
Chapter 3

Descriptive Statistics

One of the two primary reasons for using statistical methods is to summarize and describe data, to make the information easier to assimilate. This chapter presents common methods of data description. The first section discusses statistical description through the use of tables and graphs. These tools provide a summary picture of the data.

We then present ways of describing the data with numerical measures. Section 3.2 defines statistics that describe the *center* of a collection of data—in other words, a “typical” measurement in the sample. Section 3.3 introduces statistics that describe the *variation* of the data about that center. The final section distinguishes between statistics describing samples and related parameters describing populations.

3.1 Tabular and Graphical Description

Example 3.1 State Murder Rates

We use the data in Table 3.1 to illustrate descriptive methods. This table lists all 50 states in the United States and their 1993 murder rates. The murder rate measures the number of murders in that state in 1993 per 100,000 population. For instance, if a state had 120 murders and a population size of 2,300,000, its murder rate was $(120/2,300,000) \times 100,000 = 5.2$. It is difficult to learn much by simply reading through the murder rates. We will use tables, graphs, and numerical measures to understand these data more fully. \square

TABLE 3.1 List of States with 1993 Murder Rates Measured as Number of Murders per 100,000 Population

Alabama	11.6	Louisiana	20.3	Ohio	6.0
Alaska	9.0	Maine	1.6	Oklahoma	8.4
Arizona	8.6	Maryland	12.7	Oregon	4.6
Arkansas	10.2	Massachusetts	3.9	Pennsylvania	6.8
California	13.1	Michigan	9.8	Rhode Island	3.9
Colorado	5.8	Minnesota	3.4	South Carolina	10.3
Connecticut	6.3	Mississippi	13.5	South Dakota	3.4
Delaware	5.0	Missouri	11.3	Tennessee	10.2
Florida	8.9	Montana	3.0	Texas	11.9
Georgia	11.4	Nebraska	3.9	Utah	3.1
Hawaii	3.8	Nevada	10.4	Vermont	3.6
Idaho	3.5	New Hampshire	2.0	Virginia	8.3
Illinois	11.4	New Jersey	5.3	Washington	5.2
Indiana	7.5	New Mexico	8.0	West Virginia	6.9
Iowa	2.3	New York	13.3	Wisconsin	4.4
Kansas	6.4	North Carolina	11.3	Wyoming	3.4
Kentucky	6.6	North Dakota	1.7		

Frequency Distributions

Rather than simply listing all the separate observations, as Table 3.1 does, we can summarize the data. A common summary method divides the measurement scale into a set of intervals and totals the number of observations in each interval. A *frequency distribution*, defined next, does this.

Frequency Distribution

A **frequency distribution** is a listing of intervals of possible values for a variable, together with a tabulation of the number of observations in each interval.

To construct a frequency distribution for murder rate, for example, we divide the possible murder rate values into separate intervals. We then count the number (frequency) of states in each interval.

We must first select a set of intervals for murder rate. Or computer statistical software such as SAS or SPSS chooses them for us. SAS, for instance, uses the intervals {0–2.9, 3.0–5.9, 6.0–8.9, 9.0–11.9, 12.0–14.9, 15.0–17.9, 18.0–20.9} for the number of murders per 100,000 population. Counting the number of states with murder rates in each interval, we get the frequency distribution shown in Table 3.2. It is clear from looking at this frequency distribution that considerable variability exists in statewide murder rates, with one state being considerably higher than the rest. As with any summary method, some information is lost as the cost of achieving some clarity. The fre-

TABLE 3.2 Frequency Distribution of Murder Rates for the 50 States

Murder Rate (No. Murders per 100,000)	Frequency (No. States)
0.0–2.9	5
3.0–5.9	16
6.0–8.9	12
9.0–11.9	12
12.0–14.9	4
15.0–17.9	0
18.0–20.9	1
Total	50

quency distribution does not identify which states have low or high murder rates, nor are the exact murder rates known.

The intervals of values for the categories in frequency distributions are usually of equal width; the width equals 3 in Table 3.2. The intervals should include all possible values of the variable. In addition, any possible value must fit into one and only one interval; that is, they should be *mutually exclusive*.

The number of intervals in a frequency distribution depends both on the judgment of the researcher and on the number of observations to be classified. Usually, the larger the number of observations, the greater the number of intervals used. If too many intervals are used (say, more than 15), they are so narrow that the information presented is difficult to digest, and an overall pattern in the results may be obscured. If very few intervals are used, however, too much information may be lost through pooling together observations that are not very similar. Follow this general guideline: The interval should not be so wide that two measurements included in it have a difference between them that is considered major. To summarize annual income, for example, if a difference of \$5000 in income is not considered especially important, but a difference of \$10,000 is somewhat notable, we might choose intervals of width less than \$10,000, such as 0–\$7999, \$8000–\$15,999, \$16,000–\$23,999, and so forth.

Relative Frequencies

Frequency distributions are informative, but it is easier to make comparisons between different intervals using *relative frequencies*.

Relative Frequency

The *relative frequency* for an interval is the proportion of the sample observations that fall in that interval.

TABLE 3.3 Relative Frequency Distribution and Percentages for Murder Rates

Murder Rate	Frequency	Relative Frequency	Percentage
0.0–2.9	5	.10	10.0
3.0–5.9	16	.32	32.0
6.0–8.9	12	.24	24.0
9.0–11.9	12	.24	24.0
12.0–14.9	4	.08	8.0
15.0–17.9	0	.00	0.0
18.0–20.9	1	.02	2.0
Total	50	1.00	100.0

The relative frequency equals the number of observations in an interval divided by the total number of observations. For instance, for the murder rates, the relative frequency for the first interval in Table 3.2 is $5/50 = .10$; that is, 5 states out of 50, for a relative frequency of .10, had murder rates between 0 and 2.9. The relative frequency is a proportion—a number between 0 and 1 that expresses the share of the observations falling in that interval. A listing of these, by interval, provides a *relative frequency distribution*. We construct the relative frequency distribution for the data on murder rates in Table 3.2 by dividing each frequency by 50, the total number of states. Table 3.3 shows it.

More often, relative frequencies are recorded as percentages rather than proportions. A percentage is simply a relative frequency multiplied by 100; that is, the decimal place is moved two positions to the right. For example, $5/50 = .10$ is the relative frequency for the interval 0–2.9, and $100(.10) = 10$ is the percentage. Table 3.3 also shows the relative frequency distribution as a percentage distribution.

The total sum of the proportions equals 1.00, and the sum of the percentages equals 100. The process of rounding may lead to slightly different totals, such as 100.1 or 99.9. When presenting relative frequencies in a table, always include the total number of cases upon which they are based. Obviously, the statement that 60% of a sample of 1000 individuals favor a decrease in the national defense budget is much more striking than the same statement derived from a sample of 5 individuals.

Histograms and Bar Graphs

A graph of a frequency distribution for a quantitative variable is called a *histogram*. A bar is drawn over each interval of numbers, with height of the bar representing the relative number of observations in that interval. Figure 3.1 is a histogram for the murder rates, using the intervals in Table 3.2.

