Chapter 4
Frequency Distributions

Histograms and Ogives

A **histogram** is a graph of a frequency distribution of numerical data for different categories of events, individuals, or objects. A **frequency distribution** indicates the individual number of events, individuals, or objects in the separate categories. Most people easily understand histograms because they resemble bar graphs often seen in newspapers and magazines. An **ogive** is a graph of a cumulative frequency distribution of numerical data from the histogram. A **cumulative frequency distribution** indicates the successive addition of the number of events, individuals, or objects in the different categories of the histogram, which always sums to 100. An ogive graph displays numerical data in an S-shaped curve with increasing numbers or percentages that eventually reach 100%. Because cumulative frequency distributions are rarely used in newspapers and magazines, most people never see them. Frequency data from a histogram, however, can easily be displayed in a cumulative frequency ogive.

This chapter will provide you with an understanding of the histogram and its corresponding ogive. You will gain this experience quickly without the work involved in data entry and hand computation. You will be able to view the histogram and cumulative frequency distributions for different sample data sets. Histograms and ogives have different shapes and vary depending on frequency. An ogive always increases from 0% to 100% for cumulative frequencies. The shape of a histogram determines the shape of its related ogive. A uniform histogram is flat; its ogive is a straight line sloping upward. An increasing histogram has higher frequencies for successive categories; its ogive is concave and looks like part of a parabola.

A decreasing histogram has lower frequencies for successive categories; its ogive is convex and looks like part of a parabola. A uni-modal histogram contains a single mound; its ogive is S-shaped. A bi-modal histogram contains two mounds; its ogive can be either reverse S-shaped or double S-shaped depending upon the data distribution. A right-skewed histogram has a mound on the left and a long tail on the right; its ogive is S-shaped with a large concave portion.

R. Schumacker and S. Tomek, *Understanding Statistics Using R*,
DOI 10.1007/978-1-4614-6227-9_4, © Springer Science+Business Media New York 2013

A left-skewed histogram has a mound on the right and a long tail on the left; its ogive is S-shaped with a large convex portion.

FREQUENCY R Program

The FREQUENCY R program can be used to display the histogram frequency distributions and ogive cumulative frequency distributions. To simplify the graphical display and provide similar comparisons between the types of histograms, all histograms in the program will have ten categories. The data for each category are not listed; rather the categories are numbered 1 to 10. You will be asked to enter the frequency for each of the ten categories and the frequencies must be integers greater than zero. The program will print a table listing the frequencies you specified, the relative frequencies, and the less-than-or-equal cumulative relative frequencies. The program prints a histogram and a corresponding ogive, which is output in a separate window (GSD2).

The part of the program that can be changed is a list of values relating to a score distribution observed in a given classroom. The length of this list does not matter; it is never specifically referenced in the program. The *Class* object is given a value for a vector of numbers using the **c** function that was introduced in Chapter 1. Each number within the vector is divided by the sum of all values within the vector. In the FREQUENCY program, the first processing loop is replaced by the simple **sum**(*Class*), which gets a total for all of the values, and this result is then divided into each of the values within the vector by simply typing *Class*/**sum**(*Class*). No additional step is necessary.

The next program line follows the same logic, only the cumulative sum (**cumsum**) of the vector is determined at each point and these values are divided by the overall sum of the values to give a vector of values labeled C*umRelFreq*. Scaling of the histogram height is performed next so that the histogram bars are not too small compared to the vertical scaling of the graph. The "if then else" clause is used to provide vertical scaling that will be either one tenth greater than the highest relative frequency, or 1 if the value is .95 or above. The **round** function is implemented to insure that the maximum value is set to an even tenth (**digits = 1**). The **barplot** function is used to draw the histogram of the relative frequencies with the **RelFreq** vector as the specified target. The **plot** function is used to draw the ogive of the cumulative relative frequencies with **CumRelFreq** as the target. The **par(mfrow = c(2,1))** command line permits both graphs to be printed, otherwise only the last graph (ogive) will be shown.

