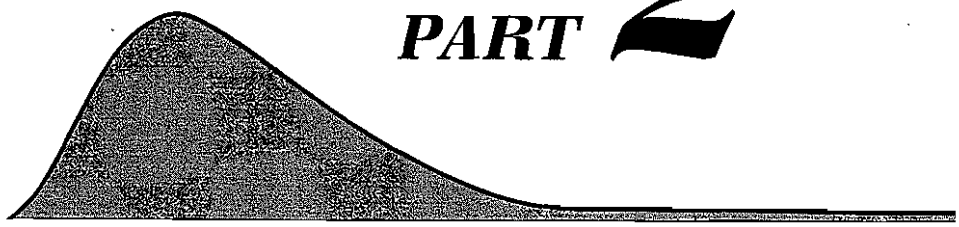
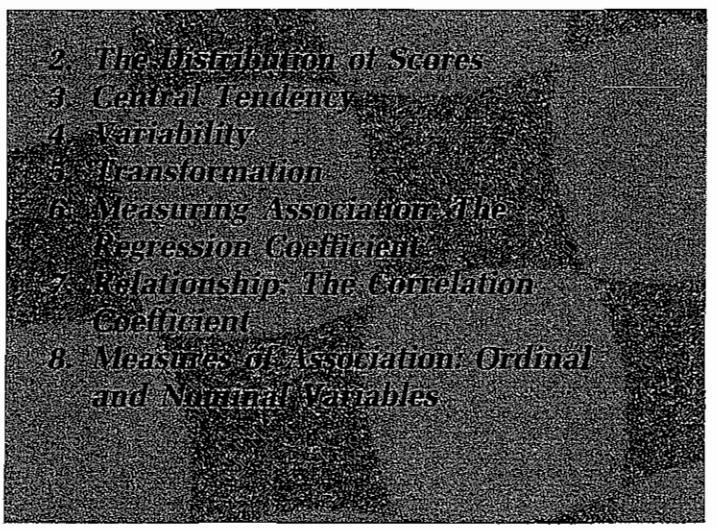
PART 2



Descriptive Statistics



2 The Distribution of Scores

I think that we all enjoy going to a shopping mall and sitting and watching people go by. It is amazing how many sizes and shapes we observe in a very brief period of time. Some people are very tall and muscular with no necks. Some are absurdly overweight and do not seem to walk but to waddle. Some are so pencil thin that it is difficult to understand what keeps their pants on. Some have fat legs and skinny arms. People certainly come in different shapes and sizes.

Though not as fascinating as people, groups of numbers or samples also come in different shapes and sizes. The technical and more general name for the shape of a sample of numbers is *distribution*.

Numbers by themselves can overwhelm us. One purpose of data analysis is to simplify the presentation of the numbers so that their meaning becomes more apparent. Systematically rearranged numbers make much more sense than raw data.

To facilitate the comprehension of how to understand the distribution of a set of numbers, consider a report by Smith (1980). She reviewed 32 studies on the gender bias of therapists.[1] These studies investigated whether clinical psychologists, psychiatrists, and counselors were prejudiced toward one gender or the other. In each study, one or more therapists advised male and female clients. It was possible to measure whether the advice given to the *male* clients was more positive than the advice given to *female* clients. For instance, one study that Smith reviewed asked whether career counselors encouraged males to enter higher-prestige occupations than females. Smith's

[1]Smith included 34 studies in her report. For both pedagogical and scientific reasons, two studies that she felt were of low quality are dropped. For studies with more than one outcome, the outcomes are simply averaged. Some of the studies in Smith's review did not compare the reaction of therapists to males and females but rather compared their reaction to a gender role stereotypic person versus a nonstereotypical person. In these studies bias toward the stereotypic person was coded positively and bias toward the nonstereotypic person was coded negatively.

index works as follows: A positive score indicates a male bias, zero indicates neutrality, and a negative score indicates a female bias. A small amount of gender bias would be indicated by a score of ±.2, a moderate gender bias by a score of ±.5, and a large bias by a score of ±.8. As an example, a score of .23 would indicate that therapists reacted more positively to males, but the difference is small. The scores for her 32 studies are contained in Table 2.1.

The Frequency Table and Histogram

It is difficult to make any sense immediately out of the 32 numbers as they are presented in Table 2.1. With a little study, however, some patterns do appear in the data. There seem to be just about as many studies with negative scores as positive scores. Thus the therapists do not seem to be consistently biased one way or the other. But this is much too coarse a judgment. A way of rearranging the numbers is needed so that their meaning can be better understood.

The first thing that can be done with sample data is to rank order them from smallest to largest. Recall that a *larger* negative number, such as −1.03, is farther from zero than a *smaller* negative number, such as −.56. The rank ordering of Smith's scores is as follows:

−1.03	−.23	.00	.14
−.56	−.22	.00	.23
−.40	−.10	.00	.24
−.36	−.03	.01	.29
−.31	.00	.01	.35
−.31	.00	.02	.56
−.23	.00	.05	.56
−.23	.00	.11	.60

The picture is now slightly less confused. Overall there is one more study that shows a bias favoring males than ones favoring females, but a fairly large number of studies show little or no gender bias.

There is still too much detail that gets in the way of understanding the

TABLE 2.1 **Data for the Studies of Gender Bias**

.29	.01	−.40	.00
.56	−.31	.00	.35
.00	.14	.02	.11
−.31	−.22	−.03	−.23
.56	.01	.00	−.56
−1.03	.00	−.23	.23
.00	−.10	.05	.00
.60	−.23	.24	−.36

numbers. One way of reducing the confusion is to remove some of the detail. It is not that crucial to know that one study has a score of .24 and another a score of .29. What is needed is to gather those studies having similar scores into groups or classes. Therefore the measure of gender bias is divided into intervals of a certain width, and the number of scores that fall within each interval is tallied. Each interval has a *lower limit,* which is the lowest possible score that can fall in the interval, and an *upper limit,* which is the highest possible score that can fall in the interval.