Although guidelines exist for drawing histograms (see Tufte, 1983), it is primarily a matter of common sense. As with frequency distributions, if too few intervals are used, too much information is lost or obscured. For example, Figure 3.2 is a histogram

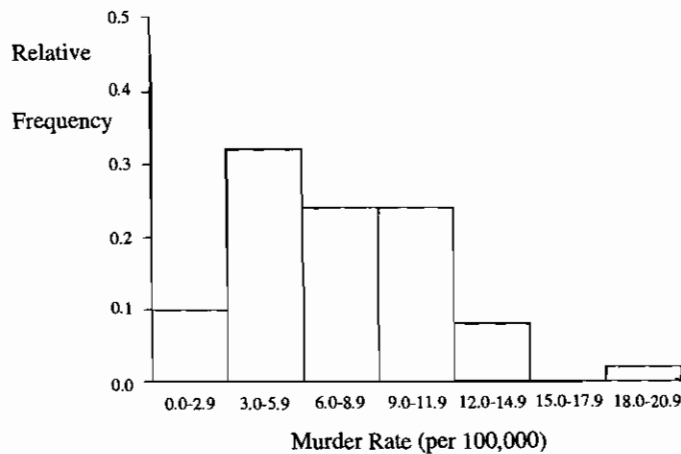


Figure 3.1 Relative Frequency Histogram for Murder Rates

of murder rates using the intervals 0.0–6.9, 7.0–13.9, 14.0–20.9. This is too crude to be very informative. On the other hand, the histogram is very irregular if too many intervals are used relative to the size of the data set. Most statistical software makes it simple to request histograms of data, and the software automatically chooses intervals that are sensible.

Relative frequencies are useful for data of any type. For categorical (nominal or ordinal) variables, instead of intervals of numbers we use the categorical scale for the variable. In that case, the graph of the relative frequencies for those categories is called a *bar graph*.

Example 3.2 Bar Graph of Family Household Structure

Table 3.4 lists percentages of different types of family households in the United States in 1994. It is sufficient in such a table to report just the percentages and the total sample size, since each frequency equals the corresponding proportion multiplied by the

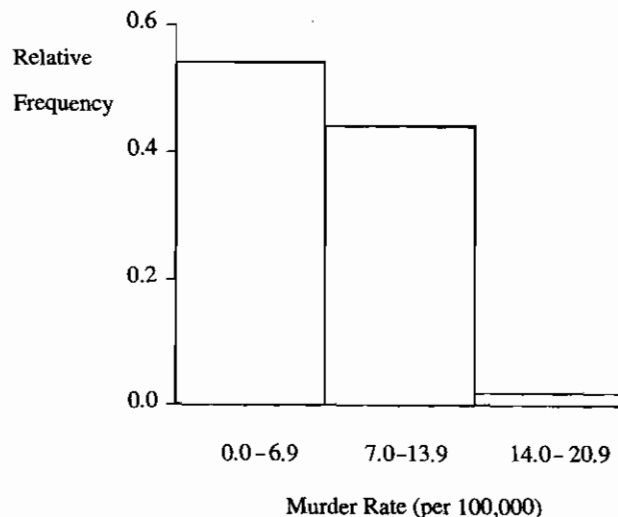


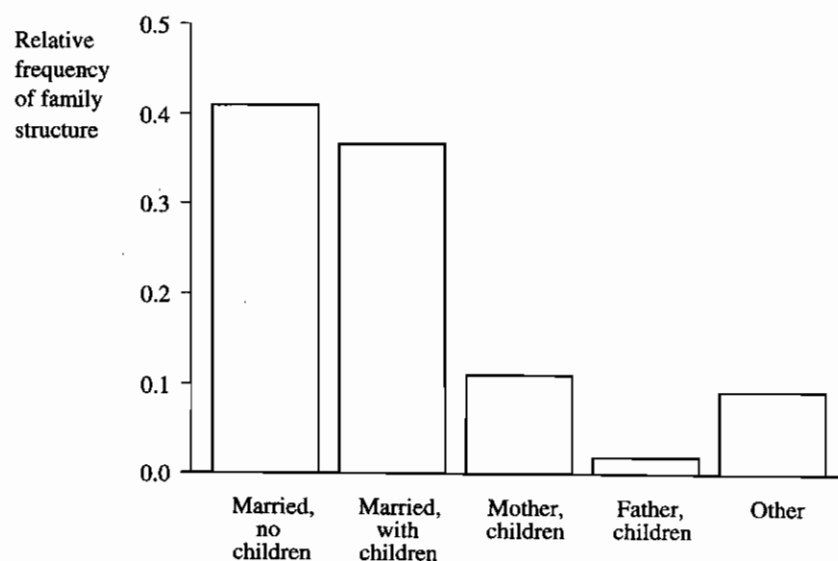
Figure 3.2 Relative Frequency Histogram for Murder Rates, Using Crude Intervals

TABLE 3.4 Family Structure, U.S. Families, 1994

Type of Family	Number (millions)	Percentage
Married couple with children	25.1	36.6
Married couple, no children	28.1	41.0
Single mother with children	7.6	11.1
Single father with children	1.3	1.9
Other families	6.4	9.3
Total	68.5	99.9

Source: U.S. Bureau of the Census, *Current Population Reports*.

total sample size. For instance, the frequency of single-mother families with children equals $.111(68.5) = 7.6$ million. Figure 3.3 presents the same data in a bar graph. Since family structure is a nominal variable, the order of the bars is not determined. By convention, they are usually ordered by frequency, except possibly for an "other" category, which is listed last. The order of presentation for an ordinal classification is the natural ordering of the levels of the variable. The bars in a bar graph, unlike in a histogram, are separated to emphasize that the variable is categorical rather than interval (quantitative). □

**Figure 3.3** Relative Frequency of Family Structure Types, U.S. Families, 1994

Stem and Leaf Plots

Figure 3.4 shows an alternative graphical representation of the murder rate data. This figure, called a *stem and leaf plot*, represents each observation by its leading digit(s) (the *stem*) and by its final digit (the *leaf*). In Figure 3.4, each stem is a number to the

left of the vertical bar and a leaf is a number to the right of it. For the murder rates, the stem is the whole part of a number, and the leaf is the fractional part. For instance, on the first line, the stem of 1 and the leaves of 6 and 7 represent the murder rates 1.6 and 1.7. On the second line, the stem of 2 has leaves of 0, 3, 9, representing the murder rates 2.0, 2.3, and 2.9.

Stem and leaf plots arrange the leaves in order on each line, from smallest to largest. Two-digit stems refer to double-digit numbers; for instance, the last line has a stem of 20 and a leaf of 3, representing the murder rate 20.3.

A stem and leaf plot conveys much of the same information as a histogram. Turned on its side, it has the same shape as the histogram. In fact, since one can recover the sample measurements from the stem and leaf plot, it displays information that is lost with a histogram. For instance, from Figure 3.4, the largest murder rate for a state was 20.3 and the smallest was 1.6. It is not possible to determine these exact values from the histograms in Figures 3.1 and 3.2.

Stem	Leaf									
1	6	7								
2	0	3	9							
3	0	1	4	4	4	6	8	9	9	9
4	4	6								
5	0	2	3	8						
6	0	3	4	6	8	9				
7	5									
8	0	3	4	6	9					
9	0	8								
10	2	2	3	4						
11	3	3	4	4	6	9				
12	7									
13	1	3	5							
14										
15										
16										
17										
18										
19										
20	3									

Figure 3.4 Stem and Leaf Plot for Murder Rate Data in Table 3.1

Stem and leaf plots are useful for quick portrayals of small data sets. As the sample size increases, you can accommodate the increase in leaves by splitting the stems. For instance, you might list each stem twice, putting leaves of 0 to 4 on one line and leaves of 5 to 9 on another. When a number has several digits, it is simplest for graphical portrayal to drop the last digit or two. For instance, for a stem and leaf plot of annual income in thousands of dollars, a value of \$27.1 thousand has a stem of 2 and a leaf of 7 and a value of \$106.4 thousand has a stem of 10 and leaf of 6.

