

# 2

## Describing, Exploring, and Comparing Data

---

- 2-1**      **Descriptive Statistics**
- 2-2**      **Histograms and Frequency Distributions**
- 2-3**      **Boxplots**
- 2-4**      **Scatter Diagrams**
- 2-5**      **Sorting Data**

*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/workbook. Be sure to understand these functions before beginning this chapter.

The main objective of Chapter 2 in the textbook is to introduce the tools needed to describe, explore, or compare the characteristics of a data set that are extremely important.

## Important Characteristics of Data

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

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

In this chapter, we learn how to use STATDISK as a tool for investigating the above important characteristics.

Chapter 2 in the textbook begins with a Chapter Problem related to the issue of secondhand smoke. That Chapter Problem includes the measured levels of cotinine for three samples of people, and those measurements are included in Table 2–1, which is reproduced on the following page. One sample consists of smokers, the second sample (ETS) consists of people who do not smoke, but are exposed to environmental tobacco smoke at home or work, and the third sample (NOETS) consists of nonsmokers who are not exposed to environmental tobacco smoke. The data are listed in Data Set 6 of Appendix B in the textbook, and the data are also included with the STATDISK program that is on the CD–ROM packaged with the textbook. (The STATDISK file names are SMKR.sdd, ETS.sdd, and NOETS.sdd.) The textbook noted that cotinine is a metabolite of nicotine, meaning that cotinine is produced in the body when nicotine is absorbed. Because it is known that nicotine is absorbed through cigarette smoke, we have a way of measuring the effective presence of cigarette smoke indirectly by measuring the amount of cotinine that is present.

Chapter 2 in the textbook presents techniques for describing, exploring, and comparing data such as those included in Table 2–1. In this manual/workbook we show how STATDISK can be used.

Table 2–1 Measured Cotinine Levels in Three Groups

**SMOKER**

1	0	131	173	265	210	44	277	32	3
35	112	477	289	227	103	222	149	313	491
130	234	164	198	17	253	87	121	266	290
123	167	250	245	48	86	284	1	208	173

**ETS**

384	0	69	19	1	0	178	2	13	1
4	0	543	17	1	0	51	0	197	3
0	3	1	45	13	3	1	1	1	0
0	551	2	1	1	1	0	74	1	241

**NOETS**

0	0	0	0	0	0	0	0	0	0
0	9	0	0	0	0	0	0	244	0
1	0	0	0	90	1	0	309	0	0
0	0	0	0	0	0	0	0	0	0

**2-1 Descriptive Statistics**

To obtain descriptive statistics (mean, median, standard deviation, and so on) for a set of sample values, follow these steps.

1. Enter or retrieve a set of sample data. (To *enter* values, use **Data/Sample Editor** as described in Section 1-2 in this manual/workbook; to *retrieve* a data set, use **File/Open** as described in Section 1-4 of this manual/workbook.) After entering or retrieving a data set, you will be in the **Sample Editor** window.
2. Click on the **Copy** bar at the bottom of the Sample Editor window. (See Section 1-5 of this manual/workbook for information about the Copy function.)
3. Click on **Data** in the main menu bar at the top.

4. Click on **Descriptive Statistics**.
5. Click on the **Paste** bar at the bottom. The data should appear in the Descriptive Statistics window.
6. Click on the **Evaluate** bar located on the bottom left side of the window.

As an example, either enter the 40 cotinine levels of the *smokers* listed in Table 2–1, or open the STATDISK file SMKR.sdd.

- *Manual entry:* If manually entering the 40 values, begin by first clicking on **Data**, then selecting **Sample Editor**. Proceed to enter the first value of 1, then press **Enter**. Now enter the second value of 0 and press **Enter**. Enter the third value of 131 and press **Enter**, and continue until all 40 values have been entered. Take great care to enter the values correctly.
- *Opening file:* To open the file SMKR.sdd, click on **File**, then **Open**. If STATDISK was properly installed, the STATDISK data sets should be listed. Scroll to the file SMKR.sdd and click on it, then click on the **Open** button. You should now see the Sample Editor screen with the first several values listed. All of the values are there, but only the first several values are visible in the list.

With the data listed in the Sample Editor window, continue with Step 2 in the above procedure, and complete the remaining steps so that the data set is copied into the Descriptive Statistics module. Click on Evaluate, and the window display should be as shown below.