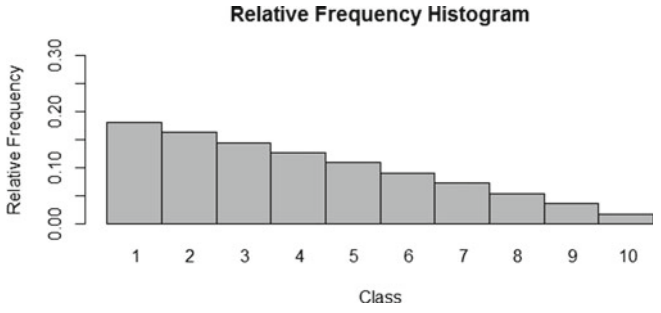
The last part of the FREQUENCY program builds a matrix of the class score distribution along with the associated relative frequencies and cumulative relative frequencies. The line beginning *TableData* **<- matrix** prepares the matrix and initializes all values within it to 0 and makes the dimensions of the matrix to be **length**(*Class*) rows and 3 columns. The **dimnames** keyword sets the labels for the

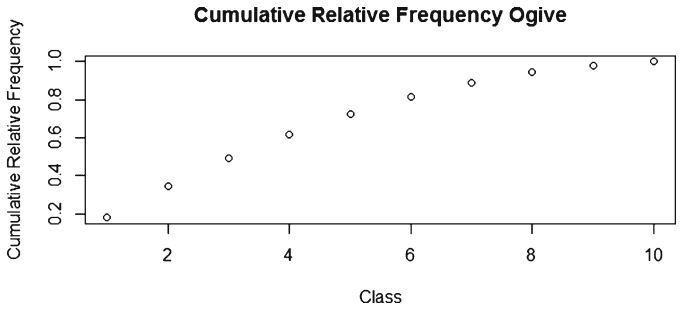
dimensions and will be used in later chapters with other vectors. The **for** loop iterates from 1 to the number of values within *Class* and assigns each row within the *TableData* matrix to the respective *Class* vector value, relative frequency, and cumulative relative frequency, rounding each frequency to three decimal places. You will see some error in the cumulative numbers due to the rounding of the cumulative values. The final line of the program simply prints out the *TableData* matrix.

You can change the values within the *Class* vector to obtain different shaped histograms and corresponding ogives. The original vector of 10 values breaks the score distribution into 10 intervals, but this can be changed to create histograms with greater resolution. You could comment out both lines of " **if** " and "**else**" statements that scale the histogram by prefixing them with " **#** " signs to see the effect of not scaling it properly to fit the plot; replace these statements with the ***PlotHeight* < -1** statement by removing the # sign in front of it. Some rounding error does occur in the program when summing the relative frequencies to obtain the cumulative relative frequencies.

FREQUENCY Program Output

```
          Freq.  Relative Freq.  Cum Rel Freq.
Class 1     50           0.182          0.182
Class 2     45           0.164          0.345
Class 3     40           0.145          0.491
Class 4     35           0.127          0.618
Class 5     30           0.109          0.727
Class 6     25           0.091          0.818
Class 7     20           0.073          0.891
Class 8     15           0.055          0.945
Class 9     10           0.036          0.982
Class 10     5           0.018          1.000
```





Histogram and Ogive Exercises

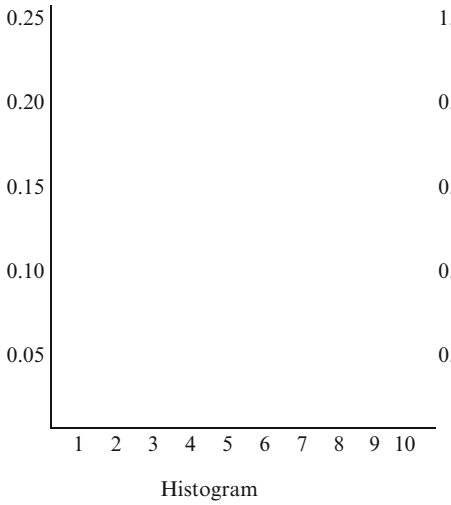
1. Run FREQUENCY program six times (a to f). Enter the frequencies listed for each type of histogram in the Class array statement. For each run, complete the frequency table and draw sketches of the histogram and corresponding ogive.