Comparing Groups

Many studies compare different groups with respect to their distribution on some variable. Relative frequency distributions, histograms, and stem and leaf plots are useful for describing differences between the groups.

Example 3.3 Comparing Canadian and U.S. Murder Rates

Table 3.5 shows recent annual murder rates for the provinces of Canada. The rates are all less than 3.0, so they would all fall in the first category of Table 3.2 or the first bar of the histogram in Figure 3.1.

TABLE 3.5 Canadian Provinces and Their Murder Rates
(Number of Murders per 100,000 Population)

Alberta	2.7	British Columbia	2.6
Manitoba	2.9	New Brunswick	1.1
Newfoundland	1.2	Nova Scotia	1.3
Ontario	2.0	Prince Edward Island	0.7
Quebec	2.3	Saskatchewan	2.2

Source: *Canada Year Book*, 1992.

Stem and leaf plots can provide simple visual comparisons of two relatively small samples on a quantitative variable. For ease of comparison, the results are plotted “back to back”; each plot uses the same stem, with leaves for one sample to its left and leaves for the other sample to its right. To illustrate, Figure 3.5 shows back-to-back stem and leaf plots of the murder rate data for the United States and Canada. From this figure, it is clear that the murder rates tend to be much lower in Canada. □

Sample and Population Distributions

Frequency distributions and histograms for a variable apply both to a population and to samples from that population. The first type is called the *population distribution* of the variable, and the second type is called a *sample distribution*. In a sense, the sample distribution is a blurry photograph of the population distribution. As the sample size increases, the sample proportion in any interval gets closer to the true population proportion. Thus, the photograph gets clearer, and the sample distribution looks more like the population distribution.

When a variable is continuous, one can choose the intervals for a histogram as narrow as desired. Now, as the sample size increases indefinitely and the number of intervals simultaneously increases, with their width narrowing, the shape of the sample histogram gradually approaches a smooth curve. This text uses such curves to represent population distributions. Figure 3.6 shows two sample histograms, one based on

Canada	Stem	United States
	0	
	1	6 7
	2	0 3 9
9 7 6 3 2 1	3	0 1 4 4 4 6 8 9 9 9
	4	4 6
	5	0 2 3 8
	6	0 3 4 6 8 9
	7	5
	8	0 3 4 6 9
	9	0 8
	10	2 2 3 4
	11	3 3 4 4 6 9
	12	7
	13	1 3 5
	14	
	15	
	16	
	17	
	18	
	19	
	20	3

Figure 3.5 Back-to-Back Stem and Leaf Plots for Murder Rate Data from U.S. and Canada

a sample of size 100 and the second based on a sample of size 500, and also a smooth curve representing the population distribution. Even if a variable is discrete, a smooth curve often approximates well the population distribution, especially when the number of possible values of the variable is large.

One way to summarize a sample or population distribution is to describe its shape. A group for which the distribution is bell-shaped is fundamentally different from a group for which the distribution is U-shaped, for example. See Figure 3.7. In the U-shaped distribution, the highest points (representing the largest frequencies) are at

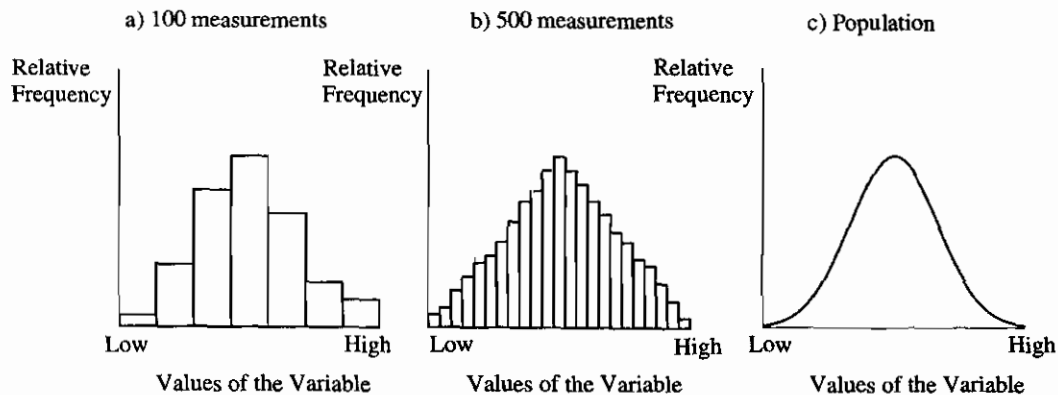


Figure 3.6 Histograms for a Continuous Variable

the lowest and highest scores, whereas in the bell-shaped distribution, the highest point is near the middle value of the variable. A U-shaped distribution indicates a polarization on the variable between two segments of the group, whereas a bell-shaped distribution indicates that most subjects tend to fall close to a central value.

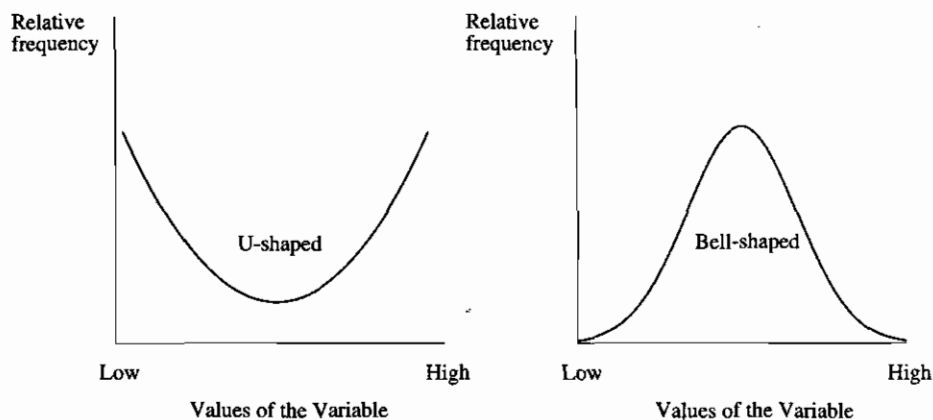


Figure 3.7 U-Shaped and Bell-Shaped Frequency Distributions

The bell-shaped and U-shaped distributions in Figure 3.7 are *symmetric*. Most distributions of variables studied in the social sciences are not exactly symmetric. Figure 3.8 illustrates. The parts of the curve for the lowest values and the highest values are called the *tails* of the distribution. A nonsymmetric distribution is said to be *skewed to the right* or *skewed to the left*, according to which tail is longer.

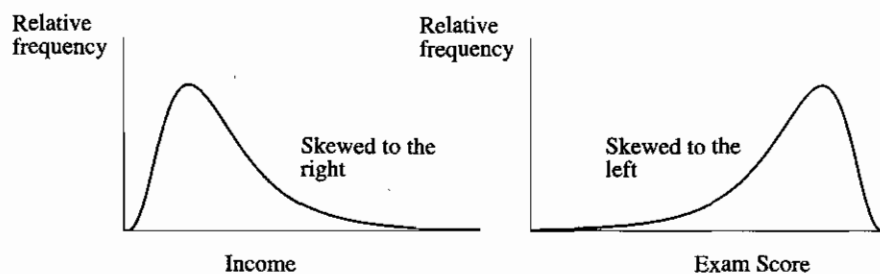


Figure 3.8 Skewed Frequency Distributions

A histogram for a sample approximates the corresponding population histogram. It is simpler to describe the difference between the two histograms, or the difference between sample distributions for two groups, using numerical descriptive methods. With these methods, one can make comparisons such as "On the average, the murder rate for U.S. states is 5.4 higher than the murder rate for Canadian provinces." We now turn our attention to ways of numerically describing data.