**Sample Descriptive Statistics**

SMKR	
Sample Size, n	40
Mean, $\bar{x}$	172.47
Median	170.00
Midrange	245.50
RMS	208.97
Variance, $s^2$	14280
St Dev, s	119.50
Mean Dev	94.775
Range	491.00
Minimum	0.0000
1 <sup>st</sup> Quartile	86.500
2 <sup>nd</sup> Quartile	170.00
3 <sup>rd</sup> Quartile	251.50
Maximum	491.00
$\sum x$	6899.0
$\sum x^2$	1746819

SMKR	
1	1
2	0
3	131
4	173
5	265
6	210
7	44
8	277
9	32
10	3
11	35
12	112
13	477
14	289
15	227
16	103
17	222
18	149
19	313
20	491

Buttons: Evaluate, Help, Clear, Copy, Paste

The preceding STATDISK display shows the values of important descriptive statistics. The value of "RMS" is the value of the *root mean square* (or quadratic mean) described in Exercise 29 of Section 2-4 in the textbook. The value listed as "Mean Dev" is the mean deviation (or *mean absolute deviation*) defined in Section 2-5. Section 2-7 of the textbook includes the definition of a "5-number summary (minimum, 1st quartile, 2nd quartile, 3rd quartile, maximum), and that summary is located near the bottom of the STATDISK display.

Based on the items included in the above STATDISK display, we now have an understanding of some of these important characteristics of data:

Center:           The mean is 172.5 and the median is 170.0 (rounded).  
Variation:        The standard deviation is 119.5, the variance is 14,280, and the range is 491.0.

To understand other important characteristics of data, such as distribution, outliers, and pattern over time, we need other tools of STATDISK. In the next section of this manual/workbook, we consider histograms and frequency distributions, which give us important insight into the nature of the distribution of the data.

## 2-2 Histograms and Frequency Distributions

In designing STATDISK, we did not include a specific menu item for generating a frequency distribution from a list of raw data, but frequency distributions can be obtained by using the ability of STATDISK to generate histograms. Section 2-3 of the textbook describes the manual construction of a histogram, but STATDISK can be used to automatically generate this important graph. The basic approach is to get the data listed in the Sample Editor window, then use Copy/Paste to copy the data to the Histogram module, where the histogram graph is generated. When using STATDISK's Histogram program, you have the option of simply accepting default settings, or you can set your own limits on the classes. 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:

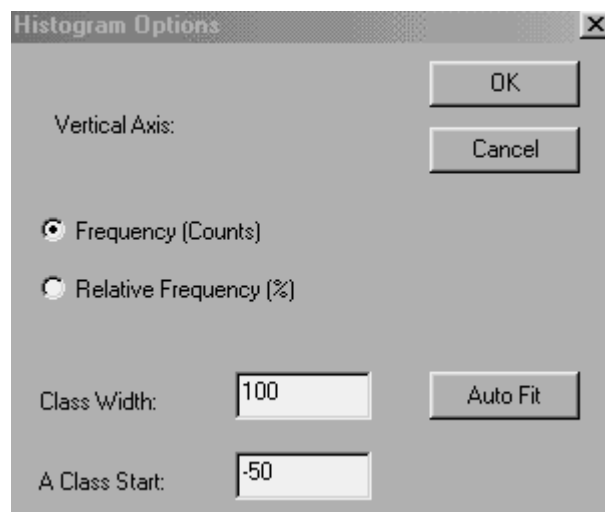
**Class width** is the difference between two consecutive lower class limits or two consecutive lower class boundaries.

See Table 2-2 in Section 2-2 of the textbook, where we note that the frequency distribution with classes of 0 – 99, 100 – 199, . . . , 400 – 499 has a class width of 100 (not 99).

