies: 1

```
Num Categories:      10
Degrees of Freedom:  9
Expected Freq:       278.40000

Test Statistic, X^2: 4490.17385
Critical X^2:        16.91895
P-Value:             0.00000
```

The *P*-value of 0.00000 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.

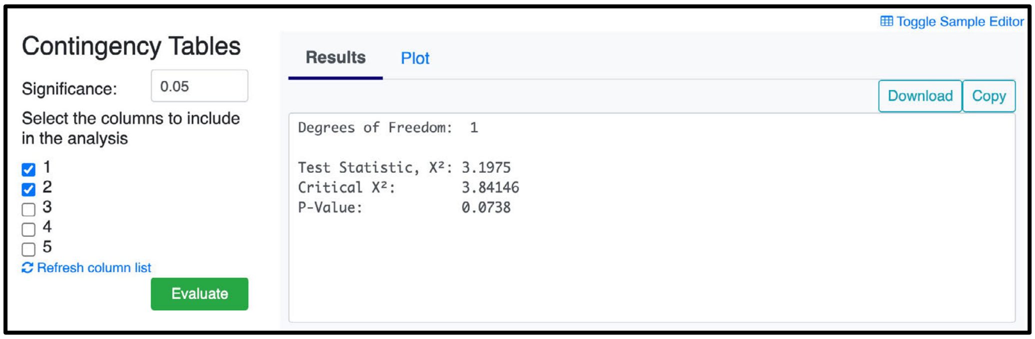
TABLE 11-6 Results From a Study of a Link Between the MMR Vaccine and Autism

	Unvaccinated	Vaccinated
Autism	25 (E = 18.340)	64 (E = 70.660)
No Autism	362 (E = 368.660)	1427 (E = 1420.340)

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** in the top menu.
3. Select **Contingency Tables** in the dropdown menu.
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 **Evaluate**.
7. Click **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 0.0738 is greater than the 0.05 significance level, so we fail to reject the null hypothesis of independence between the row and column variables. It appears that autism is not linked to the MMR vaccine.



```
Degrees of Freedom:  1

Test Statistic, X²: 3.1975
Critical X²:        3.84146
P-Value:            0.0738
```

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 on the next page, 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

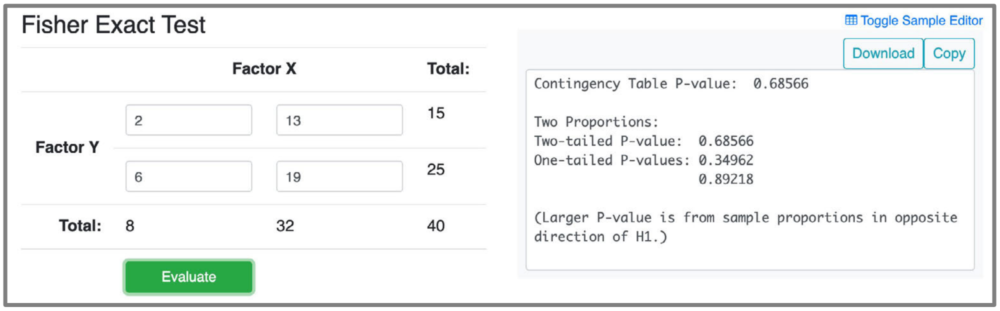
1. Select **Analysis** in the top menu.
2. Select **Fisher Exact Test** in the dropdown menu.
3. In the input box that appears, enter the four frequencies in the cells of the table. (Enter a frequency, press the **Tab** key, enter another frequency, and so on. Do not try to enter row and column totals; they will be provided by Statdisk.
   - Be careful to enter the four sample frequencies, not the *expected* frequencies.
4. Click **Evaluate**.

Consider the sample data in the table on the next page. 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.



Fisher Exact Test

	Factor X		Total:
Factor Y	2	13	15
	6	19	25
Total:	8	32	40

```
Contingency Table P-value:  0.68566

Two Proportions:
Two-tailed P-value:  0.68566
One-tailed P-values: 0.34962
                     0.89218

(Larger P-value is from sample proportions in opposite
direction of H1.)
```

11-4 McNemar's Test for Matched Pairs

The contingency table procedures in Section 11-2 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** in the top menu.
2. Select **McNemar's Test** in the dropdown menu.
3. In the dialog box that appears, enter the four frequencies in the cells of the table. (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 **Evaluate**.

Consider the following example.

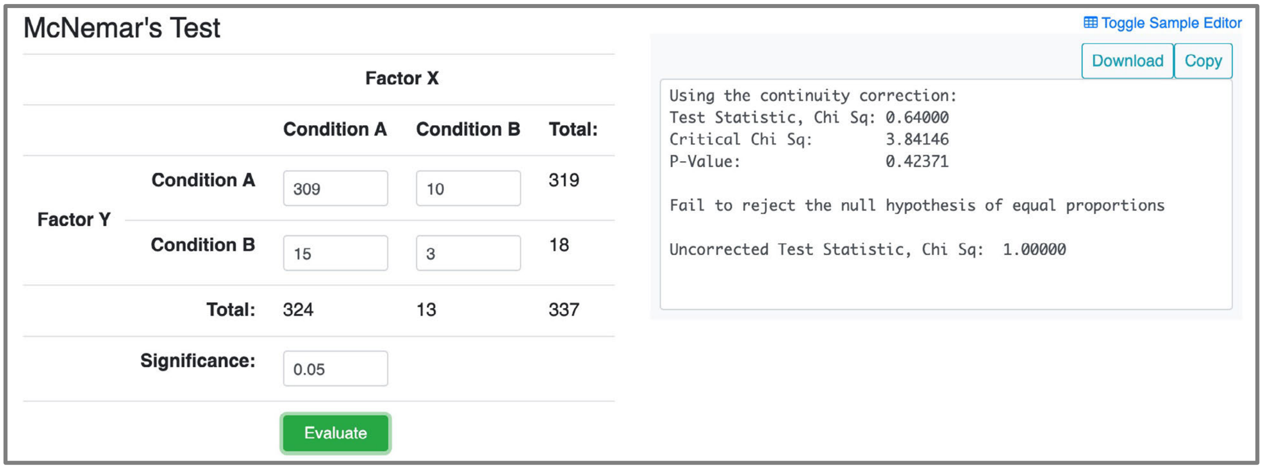
EXAMPLE ***Are Hip Protectors Effective?*** A randomized controlled trial was designed to test the effectiveness of hip protectors in preventing hip fractures in the elderly. Nursing home residents each wore protection on one hip, but not the other. Results are summarized in Table 11-10 (based on data 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.