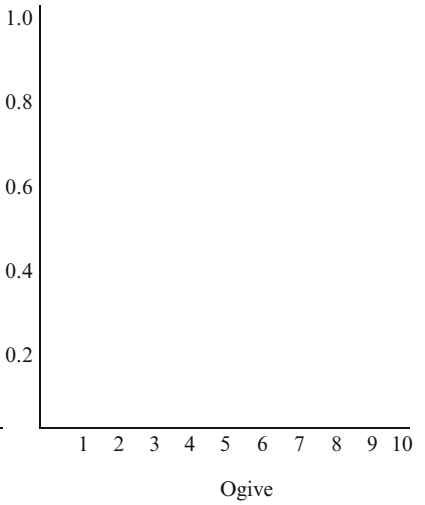
   a. A uniform histogram

CLASS	FREQ	REL FREQ	CUM REL FREQ
1	5	_____	_____
2	5	_____	_____
3	5	_____	_____
4	5	_____	_____
5	5	_____	_____
6	5	_____	_____
7	5	_____	_____
8	5	_____	_____
9	5	_____	_____
10	5	_____	_____

   b. An increasing histogram

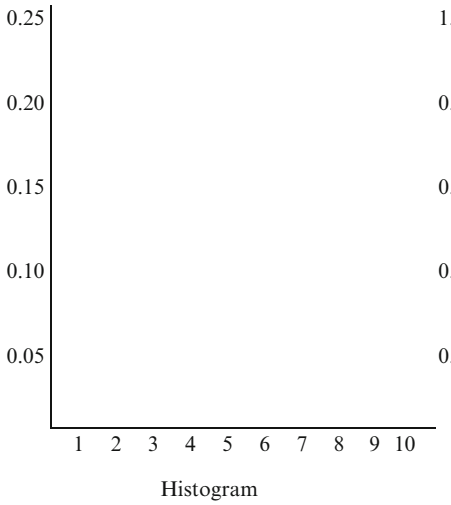
CLASS	FREQ	REL FREQ	CUM REL FREQ
1	10	_____	_____
2	12	_____	_____
3	14	_____	_____
4	16	_____	_____
5	18	_____	_____
6	20	_____	_____
7	22	_____	_____
8	24	_____	_____
9	26	_____	_____
10	28	_____	_____

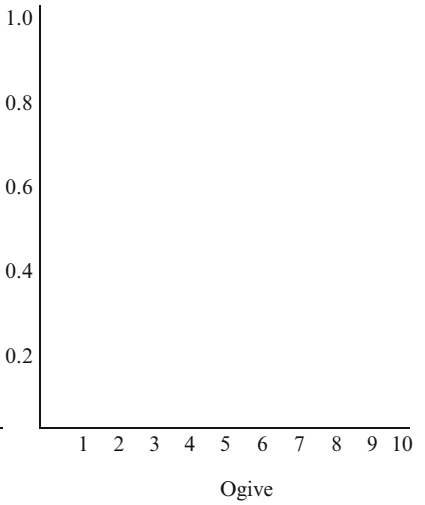




c. A decreasing histogram

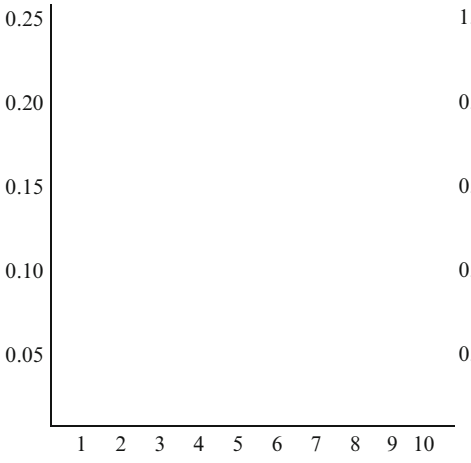
CLASS	FREQ	REL FREQ	CUM REL FREQ
1	50	_____	_____
2	45	_____	_____
3	40	_____	_____
4	35	_____	_____
5	30	_____	_____
6	25	_____	_____
7	20	_____	_____
8	15	_____	_____
9	10	_____	_____
10	5	_____	_____



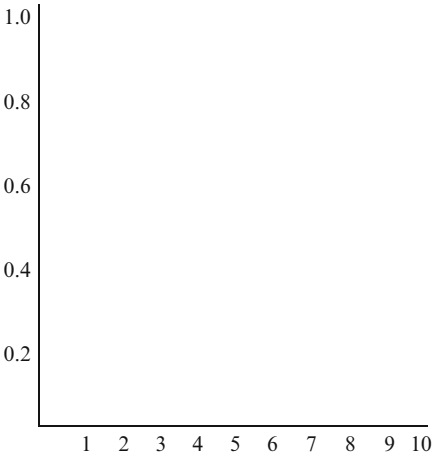


d. A unimodal histogram

CLASS	FREQ	REL FREQ	CUM REL FREQ
1	2	_____	_____
2	3	_____	_____
3	4	_____	_____
4	5	_____	_____
5	6	_____	_____
6	6	_____	_____
7	5	_____	_____
8	4	_____	_____
9	3	_____	_____
10	2	_____	_____