The *class interval* then defines the range of possible scores that can be a member of a given class. For instance, for the class interval .10 to .29, the lower limit is .10 and the upper limit is .29. So any score that falls between .10 and .29 would fall in that class. The complete set of class intervals for the gender bias data is as follows:

Class Interval	*Frequency*	*Relative Frequency*
−1.10 to −.91	1	3
−.90 to −.71	0	0
−.70 to −.51	1	3
−.50 to −.31	4	12
−.30 to −.11	4	12
−.10 to .09	13	41
.10 to .29	5	16
.30 to .49	1	3
.50 to .69	3	9
Total	32	99

In the first column are the classes, each with its lower and upper limit. The *class width* is defined as the difference between adjacent lower limits; in this case the class width is .20. The second column gives the *frequency* or number of cases in the class interval; for example, four studies have scores between −.30 and −.11. The final column gives the *relative frequency,* which is 100 times the frequency divided by sample size. For instance, for the interval −1.10 to −.91 the frequency is 1 and the sample size is 32 making the relative frequency 100 times 1 divided by 32 which equals 3, when rounded to the nearest whole number. Relative frequencies need not always be calculated, but they are especially informative when two different samples with different sample sizes are being compared.

The complete table of class intervals, frequencies, and relative frequencies is called a *frequency table.* The frequency table for Smith's scores shows that the scores cluster around zero, with about as many studies showing a male gender bias as a female gender bias. Scores between −.1 and +.1 can be considered as showing virtually no gender bias. The relative frequency of scores in the −.10 to .09 class is 41%, and therefore 41% of the studies show little or no gender bias.

Sometimes it is useful to compute the cumulative frequency. As the name implies, the *cumulative frequency* for a given class is the sum of all the

frequencies below or equal to the upper limit of that class. So for the interval of −.70 to −.51, the cumulative frequency is 1 + 0 + 1 = 2. It is also possible to compute the cumulative relative frequency by dividing the cumulative frequency by the sample size.

A graph of the frequency table is called a *histogram*. A graph has two lines that intersect at a right angle. These lines are called axes. The horizontal axis in a graph is called the X axis. The vertical axis is called the Y axis.[2] In a histogram the class intervals are on the X axis. The frequency, raw or relative, is on the vertical or Y axis. The resulting graph for the Smith data is presented in Figure 2.1. The histogram shows the shape of the 32 scores. The dominant feature of the graph is the peak in the middle of the numbers at about zero.

The basic steps to determine the shape of a sample of numbers are then

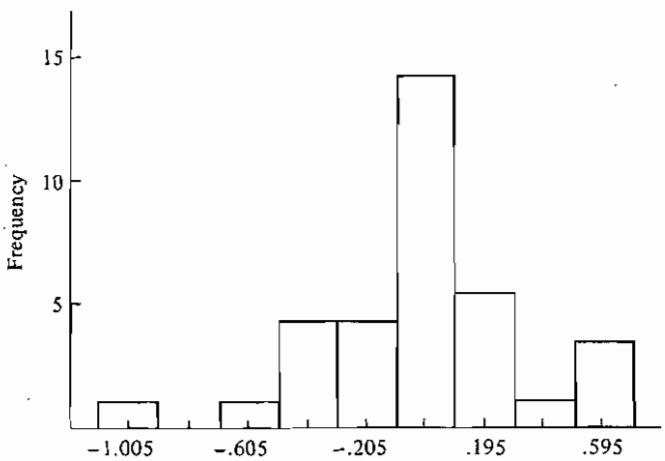
1. rank ordering the scores,
2. grouping the scores by class intervals, and
3. graphing the frequency table.

These steps are now discussed in more detail and generality.

Rank Ordering

This step is fairly simple. The numbers are ordered from smallest to largest. Although this step is not absolutely necessary, it is generally advisable to do it for the following reasons. First, the mere rank ordering already begins to describe the shape of the scores. Second, it makes the steps of creating a frequency table and a histogram easier because the frequency of scores that

FIGURE 2.1 **Histogram for 32 studies of gender bias of therapists.**



[2]The X axis is commonly called the abscissa and the Y axis the ordinate.

fall in a given interval can be quickly determined. Third, the computation of various statistical summaries that are discussed in the following chapters require a rank ordering of the numbers.

Grouping

To group the data, class intervals must be created. The first question is that of how many classes there should be or, correspondingly, how wide the class intervals should be. One reasonable guideline is to have between 8 and 15 classes. To determine the width of the class interval, the smallest score in the sample is subtracted from the largest score. This quantity is divided by 8 to determine the maximum class width and by 15 to obtain the minimum class width. For the gender bias studies, the smallest score is −1.03 and the largest is .60. The maximum class width is 1.63/8 or .20 and the minimum class width is 1.63/15 or .11. So the class width should be somewhere around .20 and .11.

The second guideline in determining class width is the sample size. The number of scores (the sample size) divided by the number of classes should be at least three. Stated differently, the average number of scores per class should be at least three. Given 32 scores, dividing by 15 classes, there are about only two observations per class. If there were 8 classes, there would be four observations per class. Thus .20 as the class width with about 8 classes seems like a good choice. It is perfectly permissible to have more than 15 classes when the sample size is large (more than 60), and with small sample sizes (less than 20) there should be fewer than 8 classes.

