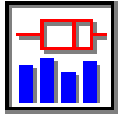


CHAPTER 1



Visualizing Univariate Data Analysis

CONCEPTS

- Exploratory Data Analysis, Central Tendency, Dispersion, Shape, Histograms, Dot Plot, Box Plot, Beam and Fulcrum, Stem and Leaf, Quantile Plot, Runs Chart, Descriptive Statistics

OBJECTIVES

- Recognize and interpret different types of histograms such as frequency, cumulative, and relative frequency histograms
- Realize how histogram setup can affect one's perception of the data
- Learn how to use common graphic data analysis tools
- Understand relationships between descriptive statistics and graphical displays

Overview of Concepts

At the beginning of a research project, statisticians feel that data should be examined and allowed to “tell its own story.” The exploratory phase of the research can help us select appropriate methods of analysis. This is important because some statistical tests are sensitive to “ill-behaved” data (especially outliers or strange distributions) that are often encountered. John W. Tukey of Princeton University pioneered an approach known as **Exploratory Data Analysis** (EDA). He invented or popularized many useful visual tools (such as the box plot, stem and leaf display, and quantile plot) and argued for reliance on robust statistics (such as quartiles) whenever possible. Traditional tools of data analysis (such as parametric statistics) and visual displays (such as histograms) are also useful in EDA.

Statisticians rely on a variety of visual and numeric tools to do this analysis. One of the first goals of the analysis is to understand **central tendency** (middle or “typical” data values), **dispersion** (“spread” of data values), and the **shape** (degree of symmetry and peakedness) of the sample.

The first step is usually to sort the data from low to high to see if any anomaly stands out. **Descriptive statistics** are usually then calculated. From these statistics we can assess the data’s central tendency, dispersion, and shape.

In order to visualize the data, statisticians use a variety of graphic displays. Visual alternatives to examining the entire sorted data array are the **dot plot** and the **stem and leaf** diagram. Visual displays of central tendency, dispersion, skewness, and peakedness are the **box plot** and the **beam and fulcrum**. The **quantile plot** provides a visual evaluation of the distribution of the sampled data (a 45-degree line for rectangular or uniformly distributed data, a lazy “S” for bell-shaped or normally distributed data, an upside-down “L” or a backwards “L” for skewed data, and an almost vertical line for very peaked data). The **runs chart** provides a visual check for randomness of the data in its original order (before sorting).

In creating a **histogram** the statistician must decide the number of intervals to display, the width of the histogram intervals, and whether the numbers on the horizontal axis are going to cover the range exactly or be aesthetically pleasing. A number of different histograms are available to the statistician. The most common is a *frequency histogram*. With this histogram we see the number of observations within each interval. A *relative frequency histogram* is used if we want to make inferences about the larger population since we see the proportion of the sample within each interval. The *cumulative histogram* is a histogram version of the quantile plot. A *frequency polygon* replaces the histogram bars with a line that connects the midpoints of the location of where the top of each histogram bar would be. It is generally used when there are a large number of intervals. The *standardized Z value histogram* is a frequency histogram which displays the standardized data (the difference between each data point and the sample mean divided by the standard deviation) that facilitates comparison between widely different data sets.

Only when we have explored the data thoroughly can we safely say that it is understood. We can draw conclusions or write a simple description of the data. If further analysis is required, the statistician is ready because he or she has allowed the data to “tell its own story.”

Illustration of Concepts

The 1996 NCAA Division I national basketball tournament paired 64 teams, whose winners advanced to the second round. Consider a brief **Exploratory Data Analysis** of the number of points scored by the 32 opening-round winners.

Figure 1 shows a sorted data list. The list showed that the winning scores ranged from 43 (Princeton) to 110 (Kentucky). Figure 2 shows a table of **descriptive statistics**. Regarding **central tendency**, the mean was 77.03, the median was 75, and the midhinge was 78 (average of Q1 and Q3). These measures are almost equal, suggesting near-symmetry. Regarding **dispersion**, the range was 67 and the standard deviation was 13.70. Regarding **shape**, skewness was -0.04 (which confirms the notion of symmetry) and kurtosis was 3.07 (which is consistent with normality). Figure 3 shows the symmetric **dot plot** with two possible outliers (the **stem and leaf** would confirm these results).

| Sorted Data List | | | |
|------------------|----------------|---------|--------|
| Row | Team | X Value | Std Z |
| 1 | Princeton | 43 | -2.473 |
| 2 | Miss. State | 56 | -1.378 |
| 3 | Temple | 61 | -1.159 |
| 4 | Va. Tech | 61 | -1.159 |
| 5 | Wake Forest | 62 | -1.086 |
| 6 | Boston College | 64 | -0.940 |
| 7 | Stanford | 66 | -0.794 |
| 8 | Cincinnati | 66 | -0.794 |
| 9 | Connecticut | 68 | -0.648 |
| 10 | Marquette | 68 | -0.648 |
| 11 | N. Mexico | 69 | -0.575 |
| 12 | Utah | 72 | -0.356 |
| 13 | Purdue | 73 | -0.283 |
| 14 | Iowa State | 74 | -0.210 |
| 15 | Texas Tech | 74 | -0.210 |
| 16 | Drexel | 75 | -0.137 |