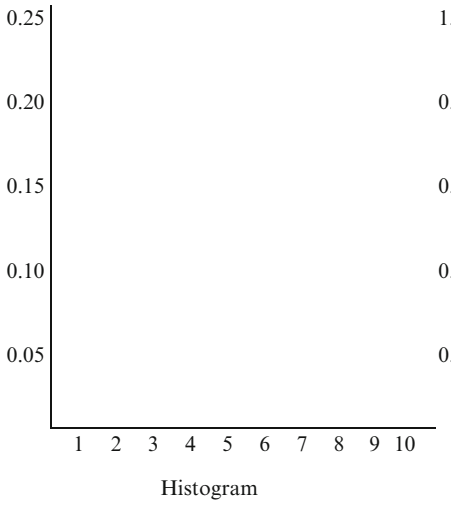
Histogram



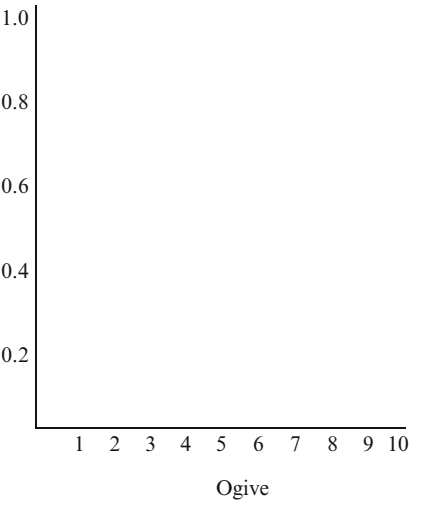
Ogive

e. A bimodal histogram

CLASS	FREQ	REL FREQ	CUM REL FREQ
1	6	_____	_____
2	5	_____	_____
3	4	_____	_____
4	3	_____	_____
5	2	_____	_____
6	2	_____	_____
7	3	_____	_____
8	4	_____	_____
9	5	_____	_____
10	6	_____	_____



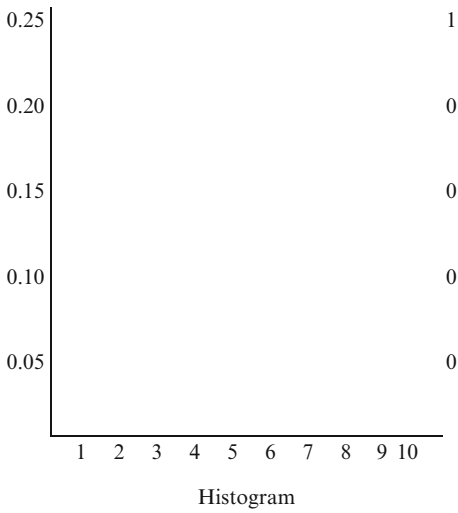
Histogram



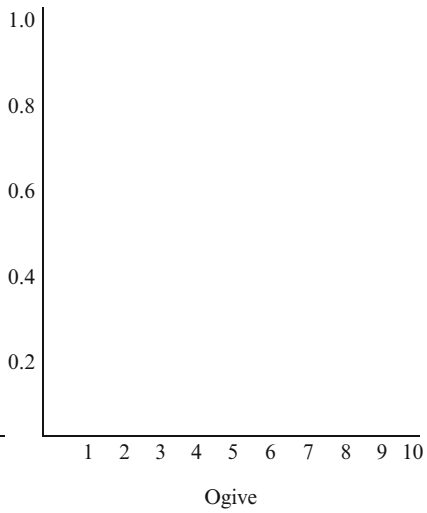
Ogive

f. A right-skewed histogram

CLASS	FREQ	REL FREQ	CUM REL FREQ
1	5	_____	_____
2	10	_____	_____
3	25	_____	_____
4	20	_____	_____
5	15	_____	_____
6	10	_____	_____
7	5	_____	_____
8	4	_____	_____
9	3	_____	_____
10	2	_____	_____