Once the class interval has been chosen, the lower limit of the lowest class interval or the *lowest lower limit* must be determined. The lowest class limit is the smallest score in the sample, rounded down to a convenient number. For the gender bias study, the smallest score is −1.03, and rounding down yields −1.10.

It is often necessary to try out a number of alternative class widths and lowest lower limits and see which works out best. Also certain features of the data must be considered when choosing these values. For instance, for the gender bias data, I made sure that zero, which indicates no bias, was near the middle of a class interval. To have zero near the middle of the interval was accomplished by having −1.10 as the lowest lower limit and not some other value such as −1.05.

In determining the class width, attention should be paid to the unit of the distribution. The *unit of the distribution* is the smallest possible difference between a pair of scores. For the gender bias study, the unit of the distribution is hundredths or .01. The lower limit must be in the same unit and the class width should be an integer multiple of the unit. For example, consider the Milgram study (1963), which is presented in detail in the next chapter. Some of the data are

300, 315, 450, 345, 330, 375

All of these scores are in multiples of 15, so the unit of the distribution is 15. The class widths should be in multiples of 15. Thus 30 and 45 are acceptable class widths, whereas 10 and 40 are not.

Graphing

Although the tally of the number of scores in an interval is helpful, a graph or histogram of the tally is even more informative. Though trite, it is still true that a picture is worth a thousand words.

In a histogram, the X axis or horizontal axis is the variable of interest divided into classes. The usual convention is to demarcate the X axis in a histogram by the class midpoints. The *midpoint of a class interval* is defined as half the sum of the upper limit of the class interval and the lower limit. So for the interval .10 to .29, the midpoint is (.29 + .10)/2, which equals .195. To have midpoints that do not have the additional trailing digit (the 5 in .195), it is advisable to have class intervals whose widths are odd. For the gender bias data, a class width of .15 or .25 might be a good alternative to .20.

The Y axis or vertical axis in a histogram is the frequency for a class. Either the frequency or the relative frequency can serve as the Y axis. Occasionally the cumulative frequency is used.

Outliers

The procedures that have been described are especially helpful in identifying outliers. An *outlier* is a score in the sample that is considerably larger or smaller than most of the other scores. In Chapter 4 a quantitative definition of "considerably" larger or smaller will be given.

As an example of outliers consider a second data set. DePaulo and Rosenthal (1979) had a number of persons, called targets, describe someone they liked. Forty persons, called perceivers, subsequently viewed videotapes of the targets' descriptions. The perceiver judged on a nine-point scale how much the target liked the person that the target was describing. These ratings of liking made by each perceiver were then averaged across the targets that the perceiver viewed. Although none of the targets lied in their descriptions, the perceivers were led to believe that some of the targets may have been lying. High average liking scores for a perceiver indicate that the perceiver correctly judged the targets as liking the persons that they were describing. Low scores indicate inaccuracy. The rank-ordered scores for the 40 perceivers are as follows:

1.37	5.70	6.16	6.55	7.05
2.21	5.70	6.16	6.63	7.05
3.21	5.80	6.26	6.63	7.26
4.65	5.95	6.35	6.65	7.30
5.05	6.05	6.40	6.73	7.35
5.55	6.05	6.42	6.75	7.35
5.65	6.10	6.45	6.89	7.89
5.70	6.15	6.52	7.00	7.94

The largest possible score is 9.00 and the smallest is 1.00. The numbers seem to cluster around 6, and there seem to be some rather small numbers.

Because the numbers are already rank ordered, the class width must now be determined. Because the largest score is 7.94 and the smallest is 1.37, the maximum class width is (7.94 – 1.37)/8 or .82 and the minimum class width is (7.94 – 1.37)/15 or .44. Two possible choices are .50 or .75. A .75 class width seems more reasonable than .50. First, .50 would result in 14 classes and an average of only 2.8 persons per class. Recall that at least three persons should be in each class. Second, because .75 is odd, the histogram would have class midpoints with two digits, not three as would happen if .50 were used. Because the lowest score is 1.37, rounding down to 1.00 yields the lowest lower limit. The resulting frequency table is

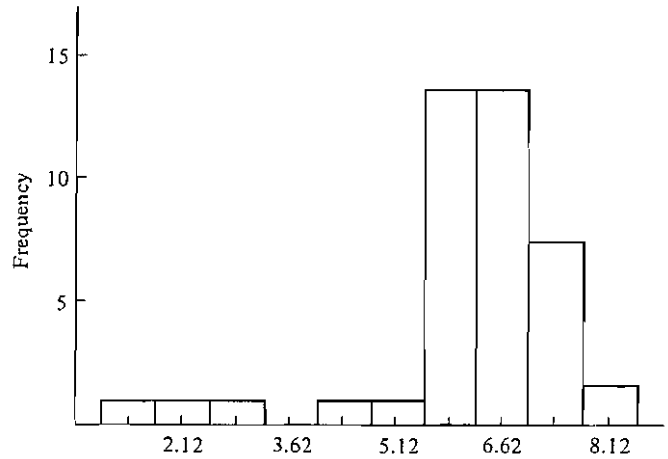
Class Interval	*Frequency*	*Relative Frequency*
1.00 to 1.74	1	2.5
1.75 to 2.49	1	2.5
2.50 to 3.24	1	2.5
3.25 to 3.99	0	0.0
4.00 to 4.74	1	2.5
4.75 to 5.49	1	2.5
5.50 to 6.24	13	32.5
6.25 to 6.99	13	32.5
7.00 to 7.74	7	17.5
7.75 to 8.49	2	5.0
Total	40	100.0

(In this case, for the relative frequency, it is sensible to round to the first decimal point because many of the numbers have a 5 at that point.) A histogram of the frequency table is presented in Figure 2.2.