Figure 1: Data List

| Descriptive Statistics | |
|------------------------|---------------------------|
| General | Central Tendency |
| Var = Points | Mean = 76.88 |
| Observations = 32 | Median = 75.00 |
| Low = 43 | Midrange = 76.50 |
| High = 110 | Midhinge = 78.00 |
| No Outliers | 5% Trim Mean = 77.00 |
| Dispersion | Shape |
| Quantile 1 = 66.50 | Skewness = -0.04 (p = 92) |
| Quantile 2 = 75.00 | Kurtosis = 3.07 (p = 55) |
| Quantile 3 = 89.50 | 66% within 1 SD |
| Std Dev = 13.70 | 94% within 2 SD |
| Coeff of Var = 17.8% | 100% within 3 SD |

Figure 2: Descriptive Statistics

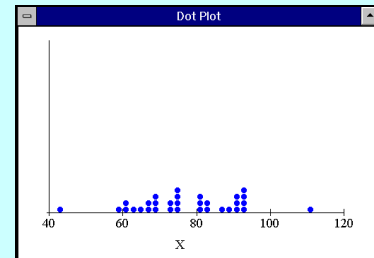


Figure 3: Dot Plot

Figure 4 shows a **box plot** and Figure 5 a **beam and fulcrum**. Both displays confirm that the data set is close to symmetric and about as peaked as a normal distribution. Similarity to a normal distribution is confirmed with the lazy “S” shape of the **quantile plot** in Figure 6. The **runs chart** in Figure 7 indicates that the data are random.

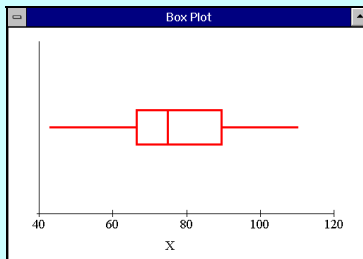


Figure 4: Box Plot

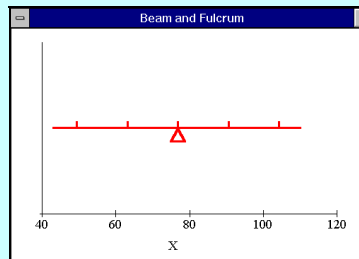


Figure 5: Beam and Fulcrum

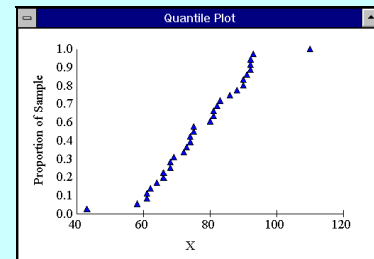


Figure 6: Quantile Plot

Two frequency **histograms** are shown. The histogram in Figure 8 uses six classes (based on Sturges' Rule) and uses nice labeling on the horizontal axis. The histogram in Figure 9 uses nine classes and covers the range exactly. It shows the two possible outliers. Which histogram is correct? Both of them! Full exploration of data requires trying many perspectives.