Histogram



Ogive

2. Describe the ogives for each of the following histograms.

HISTOGRAM	OGIVE
a. Uniform	
b. Increasing	
c. Decreasing	
d. Unimodal	
e. Bimodal	
f. Skewed right	

Population Distributions

The heights of adult men form a normal frequency distribution, i.e., a symmetrical distribution with one mode (**unimodal**). A similar population is formed by the heights of adult women. The **mode** is the score that occurs most often in a frequency distribution of data. However, because on the average women are shorter than men, the mean of the population of women's heights is less than the mean height of men. The **mean** is the score that indicates the average of all the scores in a frequency distribution. Imagine that a random sample of adults is chosen and the height of each person is determined. The sample consists of both men and women. In what way will the sample reflect the fact that it was drawn from a combination of two different populations?

In this chapter, you will learn about the shape of a histogram from a sample when the sample is drawn from a combination of two populations. The two populations in this chapter will be unimodal and symmetrical, i.e., normally distributed. To keep the chapter examples simple, the first population will always have a mean of 4. The second population mean can be changed. These mean values are also the modes of the data distributions. A histogram shows the frequency distribution of sample values. For large samples taken randomly from a normal population, the shape of the histogram is approximately normal. If a sample is taken from a combination of two populations, for which the means are far apart, the histogram of the sample will be **bimodal** (two modes). If the means of the two populations are close together, then the sample will be **unimodal** (one mode). In this case, the two populations are considered to be a single population. Large samples of data from a normal population yield a unimodal frequency distribution. The histogram is useful for summarizing data, that is, a bimodal sample could indicate that the sample is from two different populations and that the distance between the centers of two populations is related to the shape of the histogram of a sample selected from the combined populations.

COMBINATION R Program

The first time you run the COMBINATION R program, the second population will have a mean of 9. If you run the program again, you can select a different mean value for the second population. By changing the distance between the two population means, you can observe the effect of this distance on the shape of the histogram from a sample drawn from two combined populations. Each time you run the program, 500 randomly selected observations will be sampled, 250 from each of the populations. The program will print out a relative frequency table and a histogram for the sample.

The program combines the ease of random sampling with the ease of graphing. The program creates two normally distributed random samples centered at different means, with one mean fixed at 4 and the other mean different. The sample sizes can also be changed. The first population data are created from a normal distribution (**rnorm**) with a size of *SampleSize* [1], a mean of 4, and a standard deviation of 1. The values within this vector are rounded to two decimal places. The **round** function is placed within the **invisible** function so that the vector will not be printed while rounding. The second population data are created in the same manner with the mean at the value of the *CenterTwo* variable, a size of *SampleSize* [2], and is also rounded to two decimal places. The two populations are combined into a single population using the **c** function and are treated as a single vector.

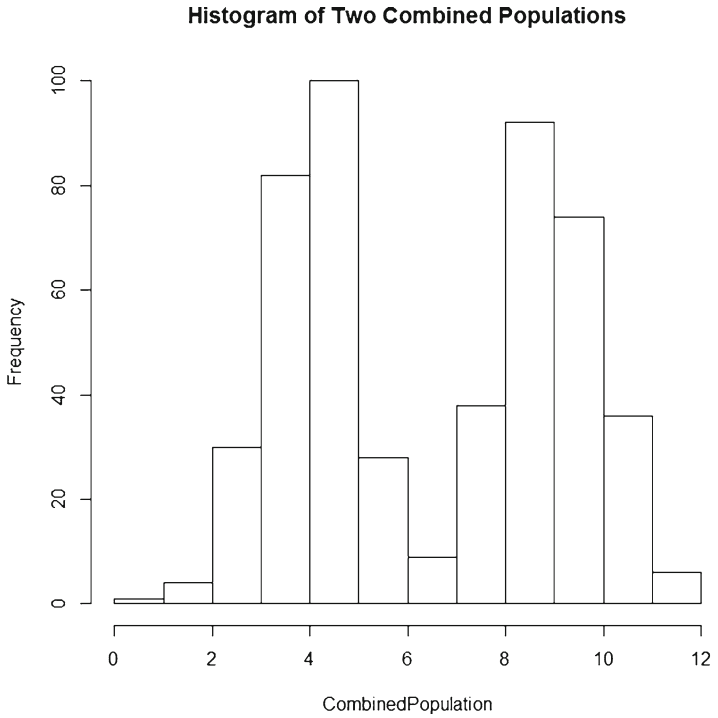
In order to create the relative frequency breakdown for various intervals within the combined population, it is necessary to use the **cut** function combined with the **factor** and **table** functions. The **cut** function breaks data into categories at given breakpoints, in this case at intervals of 0.5 from 0 to the largest value within the bimodal distribution of *CombinedPopulation*. The **factor** function takes the results of the **cut** function and assures that unused intervals are still included in the final results by choosing **levels** from 1 to twice the largest value within *CombinedPopulation* (this is because the intervals are only 0.5). These factors are then summarized by the **table** function with the number of points in each interval placed into the vector *FreqPop*. The next two lines create the labels for the intervals held in the *FreqPop* vector, with the first line being the value of the start of each interval and the second line being the value of the end of each interval. The **nsmall** keyword assures that at least one decimal place is preserved, even for whole numbers, to make the output easier to view.