3.2 Measuring Central Tendency—The Mean

The next two sections present statistics that describe the center of a frequency distribution. The statistics show what a *typical* measurement in the sample is like. They are called *measures of central tendency*.

The Mean

The best known and most frequently used measure of central tendency is the *mean*, a description of the *average* response.

Mean

The *mean* is the sum of the measurements divided by the number of subjects.

The mean is often called the *average*. We illustrate the mean and its calculation with the following example.

Example 3.4 Female Economic Activity in Europe

Table 3.6 shows an index of female economic activity for the countries of Western and Eastern Europe in 1994 (data were not available for Germany). The number reported refers to female employment, as a percentage of male employment. In Austria, for instance, the number of females in the work force was 60% of the number of males in the work force.

The table lists six observations for Eastern Europe. For these data, the sum of the measurements equals $88 + 84 + 70 + 77 + 77 + 81 = 477$. The mean economic activity for these countries equals $477/6 = 79.5$. By comparison, you can check that the mean for the Western European countries equals $722/13 = 55.5$, considerably lower. (The values in the United States and Canada were 65 and 63.) \square

We now introduce notation for the mean. We use this notation in a formula for the mean and in formulas for other statistics that use the mean.

Notation for Observations and Sample Mean

The sample size is symbolized by n . For a variable denoted by Y , its observations are denoted by Y_1, Y_2, \dots, Y_n . The sample mean is denoted by \bar{Y} .

Throughout the text, n denotes the sample size. The n sample observations on a variable Y are denoted by Y_1 for the first observation, Y_2 the second, and so forth up to Y_n , the last observation made. For example, for female economic activity in Eastern Europe, $n = 6$, and the observations are $Y_1 = 88, Y_2 = 84, \dots, Y_n = Y_6 = 81$.

TABLE 3.6 Female Economic Activity in Europe; Female Employment as a Percentage of Male Employment

Western Europe		Eastern Europe	
Country	Activity	Country	Activity
Austria	60	Bulgaria	88
Belgium	47	Czech Republic	84
Denmark	77	Hungary	70
France	64	Poland	77
Ireland	41	Romania	77
Italy	44	Slovakia	81
Netherlands	42		
Norway	68		
Portugal	51		
Spain	31		
Sweden	77		
Switzerland	60		
United Kingdom	60		

Source: *Human Development Report 1995*, United Nations Development Programme.

The symbol \bar{Y} for the sample mean is read as “Y-bar.” Other symbols are also sometimes used for variables, such as X or Z . A bar over the symbol represents the sample mean of data for that variable. For instance, \bar{X} represents the sample mean for a variable denoted by X .

The definition of the sample mean implies that it equals

$$\bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$$

The symbol Σ (uppercase Greek letter sigma) represents the process of summing. For instance, ΣY_i represents the sum $Y_1 + Y_2 + \cdots + Y_n$. This symbol stands for the sum of the Y -values, where the index i represents a typical value in the range 1 to n . To illustrate, for the Eastern European data,

$$\sum Y_i = Y_1 + Y_2 + Y_3 + Y_4 + Y_5 + Y_6 = 477$$

The symbol is sometimes even further abbreviated as ΣY . Using this summation symbol, we have the shortened expression for the sample mean of n measurements,

$$\bar{Y} = \frac{\Sigma Y_i}{n}$$

Properties of the Mean

Before presenting additional examples, we consider some basic properties of the mean.

- The formula for the mean assumes numerical values Y_1, Y_2, \dots, Y_n for the observations. Because of this, the mean is appropriate only for quantitative data. It

is not sensible to compute the mean for observations on a nominal scale. For instance, for religion measured with categories such as (Protestant, Catholic, Jewish, Other), the mean religion does not make sense, even though these levels may sometimes be coded by numbers for convenience. Similarly, we cannot find the mean of observations on an ordinal rating such as excellent, good, fair, and poor, unless we assign numbers such as 4, 3, 2, 1 to the ordered levels, treating it as quantitative.

- The mean can be highly influenced by an observation that falls far from the rest of the data, called an *outlier*.

Example 3.5 Effect of Outlier on Mean Income

The owner of a small store reports that the mean annual income of employees in the business is \$37,900. Upon closer inspection, we find that the annual incomes of the seven employees are \$10,200, \$10,400, \$10,700, \$11,200, \$11,300, \$11,500, and \$200,000. The \$200,000 income is the salary of the owner's son, who happens to be an employee. The value \$200,000 is an outlier. The mean computed for the other six observations alone equals \$10,883, quite different from the mean of \$37,900 including the outlier. □

This example shows that the mean is not always representative of the measurements in the sample. This is fairly common with small samples when one or more measurements is much larger or much smaller than the others, such as in highly skewed distributions.

- The mean is pulled in the direction of the longer tail of a skewed distribution, relative to most of the data.

In Example 3.5, the large observation \$200,000 results in an extreme skewness to the right of the income distribution. This skewness pulls the mean above six of the seven measurements. In general, the more highly skewed the frequency distribution, the less representative the mean is of a typical observation.

- The mean is the point of balance on the number line when an equal weight occurs at each measurement point. For example, Figure 3.9 shows that if an equal weight is placed at each observation from Example 3.4, then the line balances by placing a fulcrum at the point 79.5. The mean is the *center of gravity* of the observations. This property implies that the sum of the distances to the mean from the observations above the mean equals the sum of the distances to the mean from the observations below the mean.
- Denote the sample means for two sets of data with sample sizes n_1 and n_2 by \bar{Y}_1 and \bar{Y}_2 . The overall sample mean for the combined set of $(n_1 + n_2)$ measurements is the *weighted average*

$$\bar{Y} = \frac{n_1 \bar{Y}_1 + n_2 \bar{Y}_2}{n_1 + n_2}$$

The numerator $n_1\bar{Y}_1 + n_2\bar{Y}_2$ is the total sum of all the measurements, since $n\bar{Y} = \sum Y$ for each set of measurements. The denominator is the total sample size.

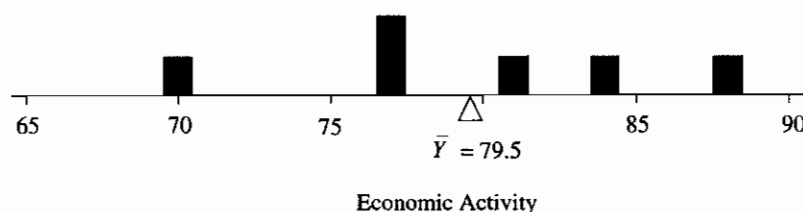


Figure 3.9 The Mean as the Center of Gravity

To illustrate, for the female economic activity data in Table 3.6, the Western European measurements have $n_1 = 13$ and $\bar{Y}_1 = 55.5$, and the Eastern European measurements have $n_2 = 6$ and $\bar{Y}_2 = 79.5$. The overall mean economic activity for the 19 nations equals

$$\bar{Y} = \frac{n_1\bar{Y}_1 + n_2\bar{Y}_2}{n_1 + n_2} = \frac{13(55.5) + 6(79.5)}{13 + 6} = \frac{(722 + 477)}{19} = \frac{1199}{19} = 63.1$$

The weighted average of 63.1 is closer to 55.5, the value for Western Europe, than to 79.5, the value for Eastern Europe, because most of the 19 observations for the overall sample come from Western Europe.

3.3 The Median and Other Measures of Central Tendency

Although the mean is a simple measure of central tendency, other measures are also informative and occasionally more appropriate than the mean.