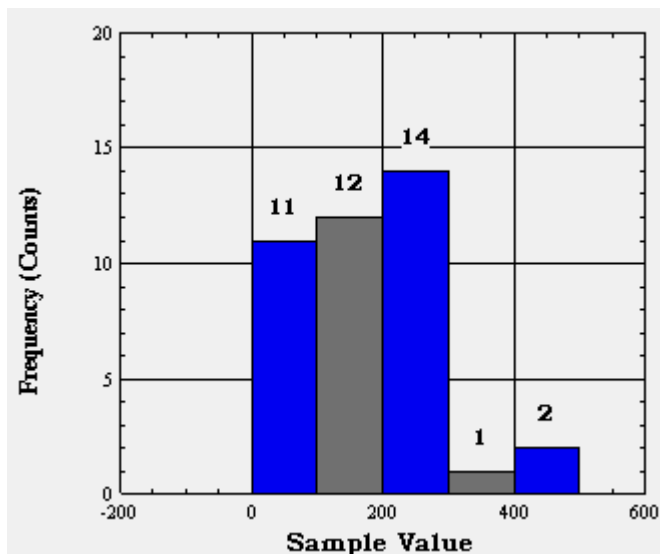
### Procedure for Generating a Histogram

1. Enter or retrieve a set of sample data so that the sample values are listed in the **Sample Editor** window. (To *enter* values, use **Data/Sample Editor** as described in Section 1-2 in this manual/workbook; to *retrieve* a data set, use **File/Open** as described in Section 1-4 of this manual/workbook.)
2. Click on the **Copy** bar at the bottom of the Sample Editor window. (See Section 1-5 of this manual/workbook for more information about the Copy function.)

3. Click on **Data** in the main menu bar at the top.
4. Click on **Histogram**.
5. Click on **Paste**. (The sample data should now be listed in the Histogram window.)
6. Click on **Evaluate** (or click on Help if you first want additional information about generating a histogram).
7. A window will appear with the title of Histogram Options. (See the display shown below.) The easy way to proceed is to simply click **OK**. (You have the option of changing the class width and starting point of the first class. You also have the option of having a vertical scale that uses relative frequencies instead of actual frequency counts. Click OK after making the desired changes.)



***Finding a Frequency Distribution:*** If we use the 40 cotinine levels of *smokers* in Table 2-1 and, instead of accepting the STATDISK default settings, we change the class start from  $-50$  to  $0$ , we get the histogram shown on the next page. Note that frequency counts are positioned above the bars of the histogram. Using a class width of  $100$  and a class start of  $0$  (instead of the default of  $-50$ ), and referring to the frequency counts shown in the STATDISK histogram, we can construct the frequency distribution included in the histogram, and that frequency distribution is the same as Table 2-2 included in Section 2-2 of the textbook. [*Technical note:* The upper class limits are shown as  $99$ ,  $199$ ,  $299$ ,  $399$ , and  $499$  (not  $100$ ,  $200$ ,  $300$ ,  $400$ ,  $500$ ), because 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. Given that the sample data are all whole numbers, the largest sample value that could fall in the first class is therefore  $99$ , not  $100$ .]



$X$	Frequency
0 – 99	11
100 – 199	12
200 – 299	14
300 – 399	1
400 – 499	2

When using STATDISK's Histogram program, it is easy to accept the program defaults, which is fine if your sole objective is to see a graph of the *distribution* of the data. If you want to use STATDISK to construct a frequency distribution, choose the option of entering your own starting point and class width (based on the range of values and the minimum value).

Among the important characteristics of data, the histogram gives us insight into the nature of the *distribution*. In later chapters, it often becomes important to determine whether sample data appear to come from a population with a normal distribution. A "normal distribution" will be discussed in Chapter 5, but for now we can consider it to be a distribution that is roughly bell-shaped when graphed in a histogram. We can therefore 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 appears to be very roughly bell-shaped, so that a requirement of a normal distribution appears to be approximately satisfied. (STATDISK can also generate *normal quantile plots*, which are helpful for determining whether distributions are normal, and those graphs are discussed in Chapter 5 of this manual/workbook.) Other more advanced tests also lead to the same conclusion that the cotinine levels of the 40 smokers appear to come from a population with a distribution that is approximately normal (or bell-shaped).