The *FreqPop* vector is placed next into a **matrix** object to make it easier to display the results. The *FreqPop* vector is the input value, there are **length**(*FreqPop*) rows, and only 1 column. The dimension names are set to the interval start and interval end labels for the rows and "Rel Freq" for the single column. The **paste** function is handy when combining strings and vectors of values into a vector of labels. The next line with *FreqPopTable* by itself prints out the matrix. An alternative approach would be to use **print**(*FreqPopTable*), but since there were no other parameters in the **print** function needed for this display, it wasn't necessary.

Most of the program code creates and displays the matrix of relative frequencies. The **hist** function creates a histogram of any vector of values with relatively few keywords which is displayed in a separate output window (GSD2). *CombinedPopulation* is the target of the histogram. Default values (**,**) are specified for the y-axis and label size. The **main** keyword specifies the title to put at the top of the histogram. It is easier to notice differences in the two distributions using a histogram.

COMBINATION Program Output

```
            Rel Freq
0.0 - 0.5       0
0.5 - 1.0       1
1.0 - 1.5       0
1.5 - 2.0       4
2.0 - 2.5       9
2.5 - 3.0      21
3.0 - 3.5      36
3.5 - 4.0      46
4.0 - 4.5      53
4.5 - 5.0      47
5.0 - 5.5      20
5.5 - 6.0       8
6.0 - 6.5       5
6.5 - 7.0       4
7.0 - 7.5      13
7.5 - 8.0      25
8.0 - 8.5      44
8.5 - 9.0      48
9.0 - 9.5      37
9.5 - 10.0     37
10.0 - 10.5    24
10.5 - 11.0    12
11.0 - 11.5     6
11.5 - 12.0     0
```



COMBINATION Exercises

1. Run the COMBINATION program. A frequency table is printed for the combined populations. The first population will always be centered at 4. The second population will initially be centered at 9. The histogram is printed in a separate output window.

   a. Describe the shape of the distribution using the relative frequencies. _____
   _____

   b. Describe the shape of the distribution using the histogram. _____
   _____

   c. Are relative frequencies or histograms better in understanding the distribution of data? _____

2. Run the COMBINATION program again. A frequency table is printed for the combined populations. The first population will always be centered at 4. Set the second population mean at 12. A histogram is printed in a separate output window.

   a. Describe the shape of the distribution using the relative frequencies. _____
   _____

b. Describe the shape of the distribution using the histogram. _____
_____

c. Are relative frequencies or histograms better in understanding the distribution of data? _____

3. Run the COMBINATION program to find the smallest distance between the two means in which a bimodal distribution is still apparent (the second mean doesn't need to be a whole number, i.e., 8.5 and 5.75). If the means of the two populations are close together, then it is appropriate to consider that the two populations are similar.

a. What is the smallest distance between the means for which a bimodal distribution is still apparent? _____.

b. Statistical analyses are generally valid only for data that are randomly sampled from a symmetrical unimodal population distribution. What can you do to verify that a certain sample of data was randomly selected from a symmetrical unimodal (normal) population? _____
_____
_____