An examination of both the frequency table and the histogram shows that the scores cluster near 6.25. Quite clearly the values 1.37, 2.21, and 3.21 are considerably smaller than the other 37 scores. They are all outliers.

After an outlier has been identified in the sample, it must be carefully considered why it is that the score is so atypical. Outliers are due to one of two reasons. First, they may be caused by a computational or data entry mistake. For instance, the recording of 6.42 as 642 would result in an outlier. Second, the outlier is not the result of a mistake, but rather it is generated by a different process than the other numbers. For instance, an abnormal physiological

FIGURE 2.2 **Histogram for the DePaulo and Rosenthal data.**



reading often indicates the presence of a disease. In industrial work, the presence of extremely large or small readings has often led to the discovery of a new manufacturing process. The outlier may be telling the researcher that the object is very different in some way from the others.

The outliers in the DePaulo and Rosenthal data are not the result of a computational mistake. Rather, they are attributable to different cognitive processes operating for the perceivers who obtained low scores. Before viewing the videotape the perceivers were led to believe in some of the descriptions the targets would be lying. That is, the target would be pretending to like someone they did not actually like. It is then plausible that the perceivers who have very low liking scores have these low scores because these perceivers felt that the targets were lying about liking the people that they were describing. Apparently most of the other perceivers took the targets at face value, but these three perceivers with low scores were very suspicious.

Although a frequency table and histogram were not necessary for the identification of the three outliers, they certainly facilitate that process. As will be seen in later chapters, the identification of outliers is an essential step in data analysis.

Features of Distributions

People have certain characteristic shapes: fat, thin, muscular, and so on. When looking at distributions of numbers, there is a parallel set of descriptive categories.

Before detailing these categories, a description of the figures that will be used to illustrate the characteristics of distributions must be presented. A

histogram of actual data, as in Figures 2.1 and 2.2, is quite jagged. However, if there were many scores and it was possible to have a very narrow class width, then the histogram would look quite smooth. When speaking about characteristics of distributions, it is helpful to consider these idealized distributions with large sample sizes and very narrow class widths. In practice, actual histograms only approximate the distributions that will be presented.

One of the first things examined in a histogram or frequency table are peaks in the distribution. A *peak* in a distribution is a frequency or set of adjacent frequencies that are larger than most of the other frequencies. So for the gender bias data, there is a peak at the class interval from −.10 to .09 because it has the highest frequency of 13.

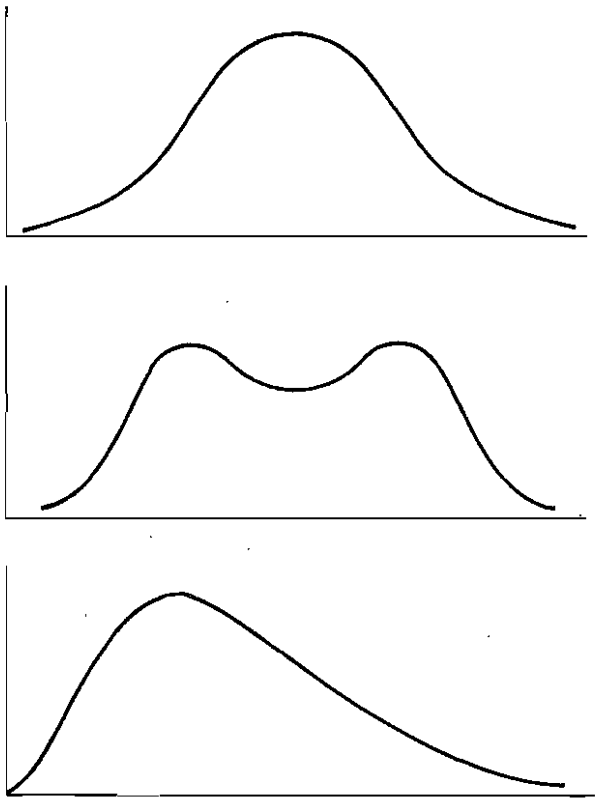
Whereas most distributions have a single peak, some have more than one peak. Distributions that have two peaks are called *bimodal* distributions. For instance, Hammersla (1983) measured the duration that college students played a game. She was interested in whether paying someone to play a game they already found fun would decrease their desire to play the game. Subjects were observed for a period of five minutes, and Hammersla measured the duration of game playing in seconds. Her histogram had two peaks, one near 0 seconds and the other near 300 seconds. Evidently subjects either played or did not play the game during the entire time period. When two different types of persons—such as those who like a game and those who no longer do—are mixed together, a bimodal distribution can result.

Examples of distributions with different types of peaks are presented in Figure 2.3. The top distribution has a single peak in the center. The middle distribution is bimodal, and the bottom has a peak on the left side of the distribution.

Besides the peaks, the low frequencies are also informative to the data analyst. For most distributions the smallest frequencies occur for the very large and very small values of the variable. For example, usually in a test few students do very well or very poorly. Most students fall in the middle. Because this pattern of dwindling frequencies at the extremes looks like a tail, the frequencies for the very large and very small values of a variable are called *tails*. The tails of distributions can be either fat or skinny. In Figure 2.4 are examples of distributions whose tails are either fat or skinny, or both. The top distribution has skinny tails, the middle one fat tails, and the one on the bottom has a fat left tail and a skinny right tail.