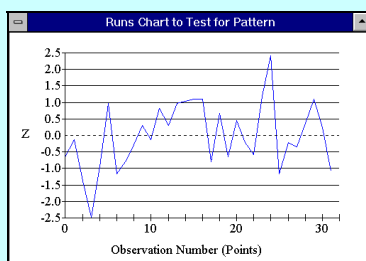


Figure 7: Runs Chart

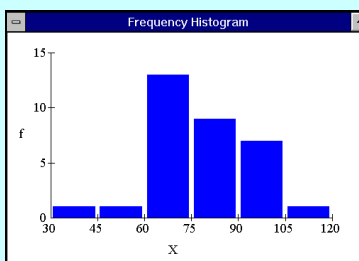


Figure 8: Nice Histogram

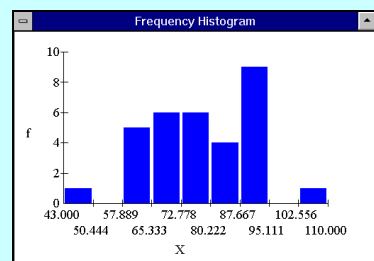


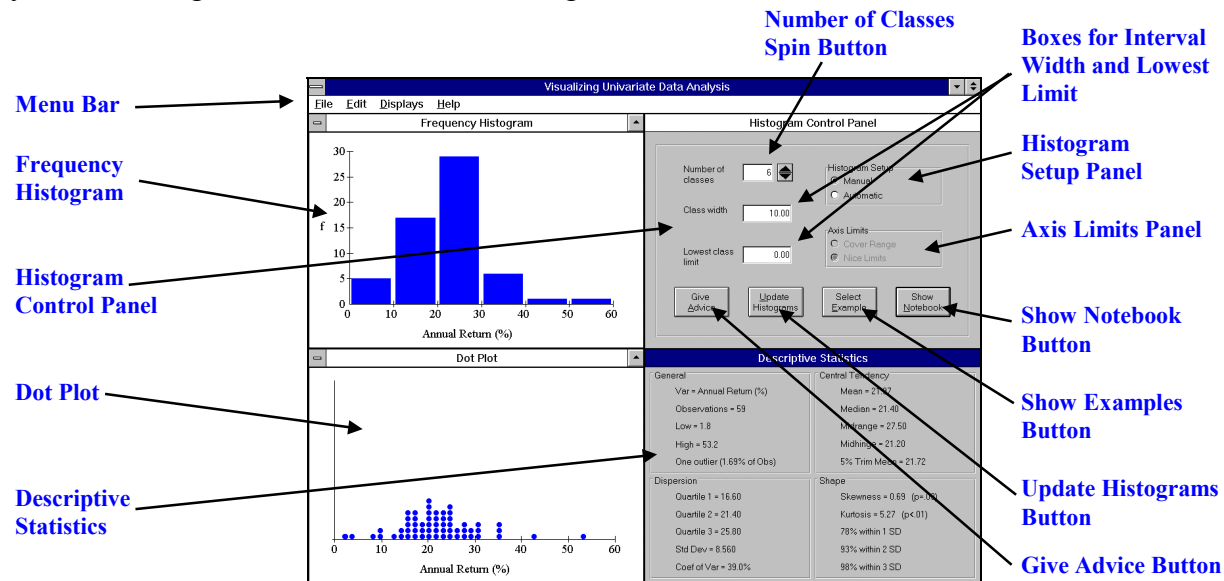
Figure 9: Exact Histogram

Orientation to Basic Features

This module familiarizes you with a variety of univariate data analysis tools. You can analyze a variety of different data sets by selecting them from the Notebook or create your own using the data editor.

1. Opening Screen

Start the module by clicking on the module's icon, title, or chapter number in the *Visual Statistics* menu and pressing the **Run Module** button. When the module is loaded, you will be on the introduction page of the Notebook. The **Introduction** and **Concepts** sections describe what will be covered in this module. Click on the **Examples** tab, click on **Financial**, select an example, and press **OK**. A Hint appears in the middle of the display. Read it and press **OK**. The upper left of the screen shows a frequency histogram. The Histogram Control Panel appears on the right. On the bottom left is the Dot Plot. On the bottom right is a table of Descriptive Statistics. Other features are controlled from the menu bar at the top of the screen. A flashing **Update Histograms** button will indicate when you have changed one or more control settings.



2. Control Panel

- The Histogram Setup panel contains option buttons to select if the histograms are to be automatically or manually created. Click on **Automatic**.
- The **Number of Classes** spin button is used to control the number of histogram intervals. Click on the spin button and watch how the shape of the histogram changes.
- The Axis Limits panel contains option buttons that are active when **Automatic** is selected in Histogram Setup. Given the number of classes, **Nice Limits** usually creates the smallest possible interval width that is divisible by 10, 2, or 5 (to an appropriate power of 10). The starting value is the largest possible value (smaller than the minimum value in the data set) divisible by the class width. This option generally creates aesthetically pleasing labels on the horizontal axis. **Cover Range** calculates interval (class) width by dividing the sample range by the number of classes. The lower limit is the smallest sample value. Although the class limits may be unaesthetic, unlike the **Nice Limits** option, it does not create beginning or ending intervals that extend beyond the data's range. Try selecting each option.

- d. Clicking on the **Give Advice** button describes the data and suggests an optimal number of intervals, the starting value for the first interval (class), and the width of each class. Click either **Cancel** or **Take Advice**.
- e. Click on the **Select Example** button to select a new data set from the list of examples. This button changes to **Select Databases** or **Edit Data** depending upon the origin of the data you are analyzing. See (f) and (g) below. It is a shortcut to the Notebook tab previously used.
- g. Click on the **Show Notebook** button to bring up the Notebook. There are two large databases that you can access with this module: U.S. States and World Nations. Select the **Databases** tab. Click on either **U.S. States** or **World Nations**. Each database is organized by categories. Click on the + symbol of any category to expand the category and list its variables. A complete discussion of the databases is given in the Introduction to this text.
- g. Click on the **Show Notebook** button to bring up the Notebook. You can use your own data by selecting the **Data Editor** tab, and pressing **OK**. A simple two-column spreadsheet appears. You can copy data from any spreadsheet program and paste it into the data editor; directions are provided under **Help**. You can save the data in *Visual Statistics* format by using the **Save** button or selecting **File** and **Save** on the menu bar. When you are finished press the **Exit and Use Data** button or **Exit and Discard Data** button or select these options under **File**. A complete discussion of the Data Editor is given in the Introduction to this text.