Stem and Leaf Graph

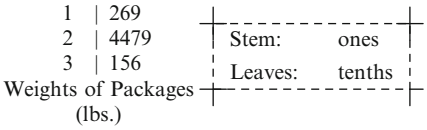
Graphical displays are often used to summarize data and usually help to uncover special characteristics of the data set. Histograms and **stem-and-leaf** plots are examples of graphical displays of data. The stem-and-leaf plots are particularly helpful in visualizing the **median** or middle value in data, the **range** or spread of data (distance between the lowest and highest data values), and **quartiles** (the first quartile is a score that separates the bottom 25% of the data in a frequency distribution from the other data and the third quartile is a score that separates the top 25% of the data in a frequency distribution from the other data). The **inter-quartile range** is the distance between the first and third quartile scores. It measures the range of scores in the middle 50% of the frequency distribution.

In graphing data, a decision must be made about how the data are grouped on the x-axis. A few large groups may obscure information and too many small groups may make summarization meaningless. This chapter addresses the problem of the number of groups in the context of stem-and-leaf plots (similar results are also true for histograms).

Consider the following data set of weights (in pounds) for UPS parcel post packages:

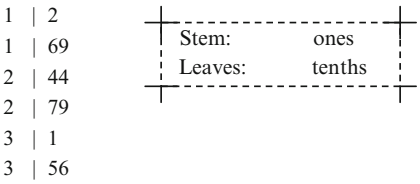
1.2	3.6	2.7	1.6	2.4
3.5	3.1	1.9	2.9	2.4

One stem-and-leaf plot for this data set is the following:



The numbers to the left of the vertical line form the **stem**. In the stem, the numbers represent the digit in the ones place. The numbers to the right of the line are the **leaves**. They represent the digit in the tenths place. Thus "1 | 269" represents the numbers 1.2, 1.6, and 1.9. There are only three groups in this stem-and-leaf plot, i.e., only one for each digit in the stem.

A finer subdivision could have been used in the stem-and-leaf plot. For example,



Here the first occurrence of 1 in the stem is used for data values from 1.0 to 1.4, and the second occurrence of 1 is used for 1.5 to 1.9. A similar approach is used for the other digits in the stem. There are six groups in this plot. Other units can be used for the stem and the leaves. A stem-and-leaf plot of adult heights in inches could have a stem in tens and leaves in ones. A plot of family incomes might have a stem in ten thousands and leaves in thousands.

The stem-and-leaf plots is a type of histogram where the median value in a stem-and-leaf plot can be viewed. The number of groups on a stem has an effect on the shape of the stem-and-leaf plot. The choice of a stem can also make the stem-and-leaf plot either unimodal or bimodal.

For some data sets, a change in the stem causes very little change in the shape of the stem-and-leaf plot.

STEM-LEAF R Program

In the STEM-LEAF R program the data sets will be student grades, which range between 0 and 100 inclusive. Two data sets of 50 student grades each are included in the program. The program will print two versions of a stem-and-leaf plot for each set of student grades. One plot has 11 groups and 6 groups, while the other plot has 15 groups and 8 groups. You will be asked to examine the two different stem and leaf plots to decide if reducing the number of groups gives a better distribution of scores that display for the median or middle value.

The program assigns two large vectors of scores to *Grades1* and *Grades2*. The scores are designed in a manner to demonstrate the importance of proper node assignment in stem-and-leaf plots and should *not* be modified. The stem-and-leaf plot is a simple command, **stem**. The variable, *LeafSpread*, is used as an argument in the command to determine the spacing of leaves.

The output and spacing is put between the script commands by use of the **cat** command with "\n\n" as the argument. The **cat** command simply prints characters to the standard output and the "\n" special sequence equates to a new line. The next line outputs the stem-and-leaf plot.

STEM-LEAF Program Output

```
   First Data Set
 [1]0  15 23 27 30 31 35 37 41 44 45 47 50 55 58 59 61 61 64
[20]64 66 68 69 70 71 71 72 72 73 74 74 85 85 85 87 88 88 88
[39]88 90 91 92 92 92 94 94 96 98 99 100

   Median=71

   11 Groups - First Data Set

   The decimal point is 1 digit(s) to the right of the |

    0 | 0
    1 | 5
    2 | 37
    3 | 0157
    4 | 1457
    5 | 0589
    6 | 1144689
    7 | 01122344
    8 | 55578888
    9 | 0122244689
   10 | 0

   6 Groups - First Data Set

   The decimal point is 1 digit(s) to the right of the |

    0 | 05
    2 | 370157
    4 | 14570589
    6 | 114468901122344
    8 | 555788880122244689
   10 | 0
```