The size of the tails and the peak are related. For a distribution that is peaked in the center, the higher the peak in the distribution the skinnier are the tails, and the lower the peak the fatter are the tails. Distributions with a very high peak in the center and skinny tails are said to be *leptokurtic*. Distributions with a low peak in the center and fat tails are said to be *platykurtic*. In Figure 2.4 the upper distribution is said to be leptokurtic and the middle one is said to be platykurtic.

FIGURE 2.3 **Examples of different types of peaks.**



Some distributions have no peak at all. Distributions in which all scores are equally likely are called *flat* or *rectangular distributions*. The following set of scores have a flat distribution

3, 3, 4, 4, 5, 5, 6, 6, 7, 7

because each score occurs twice. The histogram for this flat distribution is contained in Figure 2.5. Perfectly flat distributions of naturally occurring variables are rarely encountered in the social and behavioral sciences. The distributions of most variables have a peak.

A distribution is said to be *symmetric* if its shape is such that if the data were regraphed, reversing the order of the class intervals, the shape would not change. Stated equivalently, a distribution is symmetric if its shape does not change when its mirror image is examined. If a perfectly symmetric distribution is plotted on a piece of paper and the paper is folded vertically, the two

FIGURE 2.4 **Examples of different types of tails.**

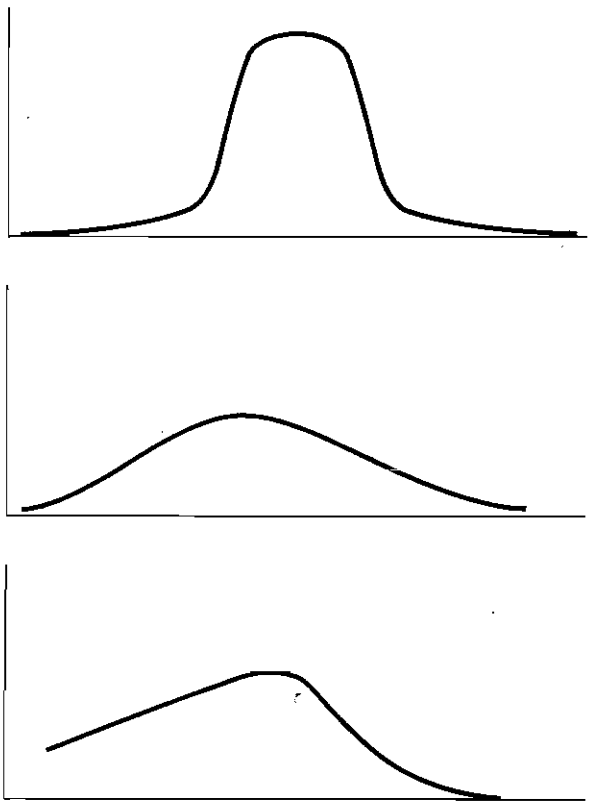
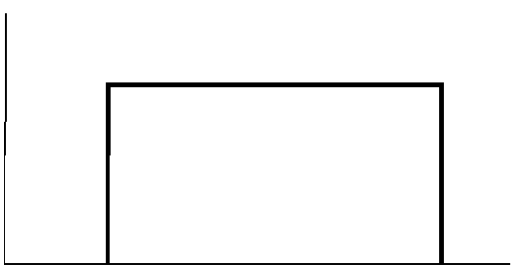


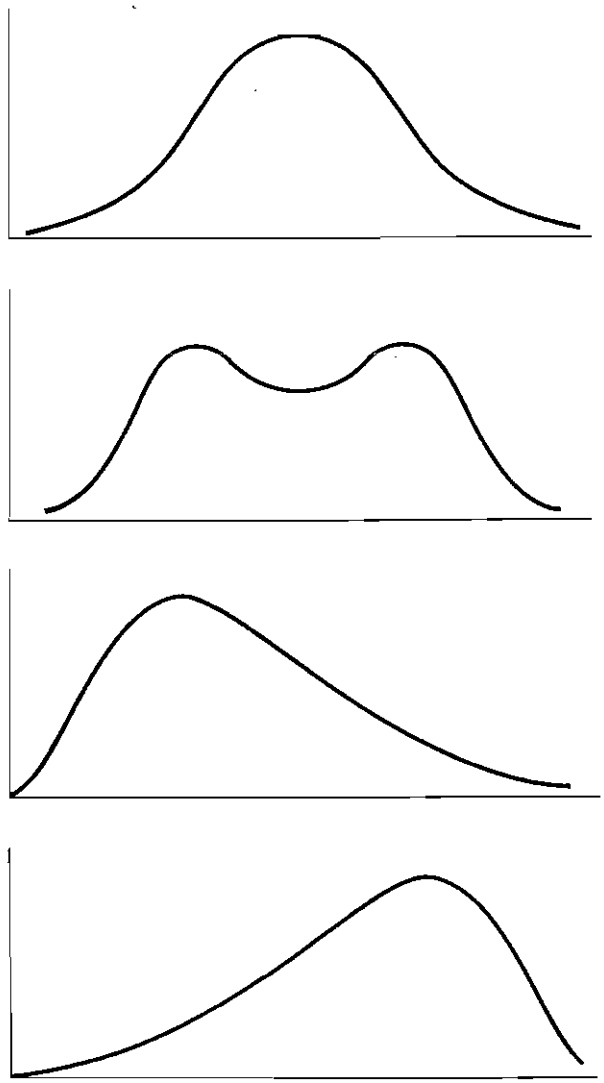
FIGURE 2.5 **Example of a flat distribution.**



sides of the histogram should completely coincide. Figure 2.6 shows examples of symmetric and asymmetric distributions. The top two distributions are symmetric, whereas the bottom two are asymmetric.