Using the Statdisk procedure and data from Table 11-10 the following results are obtained.

TABLE 11-10 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



McNemar's Test

		Factor X		
		Condition A	Condition B	Total:
Factor Y	Condition A	309	10	319
	Condition B	15	3	18
	Total:	324	13	337
	Significance:	0.05		

```
Using the continuity correction:
Test Statistic, Chi Sq: 0.64000
Critical Chi Sq:        3.84146
P-Value:                0.42371

Fail to reject the null hypothesis of equal proportions

Uncorrected Test Statistic, Chi Sq:  1.00000
```

The Statdisk results include the chi-square test statistic, the critical value, and the *P*-value. Because the *P*-value of 0.42371 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.

CHAPTER 11 WORKBOOK: 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 **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:____

Now test the above claim by using Fisher's exact test instead of using the approximating chi-square distribution. Enter the results below.

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?

____

____

11-3 **Treating Athlete's Foot** Assume that subjects are afflicted with athlete's foot on each of their two 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 McNemar's test and 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:____

____

12 Analysis of Variance

12-1 One-Way Analysis of Variance

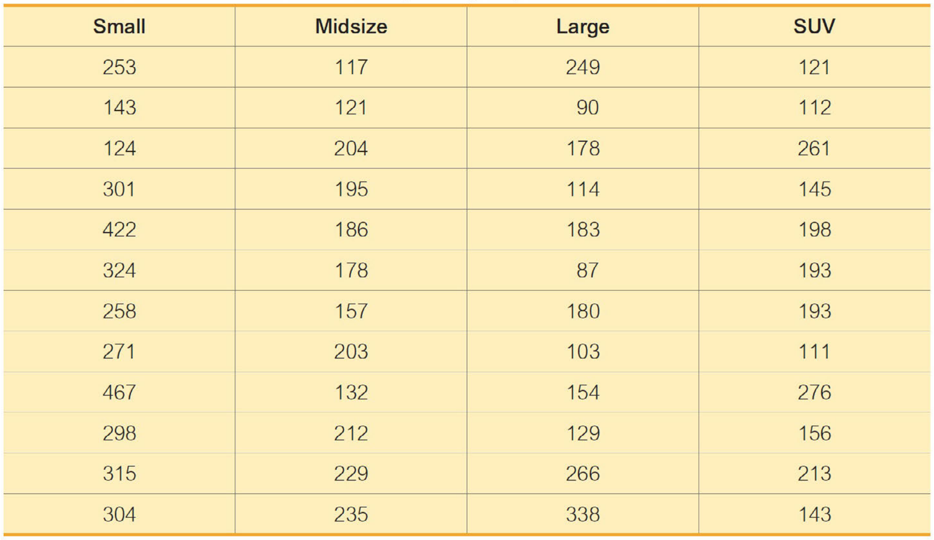
12-2 Two-Way Analysis of Variance

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 Table 12-1 shown below, there is one factor used to categorize the data: car size category. This data is from *Elementary Statistics* 14th Edition, Data Set 35 “Car Data”.

TABLE 12-1 Measurements of Head Injuries (HIC) in Car Crash Tests



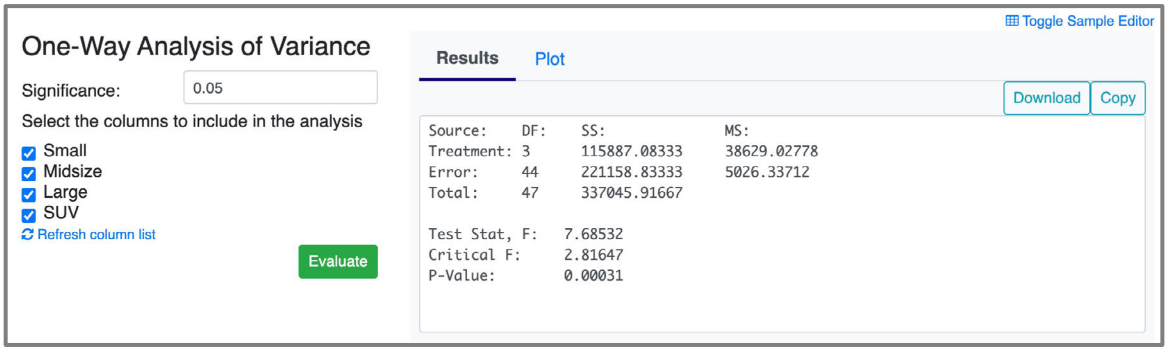
Small	Midsize	Large	SUV
253	117	249	121
143	121	90	112
124	204	178	261
301	195	114	145
422	186	183	198
324	178	87	193
258	157	180	193
271	203	103	111
467	132	154	276
298	212	129	156
315	229	266	213
304	235	338	143

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.

Statdisk Procedure for One-Way Analysis of Variance

1. Enter the data in separate columns of the Statdisk Sample Editor.
2. Select **Analysis** in the top menu.
3. Select **One-Way Analysis of Variance** in the dropdown menu.
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 checkmark (✓) and you do not want to include it, click the box to remove the checkmark. If a box does not have a checkmark and you want to include it, click the box to make the checkmark appear.
6. Click **Evaluate**.
7. Click **Plot** to obtain a graph that includes the critical value and test statistic.

If you use the above steps with the "Car Data" data set, and include all vehicle sizes in the analysis, the Statdisk result will appear as follows.