## 2-3 Boxplots

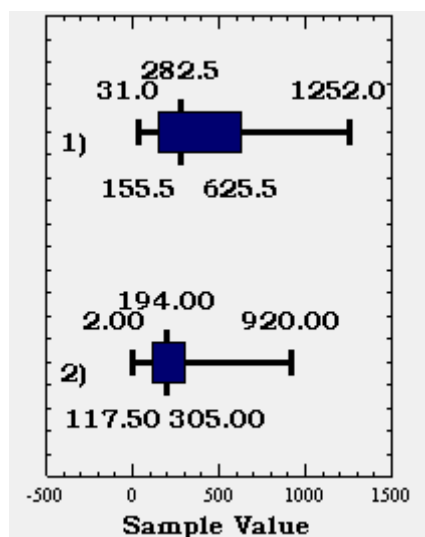
Section 2-7 in the textbook describes the construction of boxplots. They are based on the 5-number summary consisting of the minimum, first quartile, second quartile, third quartile, and maximum. The basic approach is to enter or retrieve a data set, then use Copy/Paste to copy the data to the Boxplot module, which is one of the modules under the main menu item of Data.

### Procedure for Generating a Boxplot

1. Enter or retrieve a set of sample data so that the sample values are listed in the **Sample Editor** window. (To *enter* values, use **Data/Sample Editor** as described in Section 1-2 in this manual/workbook; to *retrieve* a data set, use **File/Open** as described in Section 1-4 of this manual/workbook.)
2. Click on the **Copy** bar at the bottom of the Sample Editor window. (See Section 1-5 of this manual/workbook.)
3. Click on **Data** in the main menu bar at the top.
4. Click on **Boxplot**.
5. Click on **Paste**.
6. In the window for selecting a destination, enter column 1, then click **OK**.
7. Click on **Evaluate**.
8. In the window for boxplot options, you must select the columns to plot. (You have the option of plotting several boxplots in the same window.) If there is only one data set, select column 1, then click **OK**.

Section 2–7 in the textbook notes that one important advantage of boxplots is that they are very useful in comparing data sets. Shown below is the STATDISK display showing the two boxplots representing the cholesterol levels of men and women listed in Data Set 1 from Appendix B in the textbook. (The STATDISK file names are MCHOL.sdd and FCHOL.sdd.) These same data sets are illustrated with boxplots in Section 2–7 of the textbook. To get more than one boxplot displayed in the same window, follow steps 1-5 above to enter the data for the first set of values. Then enter or retrieve another data set and use Copy/Paste to copy it to the *same* boxplot window, but enter the second set of data in column 2. Continue until you are finished entering all of the data sets (up to 9), then continue with steps 7 and 8 above.

In the display shown below, the top boxplot depicts the cholesterol levels of 40 men, and the bottom boxplot represents the cholesterol levels of 40 women. Because the two boxplots are constructed on the same scale, a comparison becomes easier. The boxplots suggest that males have cholesterol levels that are generally higher than females, and the cholesterol levels of males appear to vary more than the cholesterol levels of females.



*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 Section 2-6 of the textbook, but other programs may use different procedures, so there may be some differences in boxplot results. The textbook states that there is not universal agreement on a single procedure for calculating quartiles, and different computer programs might yield different results. For example, if you use the data set of 1, 3, 6, 10, 15, 21, 28, and 36, you will get the results shown below. For this particular data set, STATDISK and the TI-83 Plus calculator agree, but they do not always agree.

	$Q_1$	$Q_2$	$Q_3$
STATDISK	4.5	12.5	24.5
Minitab	3.75	12.5	26.25
Excel	5.25	12.5	22.75
TI-83 Plus	4.5	12.5	24.5

## 2-4 Scatter Diagrams

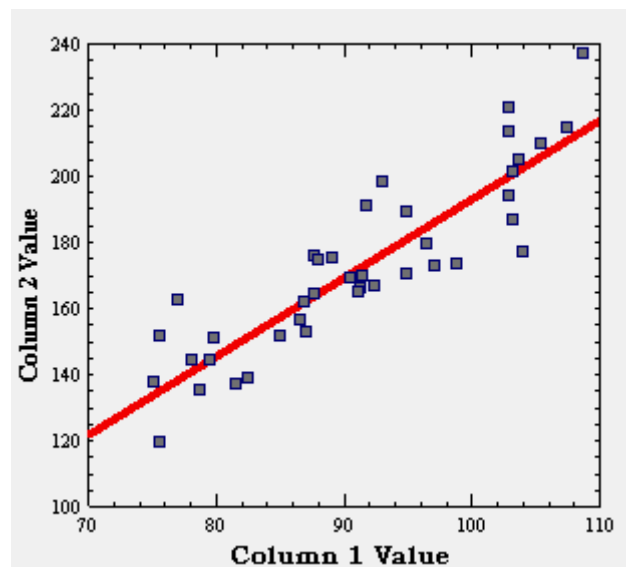
Section 2–3 of the textbook includes a scatter diagram (or scatterplot) generated by the statistical software package Minitab. A scatter diagram can be very helpful in seeing a relationship between two variables. The Minitab display included in the textbook shows that larger waist sizes of males appear to correspond to larger weights. (The sample data consist of the paired waist/weight measurements for males listed in Data Set 1 from Appendix B of the textbook.)