3. Other Displays

This module will make five different types of histograms, six different graphs, a table of descriptive statistics, and a list of the data. The Hint that was displayed as the module began said, “Click the right mouse button on a quadrant to select a different display.” Displays can also be changed by selecting a quadrant and clicking on **Displays** in the menu bar. Only the Histogram Control Panel cannot be replaced. Under **Histogram Types**, a frequency histogram, cumulative histogram, relative frequency histogram, frequency polygon, and standardized Z histogram can be selected. All of these except the standardized Z histogram are controlled by the Histogram Control Panel. Under **Graph Type**, besides a dot plot there are options for a stem and leaf plot, a box plot, a beam and fulcrum, a quantile plot and a runs chart. Under **Tables**, in addition to a numeric summary, there is an option for a data list.

4. Copying Graphs

Select **Copy** from the **Edit** menu (on the menu bar at the top of the screen), or the **Copy** option when you right click on a display, to copy the display. It can then be pasted into other applications, such as Word or WordPerfect, so it can be printed.

5. Help

Click on **Help** on the menu bar at the top of the screen. **Search for Help** lets you search an index for this module, **Contents** shows a table of contents for this module, **Using Help** gives instructions on how to use Help, and **About** gives licensing and copyright information about *Visual Statistics*. Close Help by selecting **Exit** from the **File** menu on the Help screen.

6. Exit

Close the module by selecting **Exit** in the **File** menu (or click  in the upper right-hand corner of the window). You will be returned to the *Visual Statistics* main menu.

Orientation to Additional Features

Manual Histogram

Display a frequency histogram and the descriptive statistics by using the right click feature. To create your own histograms select **Manual** in the Histogram Setup panel in the Histogram Control Panel. The **Class Width** and **Lowest Class Limit** boxes are now active and can be used. When creating your own histogram two rules should be used:

1. All of the data should be included in the histogram.
2. Neither the first interval nor last interval should be empty. Both should contain observations.

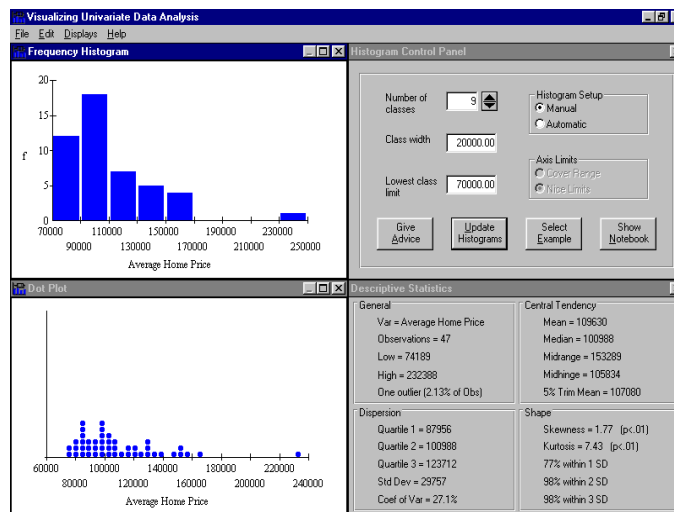
The maximum and minimum values in your data are listed as High and Low in the Descriptive Statistics table. In the **Lowest Class Limit** box enter a number no larger than the minimum value in your data. The approximate class width will be:

$$\frac{(\text{Maximum} - \text{Minimum})}{\text{Number of Classes}}$$

Enter this number, or a number slightly larger, in the **Class Width** box. If you create a histogram that does not cover the sample range, an error message will appear but the histogram will be displayed. To fix the problem, adjust one or both of the numbers.

Example

Below is an example of a manually created histogram with 9 intervals using the Home Prices data. It has 9 intervals of width 20,000. The first interval starts at 70,000. Note that 70,000 is less than the smallest observation (Low statistic = 74,189) and the last interval is larger than the largest observation (High statistic = 232,388). Therefore, all of the data points are included in the histogram. Also note that both the first and last intervals contain data (the first interval has a frequency of 12 and the last interval has a frequency of 1). Therefore, both rules are met.



Basic Learning Exercises

Name _____

Press the **Show Notebook** button and select the **Examples** tab. Click on **Consumer** and select **Home Prices**. Read the scenario.

1. a) Give the exact definition of the variable. b) What are its units of measurement? c) How do you think the *U.S. News* obtained the data? d) Are the data continuous or discrete?

Descriptive Statistics

2. Click on **OK**. Click the right mouse button to select your displays. Put the Numeric Summary in the lower right quadrant, the Data List in the lower left quadrant, and the Cumulative Frequency Histogram in the upper left quadrant. Using the sorted data list, which two cities had the lowest- and highest-priced homes? What were their standardized values?

3. What is the definition of an outlier? Get the definition from Help (click on **Help** on the menu bar and **Search for Help**. Type “outlier” and select the appropriate topic). Is either of these cities an outlier? Why or why not?

4. Give the value of each measure of centrality. If you are unfamiliar with a term, use Help. Are the measures different? What does this tell you about centrality and skewness?