NOTE: 11 Groups shows a better display of middle split for data at median = 71.

```
    Second Data Set
 [1] 30 31 36 38 42 44 45 47 50 53 53 54 56 58 58 59 61 62 63
[20] 64 65 66 67 69 70 71 72 74 75 76 77 77 80 80 83 83 85 87
[39] 88 89 91 92 93 94 95 97 97 99 100 100

    Median = 70.5

    15 Groups - Second Data Set

    The decimal point is 1 digit(s) to the right of the |
     3 | 01
     3 | 68
     4 | 24
     4 | 57
     5 | 0334
     5 | 6889
     6 | 1234
     6 | 5679
     7 | 0124
     7 | 5677
     8 | 0033
     8 | 5789
     9 | 1234
     9 | 5779
    10 | 00

8 Groups - Second Data Set

The decimal point is 1 digit(s) to the right of the |

     3 | 0168
     4 | 2457
     5 | 03346889
     6 | 12345679
     7 | 01245677
     8 | 00335789
     9 | 12345779
    10 | 00
```

NOTE: 8 Groups shows a better display of middle split for data at median = 70.5.

STEM-LEAF Exercises

1. Run STEM-LEAF program for GRADES1 with 11 groups and copy the stem-and-leaf plot here.

PLOT GRADES1

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

Stem: tens
Leaves: ones

   a. In what way does the stem-and-leaf plot resemble a histogram? _____
   _____
   _____

   b. Describe the shape. Is it symmetric, right skewed, left skewed, unimodal, or bimodal? _____

   c. The median is the middle score when an odd number of student grades are arranged in order. The median is the average of the two middle scores when an even number of student grades is arranged in order. What is the median student grade? _____

   d. What is the range of scores (distance between lowest and highest score)?
   _____

2. Run STEM-LEAF program for GRADES1 using 22 groups and copy the stem-leaf plot here.

PLOT GRADES1

```
0 |
0 |
1 |
```

Stem: tens
Leaves: ones

```
1 |
2 |
2 |
3 |
3 |
4 |
```

4 |
5 |
5 |
6 |
6 |
7 |
7 |
8 |
8 |
9 |
9 |
10 |
10 |

a. Describe the shape of the student grades in this plot. _____
_____

b. What is the median grade? _____
_____

c. What is the range? _____

3. Are there any characteristic differences in the two plots of GRADES1?
_____
_____
_____

4. Run the STEM-LEAF program for GRADES2 using 11 and 22 groups and enter the plots below.

PLOT GRADES2

0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

PLOT GRADES2

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

a. Is one of the stems more informative for the set of grades? _____
Why, or why not? _____

b. Compare the results. What does this illustrate about the effect of the number of stems (groups) in a stem-and-leaf plot? _____
_____

True or False Questions

Histograms and Ogives

T F a. The cumulative relative frequencies in a less-than-or-equal ogive are never decreasing.

T F b. Some ogives are straight lines.

T F c. An S-shaped ogive indicates a uniform histogram.

T F d. The sum of the relative frequencies in a histogram is always one.

T F e. A parabolic ogive can indicate an increasing histogram.

Population Distributions

T F a. The greater the distance between the means of two populations, the more pronounced the bimodal shape of the histogram.
T F b. If the means of two populations are close, then the histogram from the combined populations will have a single peak (unimodal) in the middle.
T F c. If a sample is taken from a combination of two populations which have means that are far apart, then the sample histogram will be bimodal.
T F d. As the means of two different populations get closer, the bimodal shape of the histogram is unchanged.
T F e. Large random samples from normal distributions have unimodal shaped histograms.

Stem and Leaf Graphs

T F a. The fewer the number of stems (groups) in a stem-and-leaf plot, the more informative the plot will be.
T F b. The number of stems in the plot does not affect the median value.
T F c. Skewness is usually apparent even if the number of stems is changed.
T F d. If a stem-and leaf plot is unimodal for one data set grouping, it will be unimodal when a finer subdivision of groups is used.
T F e. If a stem-and-leaf plot is rotated 90° counterclockwise, the plot is similar to a histogram.