### Procedure for Generating a Scatter Diagram

To use STATDISK for generating a scatter diagram, you must have a collection of *paired* data.

1. Select **Data** from the main menu.
2. Select the subdirectory item of **Scatterplot**.
3. Enter or copy the paired data into columns 1 and 2. (You can, for example, retrieve a data set, then use Copy/Paste to move it into one of the columns in the Scatterplot module. Next, you can retrieve the second data set and use Copy/Paste to move it into the other column in the Scatterplot module.)
4. After the two matched columns of data have been entered, click on **Evaluate**.

The scatter diagram will include the straight line that fits the points best. This line is discussed in Chapter 9, but it can be ignored at the present time. If we use the paired waist/weight data for the sample of 40 males from Data Set 1 in Appendix B of the textbook, we get the scatter diagram shown below. Based on the pattern of the points, we can conclude that there does appear to be a relationship between waist size and weight. Males with larger waists tend to weigh more. Such relationships (or *correlations*) will be discussed at much greater length in Chapter 9.



## 2-5 Sorting Data

To *sort* data is to arrange them in order. There are several cases in which it becomes necessary to rearrange a data set so that the values are in order (ascending from low to high, or descending from high to low). First enter or copy the data into the Sample Editor module (accessed from the main menu item of Data), then click on the Format/Sort bar and proceed to select the way that you want the data arranged. Here are the details of this procedure.

### Procedure for Sorting Data

1. Enter or retrieve a set of sample data so that the sample values are listed in the **Sample Editor** window. (To *enter* values, use **Data/Sample Editor** as described in Section 1-2 in this manual/workbook; to *retrieve* a data set, use **File/Open** as described in Section 1-4 of this manual/workbook.)
2. With the data set listed in the Sample Editor window, click on the **Format/Sort** bar.
3. A window will appear with the title of Sample Format Options.
  - Select a sort order of *ascending* if you want the data arranged in order from low to high.
  - Select *descending* if you want the data arranged in decreasing order.
4. Click **OK** and the data will be arranged in the order you specified. The sorted data set can now be saved, or copied to other modules.

**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 exceptional value when it is buried in the middle of a long list 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.*

As an example, consider the axial loads of aluminum cans 0.0111 in. thick, found in Data Set 20 in Appendix B of the textbook. We can retrieve that data set by clicking on **File**, then **Open**, then selecting CN111.sdd. With the 175 sample values displayed in the Sample Editor window, we can first sort the data, then examine the first few values and scroll through to the last few values. See the following two displays showing the first few values and the last values from the data set CN111.sdd. It becomes clear that only the maximum value of 504 is very far away from the other sample values, so 504 appears to be an outlier. If we plan to further analyze the sample data, we should be aware of any outliers, because they might dramatically affect some of our results.

*Lowest Sample Values*

CN111	
1	205
2	210
3	210
4	211
5	215
6	216
7	222
8	225
9	227
10	230
11	231
12	243
13	244
14	246
15	247
16	247
17	247
18	250

*Highest Sample Values*

CN111	
159	304
160	304
161	305
162	305
163	306
164	306
165	306
166	307
167	308
168	309
169	310
170	311
171	313
172	314
173	315
174	317
175	504
176	



**504 is an outlier.**

We noted at the beginning of this chapter that the following are extremely important characteristics of data: center, variation, distribution, outliers, and pattern over time. Chapter 13 will consider the characteristic of pattern over time, but the other characteristics can be investigated using tools of STATDISK, as we have described in this chapter. Here is a summary of the tools that are usually most relevant for the different characteristics:

1. **Center:** Use **Data/Descriptive Statistics** to find the mean and median.
2. **Variation:** Use **Data/Descriptive Statistics** to find the standard deviation, variance, and range.
3. **Distribution:** Use **Data/Histogram** and **Data/Boxplot** to generate a histogram and boxplot.
4. **Outliers:** Use **Data/Sample Editor/Format** to sort the data in ascending order, then examine the sample values to identify any that are very far away from almost all other values.
5. **Time:** See Chapter 13.