The Median

The *median* splits the sample into two parts with equal numbers of subjects, when the subjects' observations are ordered from lowest to highest. It is a measure of central tendency that better describes a typical value when the sample distribution of measurements is highly skewed.

Median

The *median* is the measurement that falls in the middle of the ordered sample. When the sample size n is odd, a single measurement occurs in the middle. When the sample size is even, two middle measurements occur, and the median is the midpoint between the two.

To illustrate, the ordered income measurements for the seven employees in Example 3.5 are \$10,200, \$10,400, \$10,700, \$11,200, \$11,300, \$11,500, and \$200,000. The median is the middle measurement, \$11,200. This is a much more typical value for this sample than the sample mean of \$37,900. In this case, the median better describes central tendency than does the mean. In Table 3.6, the ordered economic activity values for the Eastern European nations are 70, 77, 77, 81, 84, and 88. Since $n = 6$ is even, the median is the midpoint between the two middle values, 77 and 81, which is $(77 + 81)/2 = 79.0$. This is close to the sample mean of 79.5, since this small data set has no outliers.

Since a stem and leaf plot arranges the observations in order, it is easy to determine the median using such a plot. For the data in Table 3.1 on murder rates, Figure 3.4 shows the stem and leaf plot. Since the sample size $n = 50$ is even, the median is the midpoint between the middle measurements, the 25th and 26th smallest. Counting down 25 leaves from the top of the plot, we find that 25th and 26th smallest values are 6.6 and 6.8. So, the median is $(6.6 + 6.8)/2 = 6.7$. The mean is $\bar{Y} = 7.3$, somewhat larger than the median. This is partly due to the outlier observation of 20.3 for Louisiana, which is considerably higher than the other observations. Turning Figure 3.4 on its side, we see that the murder rate values are skewed to the right.

The middle observation is the one having index $(n + 1)/2$. That is, the median is the value of the $(n + 1)/2$ nd measurement in the ordered sample. For instance, when $n = 7$, $(n + 1)/2 = (7 + 1)/2 = 4$, so the median is the fourth smallest, or equivalently fourth largest, observation. When n is even, $(n + 1)/2$ falls halfway between two numbers, and the median is the midpoint of the measurements with those indices. For instance, when $n = 50$, $(n + 1)/2 = 25.5$, so the median is the midpoint between the 25th and 26th smallest observations.

Example 3.6 Median for Grouped or Ordinal Data

Table 3.7 summarizes data on the highest degree completed for a sample of subjects taken recently by the U.S. Bureau of the Census. The measurement scale grouped the possible responses into an ordered set of categories. The sample size is $n = 177,618$. The median score is the $(n + 1)/2 = (177,618 + 1)/2 = 88,809.5$ th lowest. Now, 38,012 responses fall in the first category, $(38,012 + 65,291) = 103,303$ in the first two,

TABLE 3.7 Highest Degree Completed, for a Sample of Americans

Highest Degree	Frequency	Percentage
Not a high school graduate	38,012	21.4%
High school only	65,291	36.8%
Some college, no degree	33,191	18.7%
Associate's degree	7,570	4.3%
Bachelor's degree	22,845	12.9%
Master's degree	7,599	4.3%
Doctorate or professional	3,110	1.7%

and so forth. The 38,013rd to 103,303rd lowest scores fall in category 2, which therefore contains the 88,809.5th lowest, which is the median. The median response is "High school only." Equivalently, from the percentages in the last column of the table, 21.4% fall in the first category and $(21.4\% + 36.8\%) = 58.2\%$ fall in the first two, so the 50% point falls in the second category. \square

Properties of the Median

- The median, like the mean, is appropriate for interval data. Since it requires only ordered observations to compute it, it is also valid for ordinal data, as illustrated in the previous example. It is not appropriate for nominal data, since the observations cannot be ordered.
- For symmetric distributions, such as in Figure 3.7, the median and the mean are identical. To illustrate, the sample of measurements 4, 5, 7, 9, 10 is symmetric about 7; 5 and 9 fall equally distant from it in opposite directions, as do 4 and 10. Thus, 7 is both the median and the mean.
- For skewed distributions, the mean lies toward the direction of skew (the longer tail) relative to the median, as Figure 3.10 shows. Income distributions tend to be skewed to the right, though usually not as severely as in Example 3.5. The mean household income in the United States in 1993, for example, was about \$8000 higher than the median household income of \$31,000 (U.S. Bureau of the Census, *Current Population Reports*).

Length of prison sentences tend to be highly skewed to the right. For instance, in 1994, for 67 sentences for murder imposed using U.S. Sentencing Commission guidelines, the mean length was 251 months and the median was 160 months. The distribution of grades on an exam tends to be skewed to the left when some students perform considerably poorer than the others. In this case, the mean is less than the median. For instance, suppose that an exam scored on a scale of 0 to 100 has a median of 88 and a mean of 76. Then most students performed quite

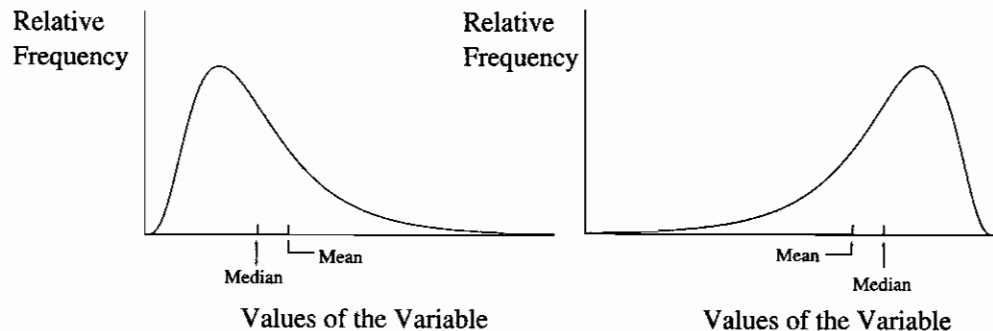


Figure 3.10 The Mean and the Median for Skewed Distributions

well (half being over 88), but apparently some scores were very much lower than the majority of students in order to bring the mean down to 76.

- The median is insensitive to the distances of the measurements from the middle, since it uses only the ordinal characteristics of the data. For example, the following four sets of measurements all have medians of 10:

Set 1:	8,	9,	10,	11,	12
Set 2:	8,	9,	10,	11,	100
Set 3:	0,	9,	10,	10,	10
Set 4:	8,	9,	10,	100,	100

- The median is unaffected by outliers. For instance, the incomes of the seven employees in Example 3.5 have a median of \$11,200 whether the largest observation is \$20,000, \$200,000, or \$2,000,000.

Example 3.7 Effect of Extreme Outlier for Murder Rate Data

Table 3.1 contains murder rates for the 50 states and has a mean of 7.3 and a median of 6.7. The data set does not include the District of Columbia (D.C.), which had a murder rate in 1993 of 78.5, nearly four times that of Louisiana. This is certainly an extreme outlier. If we include this observation in the data set, then $n = 51$. The median, the 26th largest observation, has 25 smaller and 25 larger observations. This is 6.8, so the median is barely affected by including this outlier. On the other hand, the mean changes from 7.3 to 8.7, being considerably affected by the outlier. The effect of an outlier tends to be even greater when the sample size is small, as Example 3.5 showed. □

Median Compared to Mean

The median has certain advantages, compared to the mean. For instance, the median is usually more appropriate when the distribution is highly skewed, as we have seen in Examples 3.5 and 3.7. The mean can be greatly affected by outliers, whereas the median is not.