Some asymmetric distributions are said to be skewed. A *skewed distribution* is one in which the frequencies for the class intervals trail off in one direction but not the other. The direction of skew is determined by the skinny

FIGURE 2.6 **Examples of symmetric and asymmetric distributions.**



tail. The skinny tail can be on the left- or right-hand side along the *X* axis. If the trailing values are to the left of the peak, *negative skew* is present. Trailing values to the right indicate *positive skew*. Because many distributions have lower limits of zero, distributions with a positive skew are quite common. Examples of variables with a positive skew are income, number of home runs, traffic accidents per week, number of bar presses by a laboratory rat, and the number of children per family. Scores on an easy test have a negative skew and scores on difficult test have a positive skew. The bottom distribution in Figure 2.6 has a negative skew and the one above it has a positive skew.

One type of distribution—commonly assumed in data analysis—is called the normal distribution. The *normal distribution* is a symmetric distribution with a single peak in the center. The resultant shape is often called bell-shaped. The normal distribution is discussed in detail in Chapter 10.

Stem and Leaf Display

The frequency table and the histogram are traditional ways of arranging and displaying a sample of numbers. A newer, simpler, and more elegant procedure has been developed by John W. Tukey, called a *stem and leaf display*. In a stem and leaf display the classes are called *stems*. With a stem and leaf display there may be more stems than the 8 to 15 classes that are in a frequency table. The *stem* is essentially the lower limit of a class. So for instance, for the gender bias data, the stem could be the first two digits from the left of the score: –1.0, –0.9, –0.8 and so on. Then entered are the leaves, which are the trailing digit or the next digit to the right after the stem. So for the number 0.56 the stem is 0.5 and the leaf is 6. If there were any trailing digits to the right of the 6, they would be dropped.

The stems are arranged in a vertical order and to their right a vertical line is drawn. On the right of the line, the leaves are entered, one for each score in the sample. Each leaf is entered next to its stem. The stems can be separated by commas, but the common practice is not to do so. After the display has been completed, it is customary to redraw the display by rank ordering the leaves within each stem. The stem and leaf display for the gender bias data is presented in Table 2.2. Both the unranked (entering the leaves as they are presented in Table 2.1) and the ranked displays are presented. The stem and leaf display very clearly shows the peak at zero.

There are two features of the display that must be noted. First, zero has plus and minus stems of –.0 and +.0 stems. This looks odd but it is necessary to have categories for numbers from –.09 to –.00 and from +.00 to +.09. Second, in the ranked display on the right of Table 2.2, the leaves of the negative numbers appear to be ranked backward. They are not, because –.36 is less than –.31.

For the stem and leaf display of the DePaulo and Rosenthal data, the first

TABLE 2.2 **Unranked and Ranked Stem and Leaf Displays for the Gender Bias Data**

Unranked		Ranked	
−1.0	3	−1.0	3
−.9		−.9	
−.8		−.8	
−.7		−.7	
−.6		−.6	
−.5	6	−.5	6
−.4	0	−.4	0
−.3	116	−.3	611
−.2	2333	−.2	3332
−.1	0	−.1	0
−.0	3	−.0	3
.0	00110020500	.0	00000001125
.1	41	.1	14
.2	934	.2	349
.3	5	.3	5
.4		.4	
.5	66	.5	66
.6	0	.6	0

digit from the left in the number might be used as the stem and the second digit from the left as the leaf. It is common practice to just drop any other digit. And so, the ranked stem and leaf display for the DePaulo and Rosenthal data is

1	3
2	2
3	2
4	6
5	05677789
6	0011112344455666778
7	000233389

It should be noted that for each entry the third digit from the left was dropped. Thus, some information was lost, but that always happens in descriptive statistics. Why is the trailing digit dropped and why is the digit not rounded? For a stem and leaf display rounding really does not make much difference and so it is preferable to do the simpler thing by dropping the trailing digit.

Because for the stem and leaf display of the DePaulo and Rosenthal data, almost half the numbers pile up on 6, it is better to split the stems in half. Thus, for the stem 6.0, there are two stems of 6.0 and 6.5. The leaves are the second digit of the original scores. Here is the display with the stems split in half.

1.0	3
1.5	
2.0	2
2.5	
3.0	2
3.5	
4.0	
4.5	6
5.0	0
5.5	5677789
6.0	00111123444
6.5	55666778
7.0	0002333
7.5	89

Having more stems provides a better view of the distribution and shows the outliers more clearly.

After one stem and leaf display has been constructed, it is very simple to construct another. The stems could be separated into fifths: 6.0, 6.2, 6.4, 6.6, and 6.8. To prevent the display from being too long, the three outliers are not displayed. The display can be reworked with the stems split in fifths, as follows:

4.6	6
4.8	
5.0	0
5.2	
5.4	5
5.6	6777
5.8	89
6.0	001111
6.2	23
6.4	44455
6.6	66677
6.8	8
7.0	000
7.2	2333
7.4	
7.6	
7.8	89

The stem and leaf display has some important advantages over the more traditional frequency table and histogram. First, it is easier and faster to prepare than a frequency table or a histogram. Second, except for the dropped digits, the raw data can be recovered. Third, the display can be used to compute various statistical summaries.

Smoothing the Frequencies

In creating the class intervals, the class width and the lowest lower limit must be chosen. For instance, for the gender bias data, the class width was set at .20 and the lowest lower limit at −1.10. These choices can affect the shape that is shown in the histogram. That is, a class width of .30 and a lowest lower limit of 1.20 might considerably alter the shape of the histogram. This section of the chapter describes a procedure called *smoothing*, which takes a histogram and yields a shape that would be essentially the same regardless of the choice of class width or lowest lower limit.