One-Way Analysis of Variance

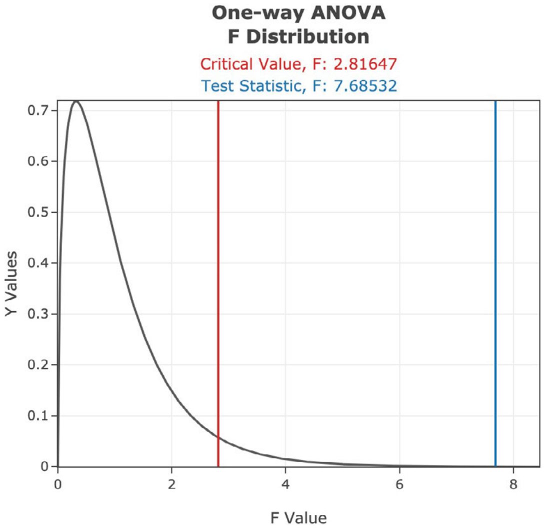
Significance: 0.05

Select the columns to include in the analysis

☑ Small
☑ Midsize
☑ Large
☑ SUV

```
Source:    DF:    SS:             MS:
Treatment: 3      115887.08333    38629.02778
Error:     44     221158.83333    5026.33712
Total:     47     337045.91667

Test Stat, F:   7.68532
Critical F:     2.81647
P-Value:        0.00031
```

The *P*-value of 0.00031 is less than the significance level of 0.05, so we reject the null hypothesis that the means are equal. On the basis of this ANOVA test, we cannot conclude that any particular mean is different from the others, but we can informally note that the sample mean for small cars is higher than the mean for the midsize, large, and SUV vehicles. The test statistic of *F* = 7.68532 is also provided along with the critical value of *F* = 2.81647. The values of the SS and MS components are also provided. Click the **Plot** button to generate a graph showing the test statistic and critical value.

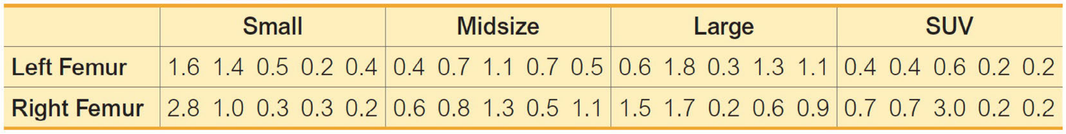


Caution: It is easy to feed Statdisk (or any other software package) data that can be processed quickly and painlessly, but we should *think* about what we are doing. We should consider the assumptions for the test being used, and we should *explore* the data before jumping into a formal procedure such 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 vehicle size (small, midsize, large SUV) and femur side (left, right) as shown in Table 12-3. 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.

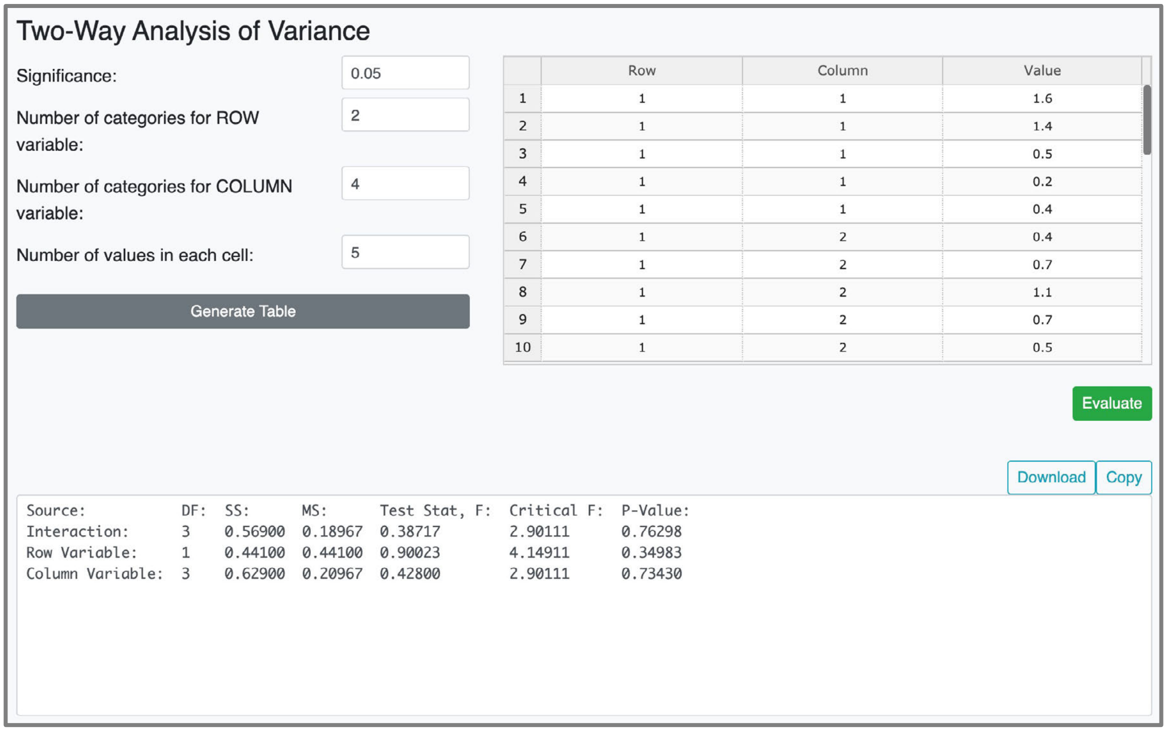
TABLE 12-3 Crash Test Force on Femur with Two Factors: Femur Side and Vehicle Size Category



	Small	Midsize	Large	SUV
Left Femur	1.6 1.4 0.5 0.2 0.4	0.4 0.7 1.1 0.7 0.5	0.6 1.8 0.3 1.3 1.1	0.4 0.4 0.6 0.2 0.2
Right Femur	2.8 1.0 0.3 0.3 0.2	0.6 0.8 1.3 0.5 1.1	1.5 1.7 0.2 0.6 0.9	0.7 0.7 3.0 0.2 0.2

Statdisk Procedure for Two-Way Analysis of Variance

1. Select **Analysis** in the top menu bar.
2. Select **Two-Way Analysis of Variance** in the dropdown menu.
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 (Left Femur, Right Femur), enter **4** for the number of categories of the column variable (Small, Midsize, Large, SUV), and enter **5** for the number of values in each cell.
   - Click **Generate Table** 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 Refer to the Statdisk display on the next page to see how the sample values from Table 12-3 are entered.
6. Click **Evaluate** after all sample values have been entered.



Two-Way Analysis of Variance

Significance: 0.05

Number of categories for ROW variable: 2

Number of categories for COLUMN variable: 4

Number of values in each cell: 5

	Row	Column	Value
1	1	1	1.6
2	1	1	1.4
3	1	1	0.5
4	1	1	0.2
5	1	1	0.4
6	1	2	0.4
7	1	2	0.7
8	1	2	1.1
9	1	2	0.7
10	1	2	0.5

```
Source:           DF:  SS:      MS:      Test Stat, F:  Critical F:  P-Value:
Interaction:      3    0.56900  0.18967  0.38717        2.90111      0.76298
Row Variable:     1    0.44100  0.44100  0.90023        4.14911      0.34983
Column Variable:  3    0.62900  0.20967  0.42800        2.90111      0.73430
```

See Section 12–2 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, *Interaction* has a *P*-value of 0.76298. Because the *P*-value is greater than the significance level of 0.05, we fail to reject the null hypothesis of no interaction between the two factors. It does not appear that femur crash force measurements are affected by an interaction between femur side (Left/Right) and vehicle size category (Small, Midsize, Large, SUV). 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 (femur side). The corresponding *P*-value is shown in the Statdisk display as 0.34983. Because that *P*-value is greater than the significance level of 0.05, we fail to reject the null hypothesis of no effects from femur side. That is, the car crash force measurements do not appear to be affected by whether the femur is in the left leg or right leg.