The mean requires quantitative data, whereas the median also applies for ordinal scales (see Example 3.6). By contrast, using the mean for ordinal data requires assigning scores to the categories. In Table 3.7, if we assign scores 10, 12, 13, 14, 16, 18, 20 to the categories of highest degree, representing approximate number of years of education, we get a sample mean of 12.8.

The median also has disadvantages, compared to the mean. For discrete data that take on relatively few values, quite different patterns of data can give the same result. For instance, consider Table 3.8, from the General Social Survey of 1991. This survey, conducted annually by the National Opinion Research Center (NORC) at the University of Chicago, asks a sample of adult American subjects about a wide variety of issues. Table 3.8 summarizes the 1514 responses in 1991 to the question, "Within the past 12

TABLE 3.8 Number of People You Know Who Have Committed Suicide

Response	Frequency	Percentage
0	1344	88.8
1	133	8.8
2	25	1.7
3	11	.7
4	1	.1

months, how many people have you known personally that have committed suicide?" Only five distinct responses occur, and 88.8% of those are 0. Since $(n + 1)/2 = 757.5$, the median is the midpoint between the 757th and 758th smallest measurements. But those are both 0 responses, so the median response is 0.

To calculate the sample mean for Table 3.8, it is unnecessary to add the 1514 separate measurements to obtain $\sum Y_i$ for the numerator of \bar{Y} , since most values occurred several times. To sum the 1514 observations, we multiply each possible value by the frequency of its occurrence, and then add; that is,

$$\sum Y_i = 1344(0) + 133(1) + 25(2) + 11(3) + 1(4) = 220$$

The sample size is $n = 1344 + 133 + 25 + 11 + 1 = 1514$, so the sample mean response is

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{220}{1514} = .15$$

If the distribution of the 1514 observations among these categories were (758, 133, 25, 11, 587) (i.e., we shift 586 responses from 0 to 4), then the median would still be 0, but the mean would shift to 1.69. The mean uses the numerical values of all the observations, not just their ordering.

A more extreme form of this problem occurs for *binary data*. Such data can take only two values, such as (0, 1) or (low, high). The median equals the most common outcome, but gives no information about the relative number of observations at the two levels.

Quartiles and Other Percentiles

The median is a special case of a more general set of measures of location called *percentiles*.

Percentile

The ***p*th percentile** is a number such that *p*% of the scores fall below it and $(100 - p)\%$ fall above it.

Substituting $p = 50$ in this definition gives the 50th percentile. This is simply the median. That is, the median is larger than 50% of the measurements and smaller than the other $(100 - 50) = 50\%$. Two other commonly used percentiles are the *lower quartile* and *upper quartile*.

Lower and Upper Quartiles

The 25th percentile is called the **lower quartile**. The 75th percentile is called the **upper quartile**.

These refer to $p = 25$ and $p = 75$ in the percentile definition. One quarter of the data fall below the lower quartile, and one quarter fall above the upper quartile. The lower quartile is the median for the observations that fall below the median, that is, for the bottom half of the data. The upper quartile is the median for the observations that fall above the median, that is, for the upper half of the data. The quartiles together with the median split the distribution into four parts, each containing one-fourth of the measurements, as Figure 3.11 shows.

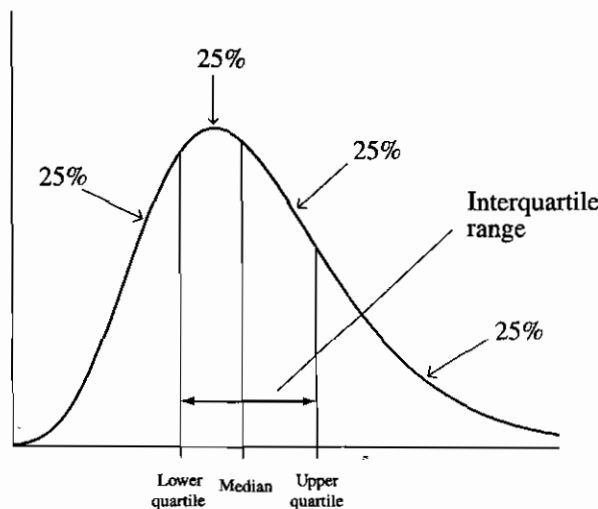


Figure 3.11 The Quartiles and Interquartile Range

We illustrate with the murder rates from Table 3.2. The sample size is $n = 50$, and the median equals 6.7. As with the median, the quartiles can easily be found from a stem and leaf plot, such as Figure 3.4. The lower quartile is the median for the 25 observations below the median, which is the 13th smallest observation, or 3.9. The upper quartile is the median for the 25 observations above the median, which is the 13th largest observation, or 10.3. This means that a quarter of the states had murder rates above 10.3. Similarly, a quarter of the states had murder rates below 3.9, between 3.9 and the median of 6.7, and between 6.7 and 10.3. The distance between the upper quartile and the median is $10.3 - 6.7 = 3.6$, which exceeds the distance $6.7 - 3.9 = 2.8$.

between the lower quartile and the median. This commonly happens when the distribution is skewed to the right.

We can summarize this information by reporting a *five-number summary*, consisting of the three quartiles and the minimum and maximum values. For instance, a popular software package reports these as follows:

100% Max	20.3
75% Q3	10.3
50% Med	6.7
25% Q1	3.9
0% Min	1.6

The five-number summary provides a simple-to-understand description of a data set.

The difference between the upper and lower quartiles is called the *interquartile range*. The middle half of the observations fall within that range. This measure describes variability of the data and is described further in Section 3.4. For the U.S. murder rates, the interquartile range equals $10.3 - 3.9 = 6.4$. The middle half of the murder rates fall within a range of 6.4.

Percentiles other than the quartiles and the median are usually reported only for fairly large data sets, and we omit rules for their calculation in this text.

The Mode

Another measure, the *mode*, describes a typical sample measurement in terms of the most common outcome.

Mode

The *mode* is the value that occurs most frequently.

In Table 3.8 on the suicide data, the mode is 0. The mode is more commonly used with categorical data or grouped frequency distributions than with ungrouped observations. The mode is then the category or interval with the highest frequency. In the data of Table 3.7 on the highest degree completed, for instance, the mode is "High school only," since the frequency for that category is higher than the frequency for any other rating.

The mode need not be near the center of the distribution. In fact, it may be the largest or the smallest value, if that is most common. Thus, it is somewhat inaccurate to call the mode a measure of central tendency. Many quantitative variables studied in the social sciences, though, have distributions in which the mode is near the center, such as in bell-shaped distributions and in slightly skewed distributions such as those in Figures 3.10 and 3.11.

Properties of the Mode

- The mode is appropriate for all types of data. For example, we might measure the modal religion (nominal level) in the United Kingdom, the modal rating (ordinal level) given a teacher, or the modal number of years of education (interval level) completed by Hispanic Americans.
- A frequency distribution is called **bimodal** if two distinct mounds occur in the distribution. Bimodal distributions often occur with attitudinal variables, when responses tend to be strongly in one direction or another, leading to polarization of the population. For instance, Figure 3.12 shows the relative frequency distribution of responses in the 1991 General Social Survey to the question, "Do you personally think it is wrong or not wrong for a woman to have an abortion if the family has a very low income and cannot afford any more children?" The relative frequencies in the two extreme categories are higher than those in the middle categories.
- The mean, median, and mode are identical for a unimodal, symmetric distribution, such as a bell-shaped distribution.

The mode is not as popular as the mean or median for describing central tendency of quantitative variables. It is useful when the most frequently occurring level of a variable is relevant, which is often true for categorical variables. The mean, median, quartiles, and mode are complementary measures. They describe different aspects of the data. In any particular example, some or all of their values may be useful.