## CHAPTER 2 EXPERIMENTS: Describing, Exploring, and Comparing Data

- 2–1. *Comparing Heights of Men and Women* In this experiment we use two small data sets as a quick introduction to using some of the basic STATDISK features. (When beginning work with new software, it is wise to first work with small data sets so that they can be entered quickly if they are lost or damaged.) The data listed below are measured heights (in inches) of random samples of men and women (taken from Data Set 1 in Appendix B of the textbook).

<b>Men</b>	70.8	66.2	71.7	68.7	67.6	69.2
<b>Women</b>	64.3	66.4	62.3	62.3	59.6	63.6

- 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.

---



---



---

2–2. **Working with Larger Data Sets** Repeat Experiment 2–1, but use the sample data for all 40 males and 40 females included in Data Set 1 that is found in Appendix B of the textbook. Instead of manually entering the 80 individual heights (which would be no fun at all), open the STATDISK files MHT.sdd and FHT.sdd (for male heights and female heights, respectively).

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.

---

---

---

d. Are there are notable differences observed from the complete sets of sample data that could not be seen with the smaller samples listed in Experiment 2–1? If so, what are they?

---

---

---

- 2–3. **Histogram** Use the same sets of data from Experiment 2–2 and print histograms for the heights of the 40 men and the heights of the 40 women. Are there any notable differences in the two sets of sample data?

---

---

---

- 2–4. **Boxplots** Use the same sets of data from Experiment 2–2 and print boxplots for the heights of the 40 men and the heights of the 40 women. The two boxplots are much easier to compare if they are constructed together on the same scale. Use STATDISK to generate the two boxplots together, then print the results. Do the boxplots suggest any notable differences in the two sets of sample data?

---

---

---

- 2–5. **Scatterplots** Section 2–4 of this manual/workbook included an example of a scatterplot depicting the paired waist and weight measurements for a sample of 40 males. Use STATDISK to print the scatterplot for the paired waist and weight measurements for the sample of 40 women. The sample values can be retrieved from the STATDISK files FFAST.sdd and FWT.sdd. How does the resulting scatterplot compare to the scatterplot shown for males? Is a similar scatterplot obtained? Does there appear to be a relationship between waist sizes and weights of females?

---

---

---

- 2–6. **Scatterplots** Now it's time to be creative. When requesting a printout of a scatterplot, Experiment 2–5 specified the variables of waist size and weight for a sample of females. Using any data set from Appendix B in the textbook (other than Data Set 1), identify two paired variables that you suspect are related, then obtain a STATDISK printout of the scatterplot. Does the graph support your belief that there is a relationship? What feature of the graph suggests that there is or is not a relationship?

---

---

---

- 2-7. **Head Circumferences** (This is from Exercise 13 in Section 2-4 of the textbook.) In order to correctly diagnose the disorder of hydrocephalus, a pediatrician investigates head circumferences of two-year-old males and females. Use the sample results listed in Data Set 3 from Appendix B in the textbook. Does there appear to be a difference between the two genders?
- 
- 

- 2-8. **Weekend Rainfall** (This is from Exercise 15 in Section 2-4 of the textbook.) Using Data Set 11 in Appendix B from the textbook, find the mean and median of the rainfall amounts in Boston on Thursday and find the mean and median of the rainfall amounts in Boston on Sunday. Media reports claimed that it rains more on weekends than during the week. Do these results appear to support that claim?
- 
- 

- 2-9. **Tobacco/Alcohol Use in Children's Movies** (This is from Exercise 16 in Section 2-4 of the textbook.) In "Tobacco and Alcohol Use in G-Rated Children's Animated Films," by Goldstein, Sobel and Newman (*Journal of the American Medical Association*, Vol. 281, No. 12), the lengths (in seconds) of scenes showing tobacco use and alcohol use were recorded for animated children's movies. Refer to Data Set 7 in Appendix B from the textbook and find the mean and median for the tobacco times, then find the mean and median for the alcohol times. Does there appear to be a difference between those times? Which appears to be the larger problem: scenes showing tobacco use or scenes showing alcohol use?
- 
- 