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 (vehicle size category). The corresponding *P*-value is shown in the Statdisk display as 0.7343. Because that P-value is not less than the significance level of 0.05, we fail to reject the null hypothesis of no effects from vehicle size category. The femur crash force measurements do not appear to be affected by the size of the vehicle.

CHAPTER 12 WORKBOOK: Analysis of Variance

12-1 **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:____

12-2 ***Pulse Rate*** The following table lists pulse rates obtained from *Elementary Statistics* 14th Edition, Data Set 1 “Body Data”. 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.

____

____

13 Nonparametric Tests

13-1 Nonparametric Methods

13-2 Sign Test

13-3 Wilcoxon Signed-Ranks Test

13-4 Wilcoxon Rank-Sum Test

13-5 Kruskal-Wallis Test

13-6 Rank Correlation

13-7 Runs Test for Randomness

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.

13-2 Sign Test

The Triola textbook (excluding *Essentials of Statistics*) describes the sign test, and the following definition is given.

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, and then we illustrate this procedure with an example.

Statdisk Procedure for the Sign Test

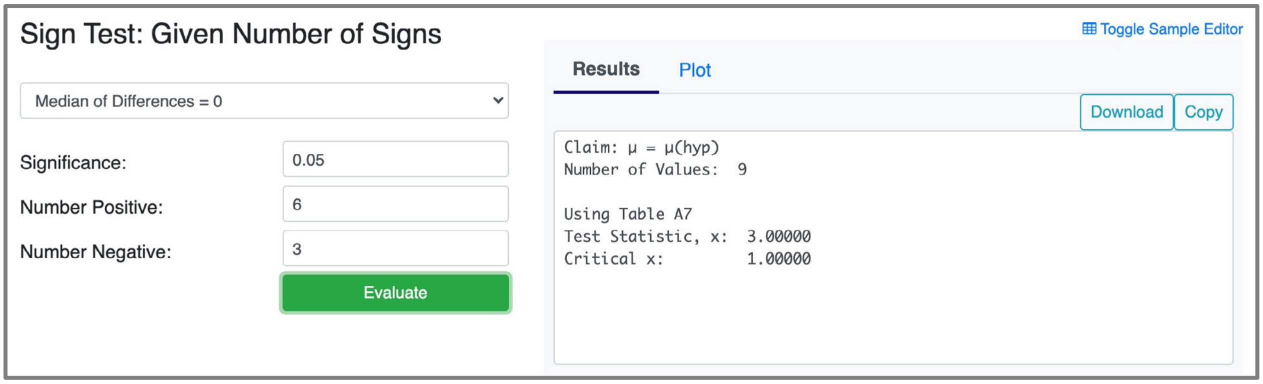
1. Either determine the number of positive signs and the number of negative signs, or enter the paired data in columns of the Statdisk Sample Editor.
2. Select **Analysis** in the top menu bar.
3. Select **Sign Test** in the dropdown menu.
4. If the number of positive and negative signs are known (as in cases involving nominal data), select **Given Number of Signs**. If using *paired* sample data, select **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 enter a significance level, such as 0.05 or 0.01. You must then enter the numbers of positive and negative signs, or select the columns containing the original pairs of data.
6. Click the **Evaluate** button.
    - Click **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. (The data are matched, because each pair of values is from the same year.) Use the sign test to test the claim that there is no difference measured and reported male weights

TABLE 13-3 Measured and Reported Male Weights

Measured	220.0	268.7	213.4	201.3	107.1	172.0	187.4	132.5	122.1	151.9
Reported	220	267	210	204	107	176	187	135	122	150
Sign of Difference	0	+	+	−	+	−	+	−	+	+

From Table 13-3 we see that there are 6 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 x = 1, so we fail to reject the null hypothesis of no difference. We conclude that there is not sufficient evidence to reject the claim that for males, there is no difference between measured weights and reported weights.



Sign Test: Given Number of Signs

Median of Differences = 0

Significance: 0.05

Number Positive: 6

Number Negative: 3

```
Claim: μ = μ(hyp)
Number of Values:  9

Using Table A7
Test Statistic, x:  3.00000
Critical x:         1.00000
```

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.

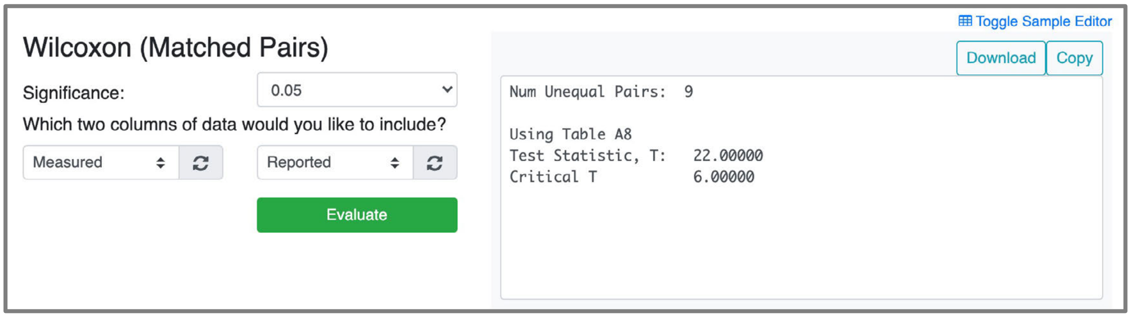
> **DEFINITION** The **Wilcoxon signed-ranks test** is a nonparametric test that uses ranks for these applications:
>
> 1. Testing a claim that the population of matched pairs has the property that the matched pairs have differences with median equal to zero.
> 2. Testing a claim that a single population of individual values has a median equal to some claimed value.

First we describe the Statdisk procedure for conducting a Wilcoxon signed-ranks test, and then we illustrate this procedure with an example.