Finally, these statistics are sometimes misused, as in Example 3.5. People who present statistical conclusions often choose the statistic giving the impression they wish to convey. Other statistics that might provide somewhat different interpretations are ignored. You should be on the lookout for misleading statistical analyses. For instance, be wary of the mean when you think that the distribution may be highly skewed.

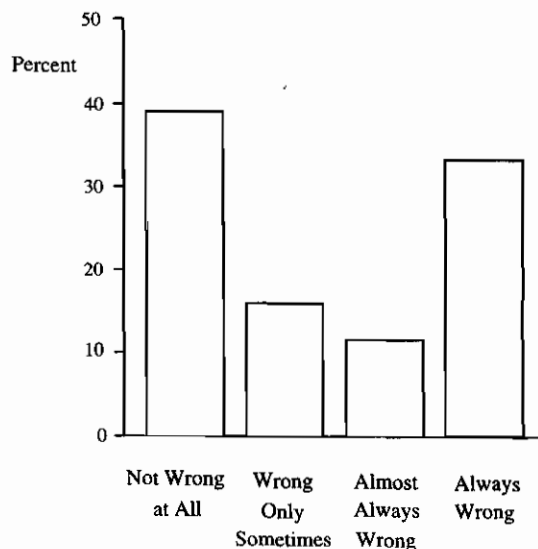


Figure 3.12 Bimodal Distribution for Opinion about Abortion

3.4 Measures of Variation

A measure of central location alone is not adequate for numerically describing a frequency distribution. It describes a typical value, but not the spread of the data about that value. The two distributions in Figure 3.13 illustrate. The citizens of nation A and the citizens of nation B have the same mean annual income (\$25,000). The distributions of those incomes differ fundamentally, however, nation B being much more homogeneous. An income of \$30,000 is extremely large for a resident of nation B, though not especially large for a resident of nation A. This section introduces statistics that describe the variability of a data set. These statistics are called *measures of variation*.

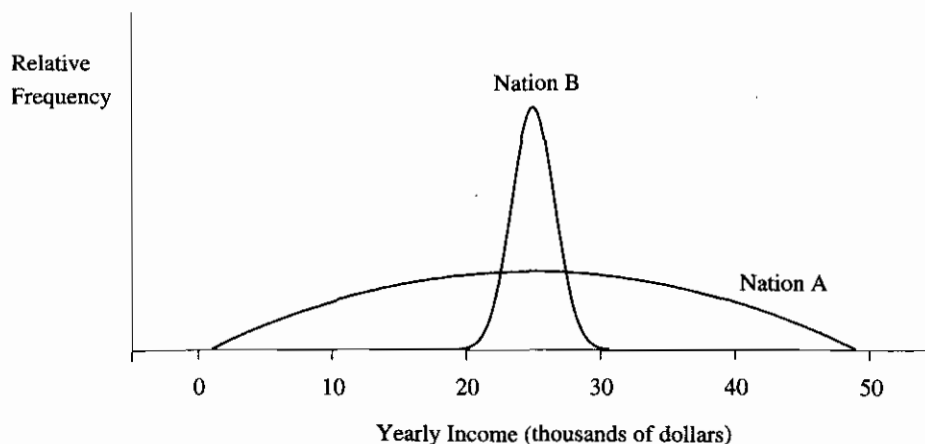


Figure 3.13 Distributions with the Same Mean but Different Variability

The Range

The difference between the largest and smallest observations in a sample is a simple measure of variation.

Range

The **range** is the difference between the largest and smallest observations.

For nation A, Figure 3.13 indicates that the range of income values is about $\$50,000 - \$0 = \$50,000$; for nation B, the range is about $\$30,000 - \$20,000 = \$10,000$. Nation A has greater variation of incomes than nation B.

The range is not, however, sensitive to other characteristics of data variability. The three distributions shown in Figure 3.14 all have the same mean (\$25,000) and range (\$50,000), yet they differ in variation about the center of the distribution. In terms of distances of measurements from the mean, nation A is the most disperse, and nation B is the least. The incomes in nation A tend to be farthest from the mean, and the incomes in nation B tend to be closest.

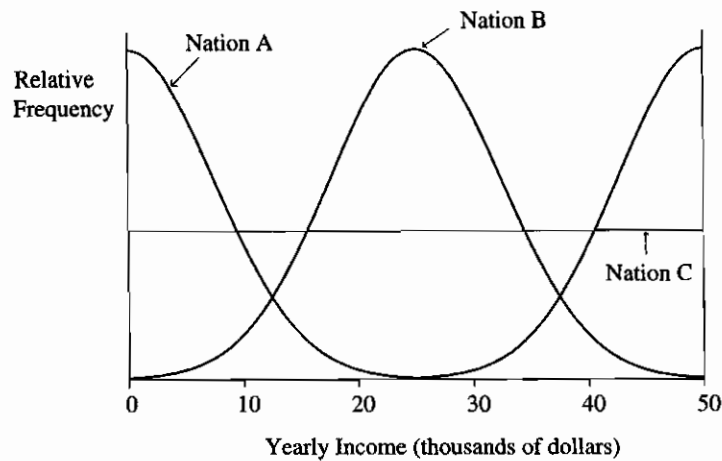


Figure 3.14 Distributions with the Same Mean and Range, but Different Variations About the Mean

Variance and Standard Deviation

Other measures of variation are based on the deviations of the data from a measure of central tendency, usually their mean.

Deviation

The **deviation** of the i th observation Y_i from the sample mean \bar{Y} is $(Y_i - \bar{Y})$, the difference between them.

Each observation has a deviation. The deviation is positive when the observation falls above the sample mean and negative when it falls below it. The interpretation of \bar{Y} as the center of gravity of the data implies that the sum of the positive deviations equals the negative of the sum of negative deviations; that is, the sum of all the deviations about the mean, $\Sigma(Y_i - \bar{Y})$, equals 0 for any sample. Because of this, summary measures of variation use either the absolute values or the squares of the deviations. The two measures we present incorporate the squares. The first measure is the **variance**.

Variance

The **variance** of n observations Y_1, \dots, Y_n is

$$s^2 = \frac{\Sigma (Y_i - \bar{Y})^2}{n - 1} = \frac{(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \dots + (Y_n - \bar{Y})^2}{n - 1}$$

The variance is approximately an average of the squared deviations. That is, it approximates the average of the squared distances from the mean. The units of measurement are the squares of those for the original data, since it uses squared deviations. This makes the variance difficult to interpret. The square root of the variance, called the *standard deviation*, is better for this purpose.

Standard Deviation

The *standard deviation* s is the positive square root of the variance:

$$s = \sqrt{\frac{\Sigma (Y_i - \bar{Y})^2}{n - 1}}$$

The expression $\Sigma (Y_i - \bar{Y})^2$ in the formulas for the variance and standard deviation is called a *sum of squares*. It represents squaring the deviations and then adding those squares. It is incorrect to first add the deviations and then square that sum; this gives a value of 0. The larger the deviations about the mean, the larger the sum of squares and the larger s and s^2 tend to be.

Example 3.8 Comparing Variability of Quiz Scores

Each of the following sets of quiz scores for two small samples of students has a mean of 5 and a range of 10:

Sample 1: 0, 4, 4, 5, 7, 10
Sample 2: 0, 0, 1, 9, 10, 10

By inspection, the scores in sample 1 show less variability about the mean than those in sample 2. Most scores in sample 1 are close to the mean of 5, whereas all the scores in sample 2 are quite far from 5.