- 2-10. **Effect of Outlier** In this experiment we study the effect of an *outlier*. Use the same heights of *men* used in Experiment 2-1, but change the first entry from 70.8 in. to 708 in. (This type of mistake often occurs when the key for the decimal point is not pressed with enough force.) The outlier of 708 in. is clearly a mistake, because a male with of height of 708 in. would be 59 feet tall, or about six stories tall. Although this outlier is a mistake, outliers are sometimes correct values that differ substantially from the other sample values.

Men: **708** 66.2 71.7 68.7 67.6 69.2

Using this modified data set with the height of 70.8 in. changed to be the outlier of 708 in., find the following.

*Center:* Mean: \_\_\_\_\_ Median: \_\_\_\_\_

*Variation:* St. Dev.: \_\_\_\_\_ Range: \_\_\_\_\_

*5-Number Summary:* Min.: \_\_\_\_\_  $Q_1$ : \_\_\_\_\_  $Q_2$ : \_\_\_\_\_  $Q_3$ : \_\_\_\_\_ Max.: \_\_\_\_\_

*Outliers:* \_\_\_\_\_

Based on a comparison of these results to those found in Experiment 2–1, how is the mean affected by the presence of an outlier?

\_\_\_\_\_

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

\_\_\_\_\_

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

\_\_\_\_\_

- 2–11. **Effect of Outlier** In Experiment 2–3 we obtained a printout of a histogram for the heights of 40 males. Change the first height from 70.8 in. to the outlier of 708 in., then obtain the histogram. How is the histogram affected by the presence of the outlier? Does the outlier disguise the true nature of the distribution of the data?

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

- 2–12. **Sorting Data** Retrieve the STATDISK file SOLTR.sdd. The sample data represent losing amounts (represented by negative values) and winning amounts (represented by positive values) when solitaire is played with Las Vegas rules. *Sort* the data by arranging them in order from lowest to highest.

How many of the sample values are negative? \_\_\_\_\_

How many of the sample values are positive? \_\_\_\_\_

What is the largest amount that was lost? \_\_\_\_\_

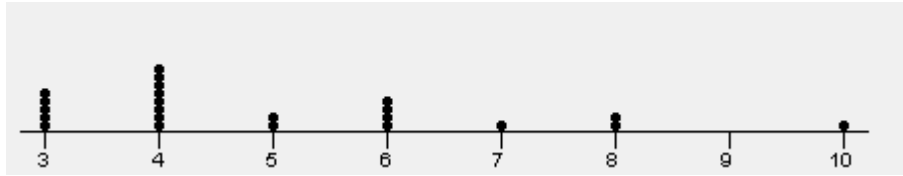
What is the largest amount that was won? \_\_\_\_\_

What is the mean amount won or lost per game? \_\_\_\_\_

What do the results suggest about this game of solitaire?

\_\_\_\_\_

- 2-13. **Interpreting Dotplot** Shown below is a dotplot of sample data. Identify the values represented in this graph, enter them in STATDISK, then find the indicated results.



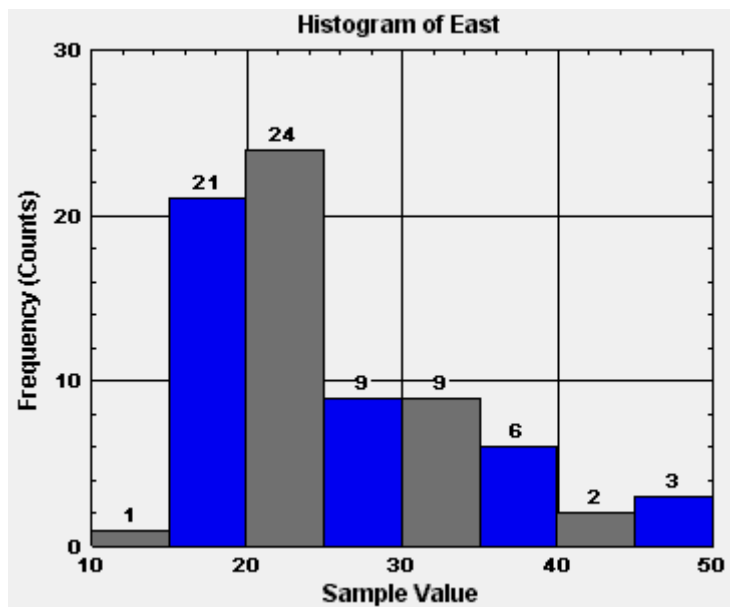
Center: Mean: \_\_\_\_\_ Median: \_\_\_\_\_