Statdisk Procedure for the Wilcoxon Signed-Ranks Test

1. Enter the paired data in columns of the Statdisk Sample Editor.
2. Select **Analysis** in the top menu bar.
3. Select **Wilcoxon Tests** in the dropdown menu and select **Wilcoxon (Matched Pairs)** in the submenu.
4. Enter a significance level, such as 0.05 or 0.01
5. Select the columns of the Sample Editor that contain the paired data.
6. Click **Evaluate**.
7. Click the **Plot** button to generate 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 on the preceding page, the Statdisk results for the Wilcoxon signed-ranks test are shown below. Based on this display we see that the test statistic is $T = 22$, and the critical value is $T = 6$. Because the test statistic of $T = 22$ is *not* less than or equal to the critical value of $T = 6$, we fail to reject the null hypothesis. Based on the small sample, we conclude that there is not sufficient evidence to support the claim that for males, there is a significant difference between measured weights and reported weights.



Significance: 0.05

Which two columns of data would you like to include?

Measured   Reported

```
Num Unequal Pairs:  9

Using Table A8
Test Statistic, T:    22.00000
Critical T            6.00000
```

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, and then we illustrate it with an example.

Statdisk Procedure for the Wilcoxon Rank-Sum Test

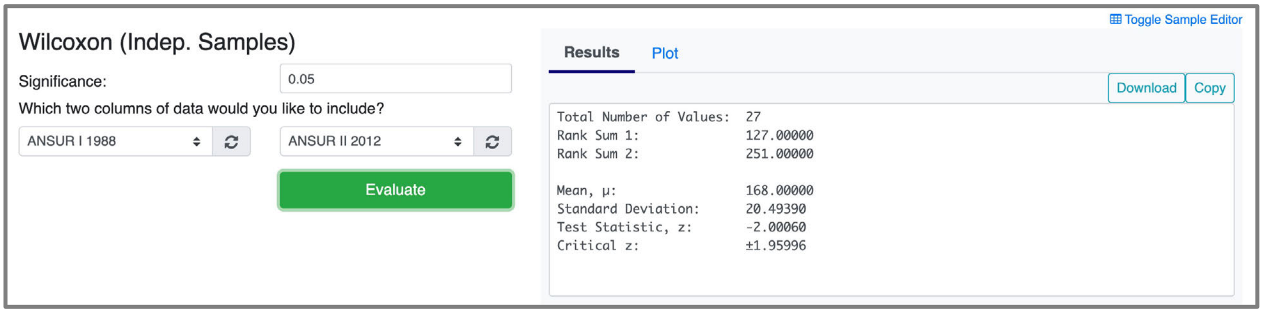
1. Enter the two lists of values from the two independent samples in columns of the Statdisk Sample Editor.
2. Select **Analysis** in the top menu bar.
3. Select **Wilcoxon Tests** in the dropdown menu.
4. Select **Wilcoxon (Independent Samples)** in the submenu.
5. Enter a significance level, such as 0.05 or 0.01.
6. Select the columns of the Sample Editor that contain the two sets of independent sample data.
7. Click the **Evaluate** button.
8. Click **Plot** to display a graph that shows the test statistic and critical values.

EXAMPLE Table 13-5 on the next page lists heights (mm) of males from *Elementary Statistics* 14th edition, Data Set 2 "ANSUR I 1988" and heights of males (mm) from Data Set 3 "ANSUR II 2012". Use a 0.05 significance level to test the claim that the two samples are from populations with the same median height.

TABLE 13-5 Heights (mm) of Males from ANSUR I and ANSUR

ANSUR I 1988	ANSUR II 2012
1698 **(5)**	1810 **(21)**
1727 **(8)**	1850 **(25)**
1734 **(11)**	1777 **(16)**
1684 **(3)**	1811 **(22)**
1667 **(1)**	1780 **(17.5)**
1680 **(2)**	1733 **(10)**
1785 **(19)**	1814 **(23)**
1885 **(27)**	1861 **(26)**
1841 **(24)**	1709 **(7)**
1702 **(6)**	1740 **(13)**
1738 **(12)**	1694 **(4)**
1732 **(9)**	1766 **(15)**
	1748 **(14)**
	1794 **(20)**
	1780 **(17.5)**
$n_1 = 12$	$n_2 = 15$
$R_1 = 127$	$R_2 = 251$

Shown below are the Statdisk results using the data from Table 13-5. Using the critical values of z = ±1.95996, we see that the test statistic of *z* = -2.00060 *does* fall within the critical region, so we reject the null hypothesis that the two samples are from populations with the same median. There is sufficient evidence to warrant rejection of the claim that the sample of male heights from ANSUR I 1988 and the sample of male heights from ANSUR II 2012 are from populations with the same median.



Wilcoxon (Indep. Samples)

Significance: 0.05

Which two columns of data would you like to include?

ANSUR I 1988 — ANSUR II 2012

```
Total Number of Values:  27
Rank Sum 1:              127.00000
Rank Sum 2:              251.00000

Mean, μ:                 168.00000
Standard Deviation:      20.49390
Test Statistic, z:       -2.00060
Critical z:              ±1.95996
```

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, and then we illustrate this procedure with an example.

Statdisk Procedure for the Kruskal-Wallis Test

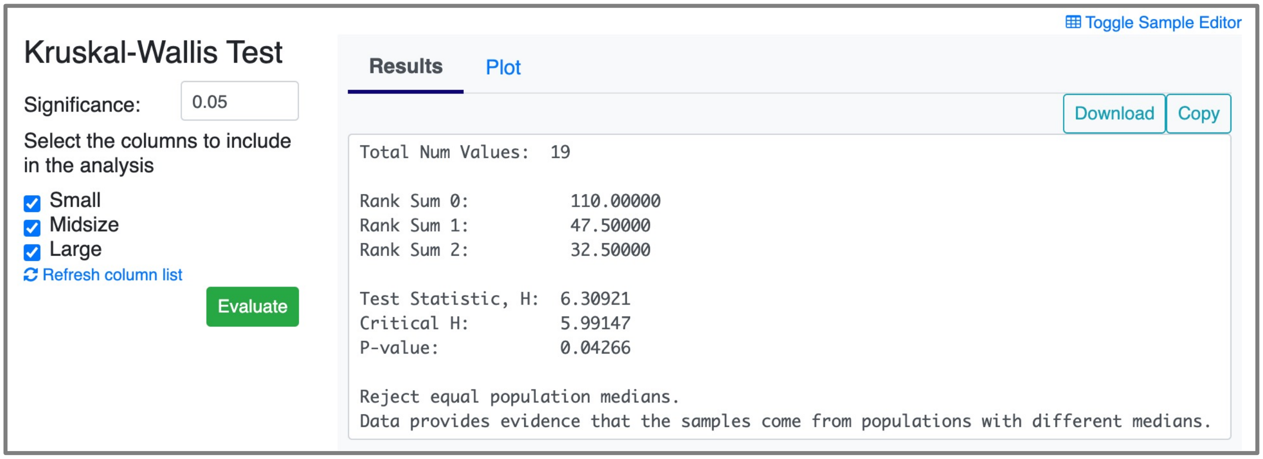
1. Enter the samples of data in columns of the Statdisk Sample Editor.
2. Select **Analysis** in the top menu.
3. Select **Kruskal-Wallis test** in the dropdown menu.
4. In the dialog box, enter a significance level, such as 0.05 or 0.01.
5. Select the columns containing the sample data. Click the boxes to insert or delete checkmarks (✓). Columns with checkmarks are included in the calculations.
6. Click the **Evaluate** button.
7. Click **Plot** to display a graph that shows the test statistic and critical values.