Mean _____ Median _____ Midrange _____ Midhinge _____
 5% Trimmed Mean _____

5. Give the value of each statistic. How does each relate to dispersion?

Standard deviation _____ Range _____ First quartile _____
 Third quartile _____ Interquartile range _____

Graphs

6. Use the right click feature to view the dot plot, stem and leaf, and box plot in the three quadrants. The dot plot, and stem and leaf both illustrate the data. When would each be used?
7. How does the box plot illustrate centrality and dispersion? Use the Help feature for assistance.

Histograms

8. Use the right click feature to bring up the frequency histogram, frequency polygon, and relative frequency histogram in the three quadrants. All three of these displays have the same general appearance. Describe this appearance. Since each is similar, why would a statistician use one instead of another?
9. Use the right click feature to display the frequency, cumulative, and the standardized Z value histograms. Why would a statistician use either the cumulative or the standardized Z value histograms?
10. Click on **Automatic** in the Histogram Setup panel and **Cover Range in the** Axis Limit panel. Type the number “2” in the **Number of Classes** box and press Enter on your keyboard. Two histograms are redrawn. Use the spin button to increase the number of classes. a) Look at the frequency histogram; why does your impression of the shape of the data change as the number of classes changes? b) Look at the cumulative histogram; does its appearance change more or less and why? c) Why doesn’t the standardized histogram change?

Intermediate Learning Exercises

Name _____

Press the **Show Notebook** button and select the **Examples** tab. Click on **Consumer** and select **Home Prices**. Display the descriptive statistics, the box plot and the beam and fulcrum diagram.

Beam and Fulcrum

11. How does the beam and fulcrum illustrate centrality and dispersion? Use Help for assistance.

Shape Statistics

12. Record the value of each shape measure. What do these statistics tell us? If you do not know how to interpret these statistics, look up “Equations: Shape” in the Contents of Help.

| | | |
|---------------------|---------------------|---------------------|
| Skewness _____ | Kurtosis _____ | |
| % within 1 SD _____ | % within 2 SD _____ | % within 3 SD _____ |

13. The skewness measure tells us that this data is positively skewed. How can we see this skewness in the box plot, and beam and fulcrum diagram?

14. Kurtosis measures the peakedness of the distribution or the relative proportion of the distribution contained in its body versus its tails. How do the box plot and beam and fulcrum illustrate kurtosis? **Hint:** The home prices are slightly peaked (leptokurtic, kurtosis $>> 3$) and positively skewed. The lottery winners (press **Select Examples**, select **Lottery Winners**) are almost flat (platykurtic, kurtosis $<< 3$). The Presidents’ ages (press **Select Examples**, click **Next Page** in the lower right three times to get to the **Miscellaneous** page, and select **Presidents’ Ages**) are almost normally distributed (mesokurtic, kurtosis near 3).

Runs Chart and Quantile Plot

15. Use Presidents' Ages from the previous question. Right click and show a quantile plot. How can we tell from the quantile plot that this data is almost bell-shaped (normal distribution)? Select **Lottery Winners** (on **Consumer Examples** page). How can we tell from the quantile plot that this data is almost rectangular (uniform distribution)?

16. Return to the **Home Prices** example. Right click and show a runs chart. How does the runs chart show that this data is random? Select the **Technology** section of the **Examples** tab of the Notebook and look at **Importance of Job Abilities**. Then select the **Miscellaneous** section of the same tab and look at **Years Served by Popes**. How does the runs chart show that these two data sets are not random?

Histograms

17. Return to the **Home Prices** example. Right click on a diagram and show a frequency histogram. Set the **Number of Classes** spin button to 9. Click on **Automatic** in the Histogram Setup panel and **Cover Range in the** Axis Limit panel. Change the option button in the Axis Limit panel to **Nice Limits**. Did your impression of the data's shape change? Why or why not?

18. Change the number of intervals from 2 to 20 using both **Cover Range and Nice Limits**. With which setting does the appearance of the frequency histogram change less? Why?

19. Given these results when would a statistician use the **Nice Limits** option for histograms, and when would he or she use the **Cover Range** option?

Advanced Learning Exercises

Name _____

Histograms

20. Choose a variable from either the U.S. States or World Nations database. Click the **Give Advice** button, read its suggestions, and click **Take Advice**. Attach a copy of the histogram to this exercise. Describe the general appearance of the frequency histogram. List its noteworthy features (for example, does it have a single modal class?).

21. In the Histogram Setup panel select **Manual**. Construct a histogram of your own using nice labels, with the same number of classes used in exercise 20, but use a different class width and/or lower class limit. Copy the histogram and attach it to this exercise. Were you able to improve on the advice? What were the tradeoffs you had to consider? If you could vary the number of classes, could you have found nicer class limits?

22. Write down some principles for an algorithm to construct nice classes. Begin by looking in the Help file for Sturges' Rule. Discuss the logic of Sturges' Rule, which says that the correct number of classes is $1 + \log_2(n)$ where n is the sample size. **Hint:** Consider that $\log_2(2) = 1$, $\log_2(4) = 2$, $\log_2(8) = 3$, and so on.