Smoothing is a way of mathematically adjusting the frequencies to remove the rough edges. The class frequencies can be denoted f_1, f_2, f_3, and so on, where f_1 is the frequency for the lowest scores, f_2 for second lowest class interval, and so on. The *smoothed frequency* for a class interval is one-half the frequency for that interval plus one-quarter the frequency of each adjacent frequency. So for the gender bias data, f_4 is 4 and its adjacent frequencies are 1 for f_3 and 4 for f_5. The smoothed frequency for the fourth class interval is (.5)(4) + (.25)(1 + 4), which equals 3.25. In terms of a formula, the smoothed frequency for the class interval i is

$$.5f_i + .25(f_{i-1} + f_{i+1})$$

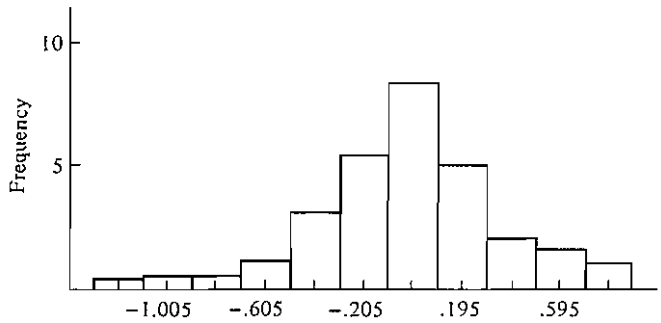
A problem arises when the first and last class frequencies are smoothed. They each have only one adjacent frequency. Two new classes must be added, one just before the smallest class interval and one just after the largest class interval. Before smoothing, these classes have zero frequencies. The frequencies of these two new classes can be smoothed by taking one-quarter of the adjacent frequency. The sum of the smoothed frequencies should always equal sample size.

The smoothed frequencies for the Smith study are as follows:

Class Interval	*Observed Frequency*	*Smoothed Frequency*	*Relative Smoothed Frequency*
−1.30 to −1.11	0	.25 = .5(0) + .25(0 + 1)	1
−1.10 to −.91	1	.50 = .5(1) + .25(0 + 0)	2
−.90 to −.71	0	.50 = .5(0) + .25(1 + 1)	2
−.70 to −.51	1	1.50 = .5(1) + .25(0 + 4)	5
−.50 to −.31	4	3.25 = .5(4) + .25(1 + 4)	10
−.30 to −.11	4	6.25 = .5(4) + .25(4 + 13)	20
−.10 to .09	13	8.75 = .5(13) + .25(4 + 5)	27
.10 to .29	5	6.00 = .5(5) + .25(13 + 1)	19
.30 to .49	1	2.50 = .5(1) + .25(5 + 3)	8
.50 to .69	3	1.75 = .5(3) + .25(1 + 0)	5
.70 to .99	0	.75 = .5(0) + .25(3 + 0)	2
Total		32.00	101

(The total relative frequencies is 101 because of rounding error.) Note how much smoother and simpler the frequencies are after smoothing. The distri-

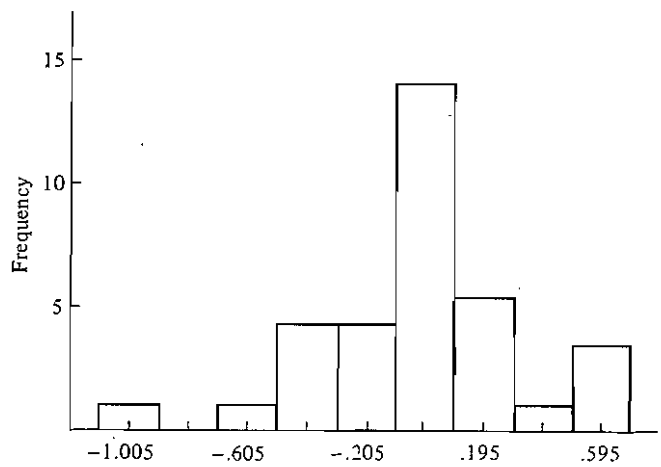
FIGURE 2.7 **Histogram of the smoothed frequencies for gender bias data.**



bution is clearly peaked near zero and is fairly symmetric. This can be seen best by graphing the smoothed frequencies in a histogram as in Figure 2.7 and comparing it with the unsmoothed histogram, which is reproduced in Figure 2.8.

Although smoothing does remove the rough edges in a frequency table, it also alters the actual frequencies. The effect of smoothing is that the peaks are lowered and the tails of the distribution are fattened. Moreover, smoothing can result in some anomalous results. For instance, if the DePaulo and Rosenthal data set is smoothed, it would happen that .25 person scored in the interval .25–.99. However, because 1.00 is the lowest possible score on the DePaulo-Rosenthal measure, the result makes no sense. Smoothing provides

FIGURE 2.8 **Histogram of the unsmoothed frequencies for gender bias data.**



a clearer picture, but at the cost of removing important details in the data and producing an anomalous class.

Summary

A sample of numbers can be summarized by grouping the numbers into a set of classes. Each class has a lower limit, which is the lowest possible score that can fall into the class and an upper limit. The *class width* is the difference between adjacent lower limits. The *class midpoint* is the average of the class's lower and upper limits. The number of scores that falls into a class is called the *class frequency*. The *relative frequency* of a class is the class frequency divided by the total number of scores. A table of class intervals, class frequencies, and relative frequencies is called a *frequency table*. A graph of the frequency table is called a *histogram*. In a histogram the X axis is the variable that is divided into classes, and the Y axis is frequency.