EXAMPLE Table 13-6 on the next page lists head injury criterion (HIC) measurements of small, midsize, and large car crash tests. Use a 0.05 significance level to test the claim that the three samples of HIC measurements are from populations with medians that are all equal.

TABLE 13-6 Head Injury Criterion (HIC) Measurements in Car Crash Tests (Ranks in parentheses)

Small	Midsize	Large
253 **(14)**	117 **(3)**	249 **(13)**
143 **(6)**	121 **(4)**	90 **(1)**
124 **(5)**	204 **(12)**	178 **(7.5)**
301 **(17)**	195 **(11)**	114 **(2)**
422 **(19)**	186 **(10)**	183 **(9)**
324 **(18)**	178 **(7.5)**	
258 **(15)**		
271 **(16)**		
$n_1 = 8$	$n_2 = 6$	$n_3 = 5$
$R_1 = 110$	$R_2 = 47.5$	$R_3 = 32.5$

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 110.0, 47.5, and 32.5, the test statistic of H = 6.30921, the critical value of H = 5.99147, and the P-value of 0.04266. Remember that the Kruskal-Wallis test is a *right-tailed* test. Because the P-value of 0.04266 is less than the significance level of 0.05, we reject the null hypothesis of equal population medians. There is sufficient evidence to reject the claim that the three samples of HIC measurements come from populations with medians that are all equal. At least one of the population medians appears to be different from the others.



Kruskal-Wallis Test

Significance: 0.05

Select the columns to include in the analysis

☑ Small
☑ Midsize
☑ Large

```
Total Num Values:  19

Rank Sum 0:          110.00000
Rank Sum 1:          47.50000
Rank Sum 2:          32.50000

Test Statistic, H:  6.30921
Critical H:         5.99147
P-value:            0.04266

Reject equal population medians.
Data provides evidence that the samples come from populations with different medians.
```

13-6 Rank Correlation

The Triola textbook (excluding *Essentials of Statistics*) 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, and then we illustrate this procedure with an example.

Statdisk Procedure for Rank Correlation

1. Enter the paired sample data in columns of the Statdisk Sample Editor.
2. Select **Analysis** in the top menu bar.
3. Select **Rank Correlation** in the dropdown menu
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 the **Evaluate** button.
7. Click **Plot** 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 $n > 30$).

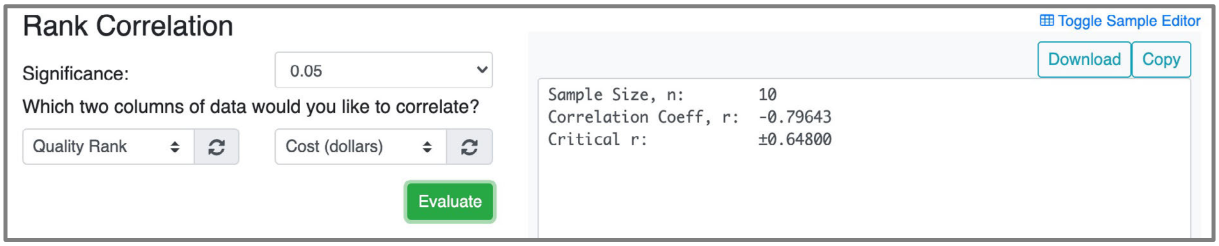
Consider the following example.

> **EXAMPLE** Table 13-1 lists ranks and costs (dollars) of smartphones (based on data from *Consumer Reports*) . Find the value of the rank correlation coefficient and use it to determine whether there is sufficient evidence to support the claim of a correlation between quality and price. Use a 0.05 significance level.

TABLE 13-1 Ranks and Costs of Smartphones

Quality Rank	1	2	3	4	5	6	7	8	9	10
Cost (dollars)	1000	1100	900	1000	750	1000	900	700	750	600

Statdisk provides the results shown below. Because the test statistic (r = -0.79643) is outside of the range between the critical r values of ±0.648, we reject the null hypothesis. There is sufficient evidence to support a claim of a correlation between quality and cost. It appears that you do get better quality by paying more, but this conclusion incorrectly implies causation.



Rank Correlation

Significance: 0.05

Which two columns of data would you like to correlate?

Quality Rank

Cost (dollars)

```
Sample Size, n:         10
Correlation Coeff, r:   -0.79643
Critical r:             ±0.64800
```

13-7 Runs Test for Randomness

The Triola textbook (excluding *Essentials of Statistics* and *Biostatistics for the Biological and Health Sciences*) discusses the runs test for randomness and includes these definitions.

DEFINITIONS 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, and then we illustrate this procedure 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** in the top menu bar.
3. Select **Runs Test** in the dropdown menu.
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 the **Evaluate** button.
6. Click **Plot** to display a graph with the test statistic and critical values.