Descriptive Statistics

23. Based on the descriptive statistics write a description of your data's shape. Explain what the skewness and kurtosis tell you. Next to the values of skewness and kurtosis is a p-value. What does this value tell you? If necessary, refer to the definitions and explanations in Help.

24. Two statistics reported in the descriptive statistics display and often not discussed in class or textbooks are the coefficient of variation and the 5% trimmed mean. Define each statistic and explain why each could be useful to a statistician.

Graphs

- [illegible]

Individual Learning Projects

Write a report on one of the three topics listed below. Use the cut-and-paste facilities of the module to place the appropriate graphs in your report.

1. Select a variable of interest and display its frequency histogram. Create five histograms with various numbers of classes, first using the **Nice Limits** option and then using the **Cover Range** option. In your paper give an exact definition of your variable, describe the shape of the distribution, explain in detail the effect the number of classes had on the appearance of your distribution, discuss the advantages and disadvantages of the **Nice Limits** and **Cover Range** options, and conclude with a decision (and reason) as to which histogram was most appropriate.
2. Select a variable of interest from one of the databases (U.S. States or World Nations) and do a complete Exploratory Data Analysis of the variable. Use the Learning Exercises and the Illustration of Concepts as a guide. Be sure to include a discussion of the units of measurement and possible inaccuracies or limitations in the data gathering process.
3. Select a variable of interest from one of the databases (U.S. States or World Nations) and explain how to construct the proper histogram manually to represent the data. You can use any other display (except automatic histogram creation) within the module to help in this task. Your paper should provide at least eight displays that illustrate the process you went through.

Team Learning Projects

Select one of the three projects listed below. In each case produce a team project that is suitable for an oral presentation. Use a large poster board(s) to display your results. Graphs should be large enough for your audience to see. Each team member should be responsible for producing some of the graphs. Ask your instructor if a written report is also expected.

1. Within one of the databases provided in this module or using another database of the team's choice, each team member should select a different variable and do an Exploratory Data Analysis of that variable. The objective of the project is to compare and contrast the central tendency, dispersion, shape, and general distribution of the variables selected. Include a discussion of the units of measurement and possible inaccuracies or limitations in the data gathering process for each variable.
2. A team of two or three should select a variable and do an in-depth analysis of constructing frequency histograms. Using **Automatic** classes with the **Nice Limits** option, create all possible histograms with 2 to 20 classes (make a display of each). Are the class limits always nice? For each that is not (but at least two), use the Manual feature to construct a histogram with nice limits that has the same number of classes (make a display of each). Of all the histograms, which histogram is most appropriate and why? Describe its overall shape. The objective of this project is to understand the importance of the number of classes in constructing histograms and the ease or difficulty in creating histograms with nice intervals.
3. A team of three should select three variables: one must be skewed with no outliers, one must be symmetric with no outliers, and one must contain outliers. For each variable the team is to construct 10 histograms varying the number of intervals and the type of labeling (five using the **Nice Limits** option and five using the **Cover Range** option). The objective of the project is to compare how different types of data have different issues involved in selecting the number of intervals and labeling. For each variable the team should select a "best" histogram.

Self-Evaluation Quiz

1. Which statistics offer robust measures of central tendency when outliers are present?
 - a. Mean, midrange, and mode.
 - b. Median, midhinge, and trimmed mean.
 - c. Mean, midrange, and midhinge.
 - d. Mean, mode, and quartiles.
 - e. None of the above.
2. The quartiles of a distribution are most clearly revealed in which display?
 - a. Box plot.
 - b. Dot plot.
 - c. Stem and leaf.
 - d. Frequency histogram.
 - e. Standardized Z histogram.
3. The frequency of outliers can always be seen on which display?
 - a. Box plot.
 - b. Standardized Z histogram.
 - c. Dot plot.
 - d. Frequency histogram.
 - e. None of the above.
4. The sorted stem and leaf display does *not* reveal
 - a. the modal groupings.
 - b. the mean.
 - c. all of the data values.
 - d. the mode(s).
 - e. the lowest and highest values.
5. Which display(s) will show the position of each data item?
 - a. Box plot.
 - b. Standardized Z histogram.
 - c. Dot plot.
 - d. Frequency histogram.
 - e. None of the above.
6. As we increase the number of classes in a histogram
 - a. central tendency becomes more obvious.
 - b. class interval width increases.
 - c. class intervals become rounder.
 - d. all of the above would be likely to happen.
 - e. none of the above would be likely to happen.