Highly deviant scores are called *outliers*, and they should be noted. The researcher should discover whether the outliers are due to a computational error or to a different process.

When a distribution is examined, the number and location of peaks should be noted. A distribution with two peaks is said to be *bimodal*. The *tails* of the distribution are the frequencies on the far left and right of the distribution. Distributions are characterized as having fat or skinny tails. A distribution with skinny tails is said to be *leptokurtic*. A distribution with fat tails is said to be *platykurtic*. A distribution with no peak at all is said to be *flat* or *rectangular*.

A distribution is said to be *symmetric* if, when the X axis is reversed, the shape does not change. A distribution with a peak on one side and a skinny tail on the other is said to be *skewed*. A *positive skew* has a skinny tail on the right, and a *negative skew* has a skinny tail on the left. A *normal distribution* is a unimodal and symmetric distribution that looks bell shaped.

A set of data can also be summarized by a *stem and leaf display*, which is a type of vertical histogram. The stems correspond to lower class limits and the leaves to the scores.

The shape of this histogram can be smoothed so that its true shape can be better revealed and so that chance fluctuation due to grouping is reduced.

Problems

1. Prepare the frequency table for the following data using 46 as the lowest lower limit and 5 as the class width:

   68, 73, 81, 76, 83, 96, 76, 83, 95, 81, 48, 56,
   75, 79, 90, 73, 76, 77, 84, 63, 68, 65, 62, 70

2. Draw histograms for the following shapes.
   a. a single-peaked asymmetric distribution
   b. an asymmetric bimodal distribution
   c. a flat distribution
   d. a unimodal leptokurtic distribution
   e. a symmetric bimodal distribution
3. Why must a single-peaked, symmetric distribution be peaked in the middle?
4. Can a flat distribution be bimodal?
5. For the data in Table 2.1 prepare a frequency table using a class width of .20 and a lowest lower limit of −1.20. Smooth the frequency table.
6. The following sample of numbers consists of the rents of apartments listed for rent in a university town.

298	288	300	300	385
310	230	385	325	375
350	300	265	340	310
285	260	425	275	300
320	275	300	310	285
260	375	295	250	275
385	310	380	265	285
310	300	310		

   a. Discuss the choice of class width and lowest lower limit.
   b. Construct a frequency table showing the frequency of rents in each class interval. Use 25 as the class width and 226 as the lowest lower limit.
   c. Describe the shape of the distribution.
7. For the data in problem 6, construct a histogram of the frequencies.
8. Below is the life expectancy in years at birth of males and females in the 30 most populous countries.

Male		*Female*	
68.7	65.2	76.5	71.4
45.8	57.6	46.6	61.0
48.6	59.9	51.5	63.3
59.2	51.6	62.7	53.8
36.5	69.0	39.6	76.9
68.3	41.9	74.8	40.6
47.5	57.6	47.5	57.4
69.0	72.2	74.9	77.4
63.0	62.8	67.0	66.6
37.2	53.7	36.7	48.8
56.9	66.9	60.0	74.6
49.8	69.7	53.3	75.0
53.6	53.7	58.7	53.7
64.0	67.8	74.0	73.0
43.2	41.9	46.0	45.1

a. Construct a frequency table for the males and another for the females using 35.0 as the lowest lower limit and 5.0 as the class width.
b. Compare the two frequency tables. Do women live longer than men?

9. For the data in problem 8 construct a stem and leaf display for both males and females.

10. Smooth the frequencies for the data in problem 8.

11. Prepare a frequency table for the DePaulo and Rosenthal data using 1.25 as the lowest lower limit and .75 as the class width.

12. Near the end of the fall semester, Harrison (1984) gave the UCLA Loneliness Scale to freshman women living in dormitories. The possible scores on the scale range from 20 to 80, higher scores indicating more loneliness. Below are the results for the women who were assigned to their dormitories.

Class Interval	*Frequency*
23 to 25	2
26 to 28	9
29 to 31	11
32 to 34	6
35 to 37	2
38 to 40	3
41 to 43	1
44 to 46	5
47 to 49	2
50 to 52	3
53 to 55	2

a. Compute the relative frequencies.
b. Smooth the observed frequencies.
c. Compute the relative smoothed frequencies.
d. Compute the cumulative frequencies (of the unsmoothed data).

13. Below are the loneliness scores for the women who chose their dormitories.

Class Interval	*Frequency*
20 to 22	1
23 to 25	1
26 to 28	1
29 to 31	5
32 to 34	4
35 to 37	4
38 to 40	4
41 to 43	3
44 to 46	2

a. Compute the relative frequencies.
b. Compare this distribution with the distribution in problem 12. Where does each distribution peak?

c. Are there relatively more residents with low loneliness scores (31 and below) in the assigned dorm group or in the group that chose their dorms? What about those with scores of 44 or higher? What about those with scores in the middle range, from 32 to 43?

14. For the following samples, state the unit of the distribution.

    a. 1.75, .25, 3.50, 4.50, 6.50, 7.00, 3.75
    b. 40, 120, 80, 160, 60, 100, 200
    c. 12, 18, 10, 26, 14, 18, 16, 14
    d. 1.33, 3.67, .67, 4.00, 3.33, 1.67

15. Prepare a stem and leaf display for data in problem 1 of this chapter.

16. Prepare a stem and leaf display for the data in problem 6 of this chapter.

17. Construct a histogram of the frequencies for the data in problem 12 of this chapter.