Variation: St. Dev.: \_\_\_\_\_ Range: \_\_\_\_\_

5-Number Summary: Min.: \_\_\_\_\_  $Q_1$ : \_\_\_\_\_  $Q_2$ : \_\_\_\_\_  $Q_3$ : \_\_\_\_\_ Max.: \_\_\_\_\_

Outliers: \_\_\_\_\_

- 2-14. **Frequency Distribution** Shown below is a STATDISK-generated histogram representing the ages (n years) of eastbound stowaways on the Queen Mary. The data are listed in Data Set 15 from Appendix B of the textbook. Use the displayed histogram to construct a table representing the frequency distribution, and identify the value of the class width. Enter the results in the space to the left of the histogram.



2–15. **Comparing Data** Readability data were compiled from randomly selected pages of the following books:

- Tom Clancy's *The Bear and the Dragon*
- J. K. Rowling's *Harry Potter and the Sorcerer's Stone*
- Leo Tolstoy's *War and Peace*

The data are listed in Data Set 14 of Appendix B in the textbook, and they are also available as STATDISK files. Refer to Data Set 14 in Appendix B for the STATDISK file names. Use STATDISK to compare the readability of the three books. Provide relevant printed displays, and write a brief report stating your conclusions.

2–16. **Comparing Data** The ages of actors and actresses when they won Oscars are listed below. (The same data are used for Exercise 34 in Section 2–3 of the textbook.) Use STATDISK to compare the two data sets. Provide relevant printed displays, and write a brief report stating your conclusions.

Actors

32	37	36	32	51	53	33	61	35	45
55	39	76	37	42	40	32	60	38	56
48	48	40	43	62	43	42	44	41	56
39	46	31	47	45	60	46	40	36	

Actresses

50	44	35	80	26	28	41	21	61	38
49	33	74	30	33	41	31	35	41	42
37	26	34	34	35	26	61	60	34	24
30	37	31	27	39	34	26	25	33	

2-17. **Exploring Distributions** Retrieve each of the indicated data sets and obtain a printed copy of a STATDISK–generated histogram. Examine the histogram and described its general shape. Determine whether the shape of the distribution is approximately bell-shaped.

Rainfall amounts for Boston on Tuesday (STATDISK file RNTUE.sdd):

---

Weights of Domino sugar packets (STATDISK file SUGAR.sdd):

---

Weights of quarters in circulation (STATDISK file QRTRS.sdd):

---

2-18. **Working with Your Own Data** Through observation or experimentation, collect your own set of sample data. Obtain at least 40 values and try to select data from an interesting population. Use STATDISK for help in answering the following questions.

a. Describe the nature of the data. That is, what do the values represent?

---

---

b. What method was used to collect the values?

---

---

c. What are some of the possible reasons why the data might not be representative of the true population? That is, what are some of the possible sources of bias?

---

---

d. Enter the appropriate values in the spaces provided and obtain printouts of a histogram and boxplot.

*Center:* Mean: \_\_\_\_\_ Median: \_\_\_\_\_

*Variation:* St. Dev.: \_\_\_\_\_ Variance: \_\_\_\_\_ Range: \_\_\_\_\_

*5-Number Summary:* Min.: \_\_\_\_\_  $Q_1$ : \_\_\_\_\_  $Q_2$ : \_\_\_\_\_  $Q_3$ : \_\_\_\_\_ Max.: \_\_\_\_\_

*Outliers:* \_\_\_\_\_

e. Describe the shape of the distribution. Is the distribution approximately bell-shaped?

---

---

f. What particular characteristics of the data are noteworthy?

---

---