7. The mean and standard deviation are most easily seen in which display?
 - a. Box plot.
 - b. Beam and fulcrum.
 - c. Stem and leaf.
 - d. Dot plot.
 - e. Histogram.
8. If nice classes are used in a histogram
 - a. the scale range may not equal the true data range.
 - b. the histogram classes may be easier to interpret.
 - c. Sturges' Rule may be of secondary importance.
 - d. all of the above are likely.
 - e. none of the above is likely.
9. To ascertain the approximate range of a distribution we could use which display(s)?
 - a. Box plot.
 - b. Beam and fulcrum.
 - c. Stem and leaf.
 - d. Quantile plot.
 - e. All of the above.
10. The quantile plot for a sample from a normal population would have what shape?
 - a. L-shaped.
 - b. Backward Z-shaped.
 - c. Lazy S-shaped.
 - d. Inverted U-shaped.
 - e. Either a or b.
11. The runs chart is most useful to
 - a. check for randomness.
 - b. check for normality.
 - c. check for skewness.
 - d. check the interquartile range.
 - e. check none of the above.
12. Sturges' Rule would be most helpful in constructing which display?
 - a. Runs chart.
 - b. Beam and fulcrum.
 - c. Histogram.
 - d. Stem and leaf.
 - e. Dot plot.

Glossary of Terms

Beam and fulcrum A display that plots the position of the sample mean (the “fulcrum”) and the standard deviation points ($\text{Mean} \pm 1 \text{ SD}$, $\text{Mean} \pm 2 \text{ SD}$, $\text{Mean} \pm 3 \text{ SD}$, etc.). This display reveals skewness (the longer tail will indicate the direction of skewness) and kurtosis (the more standard deviations displayed along the beam the more peaked the data).

Bimodal When a sample contains two modes, the data are bimodal.

Box plot Five-number graphical display plotting the positions of the minimum, quartiles (first, second, third), and maximum along a scale representing data values.

Central tendency General reference to the attempt to characterize the location of the middle or “typical” values in a distribution (mean, median, midrange, midhinge, trimmed mean, modal class).

Coefficient of variation The ratio of the standard deviation to the mean. It is often multiplied by 100 so that it can be expressed as a percentage. It shows dispersion in relative terms. The formula fails if the mean is zero. It is unit-free and thereby allows comparison of samples with different means.

Cumulative histogram Histogram showing accumulated frequencies of data values. It begins at zero and rises to the sample size as we move to the right.

Descriptive Statistics A variety of statistics that are use to summarize or describe a data set.

Dispersion General reference to the “spread” of data values around the center of a distribution (variance, standard deviation, range, interquartile range, and coefficient of variation).

Dot plot Display of each data point as a dot along a horizontal scale. It is a kind of histogram with many bins. Dots are stacked when they are very close to the same horizontal position.

EDA Acronym for Exploratory Data Analysis.

Exploratory Data Analysis A broad term encompassing a variety of methods of looking at data to understand its characteristics. Its common acronym is EDA.

Frequency histogram Histogram showing the frequency of individual data values on the vertical axis and the data value along the horizontal axis.

Frequency polygon Frequency histogram that connects the midpoints of its bar tops and then omits the bars, producing a line graph.

Histogram Bar chart showing on the horizontal axis the values of a variable grouped into discrete classes (intervals or bins) and on the vertical axis the frequency of occurrence within each class.

Interquartile range The difference between the third and first quartile. It is robust to outliers and extreme values.

Kurtosis Measure of relative peakedness. The Pearson coefficient of kurtosis is the ratio of the fourth sample moment about the mean to the square of the second sample moment about the mean. If a distribution is unimodal and symmetric, then $K = 3$ indicates a normal, bell-shaped distribution (mesokurtic); $K < 3$ indicates a platykurtic distribution (flatter than normal, with shorter tails); and $K > 3$ indicates a leptokurtic distribution (more peaked than normal, with longer tails).

Mean Average of the sample data. It may be interpreted as the fulcrum (balancing point) of the sample data if the n data points are plotted along the X-axis. It is the most common measure of central tendency. It is not robust to outliers and extreme values.

Median The point along the X-axis that defines the upper and lower 50 percent of the sample. If n is odd the median is a member of the data set, while if n is even the median is the average of two adjacent values. It is a robust measure of central tendency because it is insensitive to outliers and extreme values.

Midhinge The average of the first quartile and third quartile. It is a robust measure of central tendency when the data contain outliers or extreme values.

Midrange The average of the smallest and largest observation. It is a measure of central tendency that is easily affected by outliers or extreme values.

Modal class Bar on the histogram that is higher than the bars on either side. Histograms may have two modes (bimodal) or more than two modes (multimodal). Axis tick marks indicate the limits of the modal class. The modal class is a more useful indicator of central tendency than the mode when the data are continuous or have a large range.

Mode The data value that occurs most frequently in a sample. It is not necessarily unique. If there are two modes, the data are called bimodal. The mode is most useful for discrete data with a small range.

Moment See **Sample moment about the mean**.

Outlier Any sample observation that is more than 3 standard deviations from the mean. In general, it is an observation that may be from a different population because it differs markedly from the others in the sample. In a normal population, an observation more than 3 standard deviations from the mean is expected to occur only about 27 times in 10,000 observations.

Percent within k standard deviations In a normal population, we expect 68.26% of the observations within 1 standard deviation of the mean, 95.44% within 2 standard deviations of the mean, and 99.73% within 3 standard deviations of the mean.