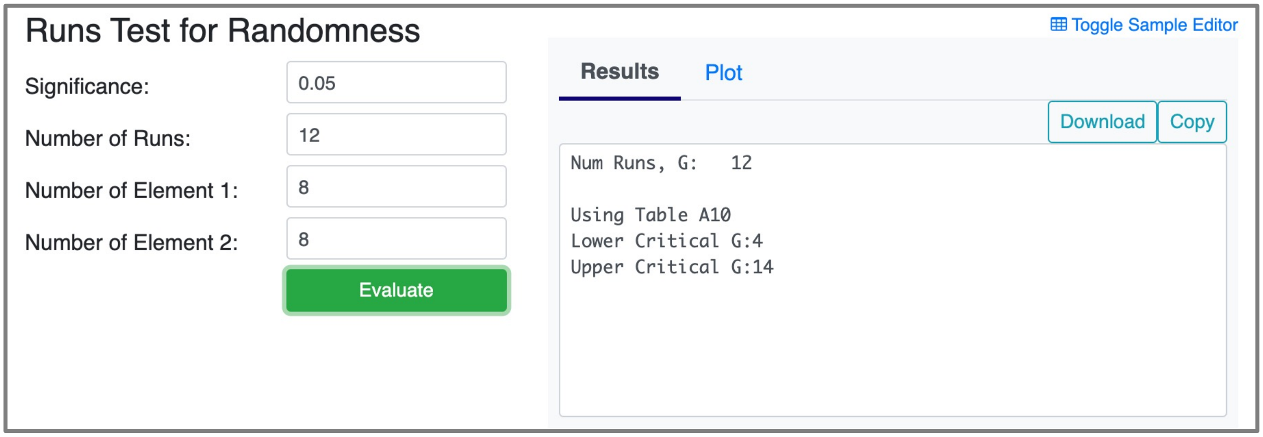
EXAMPLE Listed below are the recent political parties of presidents of the United States. The letter R represents a Republican president and the letter D represents a Democratic president. Does it appear that we elect Democrat and Republican presidents in a random sequence?

R D D R D D R R D R R D R D R D

The textbook describes the procedure for examining the above sequence to find these results:

G = number of runs = 12
n_1 = number of Republican presidents = 8
n_2 = number of Democratic president = 8

Using Statdisk for the runs test for randomness, we obtain the display shown below. Because G = 12 is neither less than or equal to the critical value of 4, nor is it greater than or equal to the critical value of 14, we do not reject randomness. There is not sufficient evidence to reject randomness in the sequence of political parties of recent presidents.



Runs Test for Randomness

Significance: 0.05

Number of Runs: 12

Number of Element 1: 8

Number of Element 2: 8

```
Num Runs, G:   12

Using Table A10
Lower Critical G:4
Upper Critical G:14
```

Runs Test with Large Samples

Elementary Statistics 14th Edition, Data Set 6 “Births” involves a large sample of genders resulting in 213 runs with n_1 = 205 and n_2 = 195. Instead of performing complicated calculations for μ_G and σ_G and the test statistic, Statdisk automatically displays results of μ_G = 200.875, σ_G = 9.981203, the test statistic z = 1.2148, and the critical values of z = ±1.95996.

CHAPTER 13 WORKBOOK: Nonparametric Statistics

13-1 **Flight Data** Use the sign test with the following times for American Airlines Flight 19 from New York (JFK) to Los Angeles (LAX). The times in each column are from the same day. 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:____

____

13-2 **Sign Test vs. Wilcoxon Signed–Ranks Test** Repeat 13-1 "Flight Data" using the Wilcoxon signed-ranks test for matched pairs. Enter the Statdisk results below, and compare them to the sign test results obtained in exercise 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: ____

____

13-3 **Do All Colors of M&Ms Weigh the Same?** Refer to *Elementary Statistics* 14th Edition, Data Set 38 "Candies." 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 (red, orange, yellow, brown, blue, green).

Test statistic:__________ Critical value: __________

Conclusion: ____

____

13-4 **IQ and Brain Volume** Refer to *Elementary Statistics* 14th Edition, Data Set 12 "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: ____

____

13-5 **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: ____

14 Statistical Process Control

Statdisk

14-1 Run Charts

We define **process data** to be data arranged according to some time sequence, such as the data in Table 14-1 below. Table 14-1 lists the global mean surface temperature (in °C) of Earth for each year from 1880, with projections used for the last two years. This data set is based on measurements provided by NASA Goddard's Global Surface Temperature Analysis (GISTEMP).