P-value for shape A large-sample test for departure from normality may be used to yield a two-tailed p-value for normal skewness (if $n > 8$) and kurtosis (if $n > 20$). Small p-values (for instance, below 0.05) tend to indicate departure from normality.

Q_1 , Q_2 , Q_3 Abbreviation for the first, second and third quartiles.

Quantile plot Cumulative frequency plotted against original data values. Normally distributed data will resemble a lazy S-shaped curve; skewed data will more closely resemble an upside-down “L” or a backwards “L”; uniformly distributed data will resemble a 45-degree line; and peaked data will have a steeper line.

Quartiles The first quartile (Q_1) is the point along the X-axis which defines the lower 25 percent of the sample, located at observation $(n + 1)/4$. The second quartile (Q_2) is the median (see above). The third quartile (Q_3) is the point along the X-axis that defines the upper 25 percent of the sample, located at observation $3(n + 1)/4$. If $n + 1$ is a multiple of 4, all quartiles are members of the data set; otherwise, we interpolate between adjacent observations.

Range The difference between the maximum and the minimum value in a sample. It is a measure of dispersion. It is not robust to outliers and extreme values.

Relative frequency histogram Frequency histogram that expresses frequencies as a fraction of sample size. Except for scaling of the frequency axis, it is identical in appearance to the frequency histogram.

Robust The quality of being unaffected by a particular factor. For example, the sample median is robust to the existence of an outlier (the sample median changes very little if an outlier is added to a data set).

Runs chart Plot of standardized data against the original order of entry in the sample, used to check for patterns that may occur when data are collected over time or in a systematic way.

Sample moment about the mean The k^{th} sample moment about the mean is obtained by taking the average of all of the deviations about the mean raised to the k^{th} power. The moments are used in the calculation of the Pearson coefficients of skewness and kurtosis.

Shape General reference to the degree of symmetry or asymmetry of a distribution, and to its degree of peakedness or flatness (see skewness and kurtosis).

Skewness Measure of relative symmetry. The Pearson coefficient of skewness is the ratio of the third sample moment about the mean to the square root of the second sample moment about the mean cubed. Zero indicates symmetry. Positive values show a long right tail. Negative values show a long left tail.

Standard deviation The square root of the sample variance. It is a measure of dispersion about the mean. It is measured in the same units as the mean (pounds, dollars, or whatever).

Standardized Z values Obtained from each observation when we subtract the mean and divide by the sample standard deviation. These transformed data are called Z values because they may be used to see how closely the sample resembles a standard normal distribution, to spot outliers, and to check symmetry about the mean. See **Percent within k standard deviations**.

Standardized Z histogram Histogram whose horizontal axis is scaled in terms of standard deviations instead of the original data units (usually from -4 to $+4$). It is useful for making comparisons with a standard normal distribution.

Stem and leaf Frequency tally in which each data point is tallied by displaying its “leaf” (its second significant digit) beside its “stem” (its first significant digit). The stem may be split into “high” and “low” stems if leaf frequencies become large. Data with more than two digits are expressed in appropriate units (10, 100, etc.). Leaf items are often sorted from low to high.

Sturges’ Rule The number of histogram classes should be $1 + \log_2(n)$ where n is the sample size (e.g., 4 classes for 8 observations, 5 classes for 16 observations, 6 classes for 32 observations). It is only a suggestion to avoid having too many or too few classes. If the data are skewed, Sturges’ Rule may not provide enough classes to reveal adequate detail.

Trimmed mean Mean calculated by omitting the highest 5% of the observations and the lowest 5% of the observations to mitigate the effect of extreme values. If 5% of the sample size is not an integer, we round the number of removed observations to the next smaller integer. It is generally robust to outliers and extreme values.

Univariate data Referring to any sample of observed values on a single variable (as opposed to bivariate or multivariate data).

Variance The sum of the squared deviations about the mean divided by the sample size minus one. The larger the variance the greater the dispersion or “spread” around the mean. It is not robust to outliers and extreme values.

Solutions to Self-Evaluation Quiz

1. b Do Exercise 2–5. Read the Overview of Concepts.
2. a Do Exercises 6–10. Consult the Glossary. Read the Overview of Concepts.
3. b Do Exercises 6–10. Consult the Glossary. Read the Overview of Concepts.
4. b Do Exercises 6. Consult the Glossary. Read the Overview of Concepts.
5. c Do Exercises 6–10. Consult the Glossary. Read the Overview of Concepts.
6. e Do Exercise 8–10. Do Exercise 10. Do Team Learning Project 2.
7. b Do Exercises 6–11, 15, 16. Consult the Glossary.
8. d Do Exercise 8–10, 17–19. Do Individual Learning Project 1 or Team Learning Project 2.
9. e Do Exercises 11, 15, 16. Consult the Glossary. Read the Overview of Concepts.
10. c Do Exercise 15. Consult the Glossary. Read the Overview of Concepts.
11. a Do Exercise 16. Consult the Glossary. Read the Overview of Concepts.
12. c Do Exercise 22. Consult the Glossary.