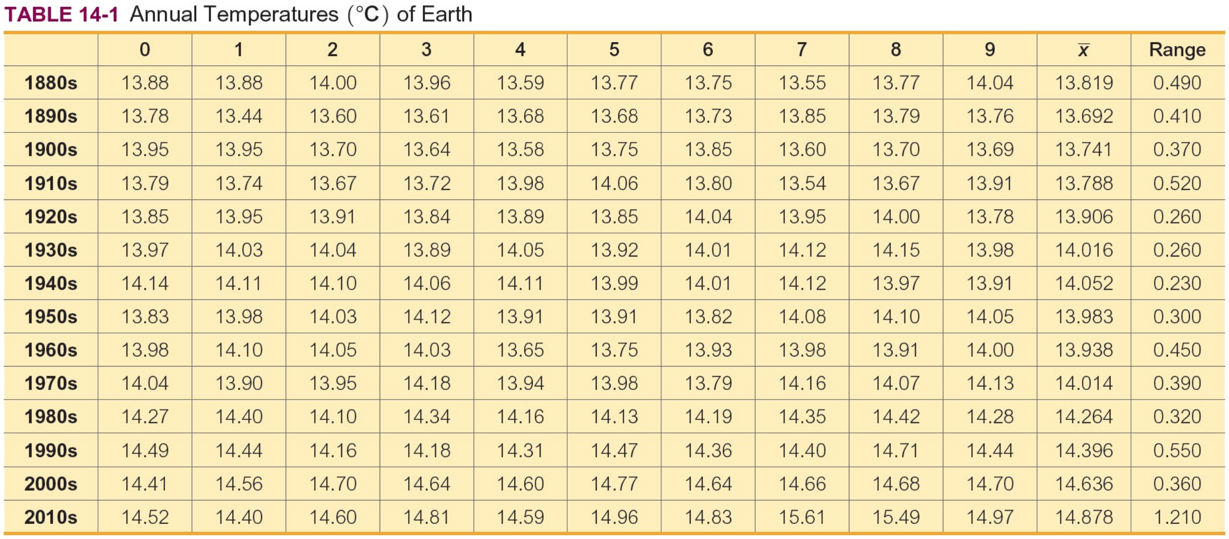


TABLE 14-1 Annual Temperatures (°C) of Earth

	0	1	2	3	4	5	6	7	8	9	$\bar{x}$	Range
1880s	13.88	13.88	14.00	13.96	13.59	13.77	13.75	13.55	13.77	14.04	13.819	0.490
1890s	13.78	13.44	13.60	13.61	13.68	13.68	13.73	13.85	13.79	13.76	13.692	0.410
1900s	13.95	13.95	13.70	13.64	13.58	13.75	13.85	13.60	13.70	13.69	13.741	0.370
1910s	13.79	13.74	13.67	13.72	13.98	14.06	13.80	13.54	13.67	13.91	13.788	0.520
1920s	13.85	13.95	13.91	13.84	13.89	13.85	14.04	13.95	14.00	13.78	13.906	0.260
1930s	13.97	14.03	14.04	13.89	14.05	13.92	14.01	14.12	14.15	13.98	14.016	0.260
1940s	14.14	14.11	14.10	14.06	14.11	13.99	14.01	14.12	13.97	13.91	14.052	0.230
1950s	13.83	13.98	14.03	14.12	13.91	13.91	13.82	14.08	14.10	14.05	13.983	0.300
1960s	13.98	14.10	14.05	14.03	13.65	13.75	13.93	13.98	13.91	14.00	13.938	0.450
1970s	14.04	13.90	13.95	14.18	13.94	13.98	13.79	14.16	14.07	14.13	14.014	0.390
1980s	14.27	14.40	14.10	14.34	14.16	14.13	14.19	14.35	14.42	14.28	14.264	0.320
1990s	14.49	14.44	14.16	14.18	14.31	14.47	14.36	14.40	14.71	14.44	14.396	0.550
2000s	14.41	14.56	14.70	14.64	14.60	14.77	14.64	14.66	14.68	14.70	14.636	0.360
2010s	14.52	14.40	14.60	14.81	14.59	14.96	14.83	15.61	15.49	14.97	14.878	1.210

A *run chart* is a sequential plot of *individual* data values over time. Statdisk has a function for creating run charts as described below.

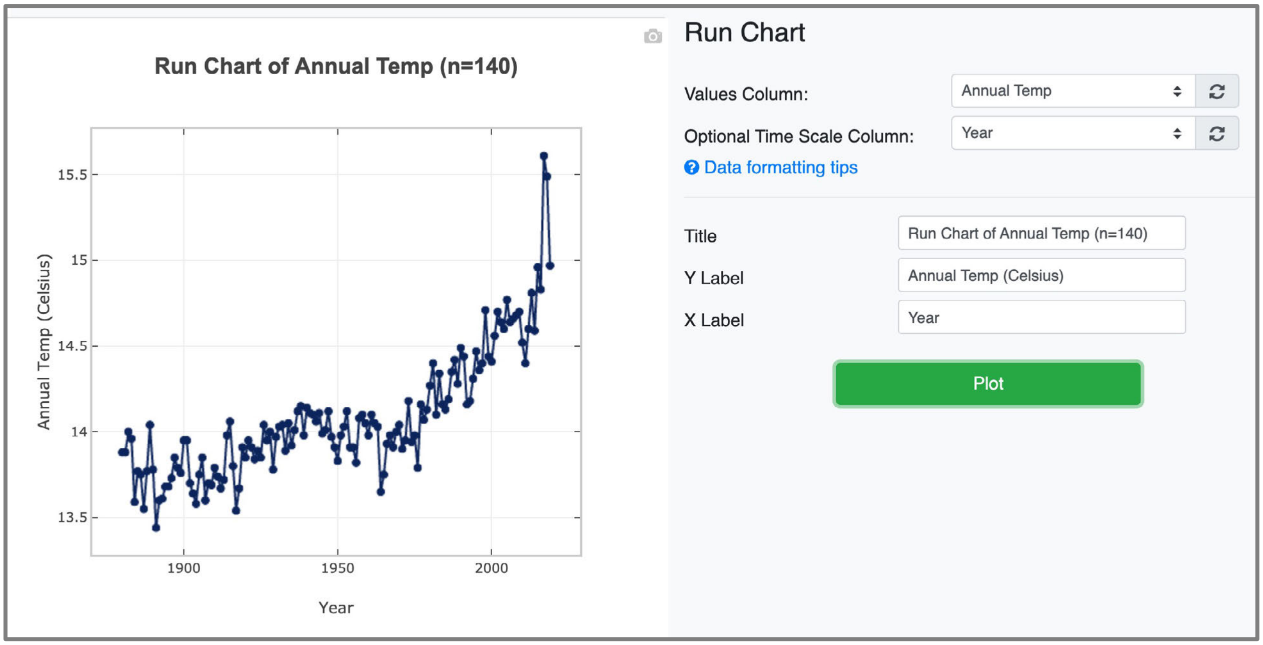
Statdisk Procedure for Creating a Run Chart

To generate a run chart in Statdisk, we use the following procedure.

1. Enter the individual data values in one column of the Statdisk Sample Editor. If time/sequential data is available, enter this data in a second column. (If there is no time/sequential data, the individual data values will be graphed in sequence.)
2. Click **Data** in the top menu bar.
3. Select **Run Chart** in the dropdown menu. The *Run Chart* dialog box appears in Statdisk.
4. For *Values Column*, select the column containing the individual data values. *OPTIONAL:* For *Optional Sequence/Time Column*, select the column containing the sequence/time data.
5. Modify the run chart *Title*, *Y label* and *X label* as desired.
6. Click the **Plot** button.

Shown on the next page is the run chart of the 140 sample values from Table 14-1 paired with the year for each data value

.



Run Chart

Values Column: Annual Temp

Optional Time Scale Column: Year

Title: Run Chart of Annual Temp (n=140)

Y Label: Annual Temp (Celsius)

X Label: Year

Examine the run chart above and note that it reveals this problem: as time progresses from left to right, the points appear to be rising. If this pattern continues, rising temperatures will cause melting of large ice formations, widespread flooding, and many other climatic changes.

14-2 Control Charts

Statdisk is not programmed to generate control charts, and it is recommended that you use a different software package, such as Minitab or StatCrunch, to create *R* Charts and $\overline{x}$ Charts.

CHAPTER 14 WORKBOOK: Statistical Process Control

14-1 **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?

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

Conclusion____

____

____

14-2 **Weights of Minted Quarters** Create a run chart using *Elementary Statistics*, 14th edition, Data Set 44 "Weights of Minted Quarters". Treat the 100 consecutive weight measurements from the 20 days as individual values. What does the result suggest?

Conclusion____

____

____

14-3 **Cola Cans** In each 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

Conclusion____

____

____