For sample 1,

$$\Sigma (Y_i - \bar{Y})^2 = (0 - 5)^2 + (4 - 5)^2 + (4 - 5)^2 + (5 - 5)^2 + (7 - 5)^2 + (10 - 5)^2 = 56$$

so that the variance equals

$$s^2 = \frac{\Sigma (Y_i - \bar{Y})^2}{n - 1} = \frac{56}{6 - 1} = \frac{56}{5} = 11.2$$

Likewise, you can verify that for sample 2, $s^2 = 26.4$. The average squared distance from the mean is about 11 in sample 1 and 26 in sample 2. The standard deviation for sample 1 equals $s = \sqrt{11.2} = 3.3$, whereas for sample 2 it equals $s = \sqrt{26.4} = 5.1$.

Since $5.1 > 3.3$, the performances in sample 2 were more variable than those in sample 1, as expected. \square

Similarly, if s_A , s_B , and s_C denote the standard deviations of the three distributions in Figure 3.14, then $s_B < s_C < s_A$; that is, s_B is less than s_C , which is less than s_A .

Statistical software and many hand calculators can calculate the standard deviation for you. You should do the calculation yourself for a few small data sets to help you understand what this measure represents. The answer you get may differ slightly from the value reported by computer software, depending on how much you round off the mean before inserting it into the sum of squares part of the calculation.

Properties of the Standard Deviation

- $s \geq 0$.
- $s = 0$ only when all observations have the same value. For instance, if the ages in a sample of five students are 19, 19, 19, 19, 19, then the sample mean equals 19, each of the five deviations equals 0, and $s^2 = s = 0$. This is the minimum possible variation for a sample.
- The greater the variation about the mean, the larger is the value of s . Example 3.8 illustrated this property. For another example, we refer back to the U.S. and Canadian murder rates shown in Figure 3.5. The plot suggests that murder rates are much more variable in the U.S. In fact, the standard deviations are $s = 4.0$ for the United States and $s = .8$ for Canada.
- The reason for using $(n - 1)$, rather than n , in the denominator of s and s^2 is a technical one (discussed later in the text) concerning the use of these statistics to estimate population parameters. In the (rare) instances when we have data for the entire population, we replace $(n - 1)$ in these definitions by the actual population size. In this case, the standard deviation can be no larger than half the size of the range.
- Problem 3.64 at the end of this chapter presents two properties of standard deviations that refer to the effect of rescaling the data. Basically, if the data are rescaled, the standard deviation is also rescaled. For instance, if we double the scores, thus doubling the variation, then s doubles. If we change data on annual incomes from dollars (such as 34,000) to thousands of dollars (such as 34.0), the standard deviation also changes by a factor of 1000 (such as from 11,800 to 11.8).

Interpreting the Magnitude of s

Thus far, we have not discussed the magnitude of the standard deviation s other than in a comparative sense. A distribution with $s = 5.1$ has greater variation than one with $s = 3.3$, but how do we interpret *how large* $s = 5.1$ is? A very rough answer to this question is that s is a type of *average distance* of an observation from the mean. To illustrate, suppose the first exam in this course is graded on a scale of 0 to 100, and the sample mean for the students is 77. A value of $s = 0$ is very unlikely, since every

student must then score 77; a value of $s = 50$ seems implausibly large for a typical distance from the mean; values of s such as 8 or 11 or 14 seem much more realistic.

More precise ways to interpret s require further knowledge of the mathematical form of a frequency distribution. The following rule provides an approximate interpretation for many data sets.

Empirical Rule

If the histogram of the data is approximately bell-shaped, then

1. About 68% of the data fall between $\bar{Y} - s$ and $\bar{Y} + s$.
2. About 95% of the data fall between $\bar{Y} - 2s$ and $\bar{Y} + 2s$.
3. All or nearly all the data fall between $\bar{Y} - 3s$ and $\bar{Y} + 3s$.

The rule is called the Empirical Rule because many distributions encountered in practice (that is, *empirically*) are approximately bell-shaped. Figure 3.15 is a graphical portrayal of the rule.

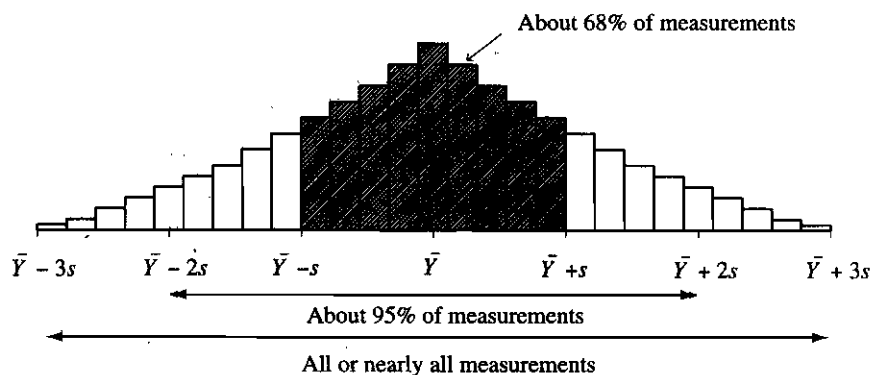


Figure 3.15 Empirical Rule: Interpretation of the Standard Deviation for a Bell-Shaped Distribution

Example 3.9 Describing the Distribution of SAT Scores

The distribution of scores on the verbal or math portion of the Scholastic Aptitude Test (SAT) is now scaled so it is approximately bell-shaped with mean of 500 and standard deviation of 100, as portrayed in Figure 3.16. By the Empirical Rule, about 68% of the scores fall between 400 and 600 on each test, since 400 and 600 are the numbers that are one standard deviation below and above the mean of 500. Similarly, about 95% of the scores fall between 300 and 700, the numbers that are two standard deviations from the mean. The remaining 5% fall either below 300 or above 700. The distribution is roughly symmetric about 500, so about 2.5% of the scores fall above 700 and about 2.5% fall below 300. \square

The percentages stated in the Empirical Rule are approximate and refer only to distributions that are approximately bell-shaped. In the bell-shaped case, for instance,

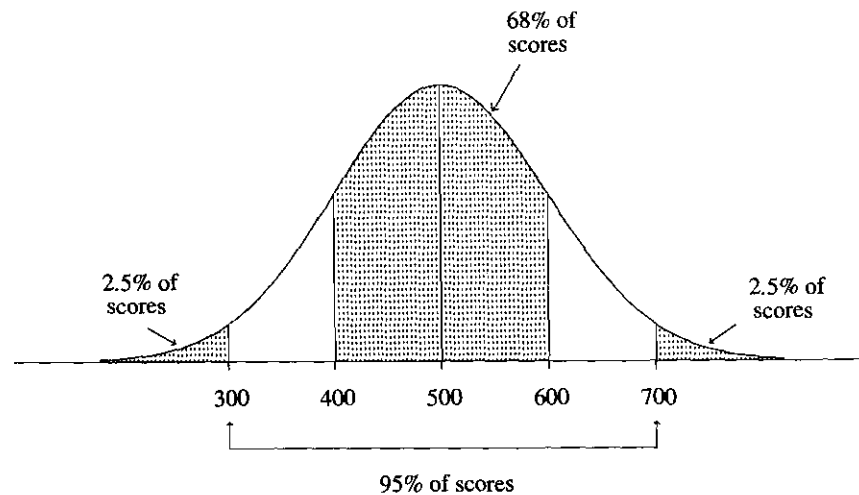


Figure 3.16 A Bell-Shaped Distribution of Test Scores with Mean 500 and Standard Deviation 100

the percentage of the distribution falling within two standard deviations of the mean is 95%, but this could change to as low as 75% or as high as 100% for other distributions. The Empirical Rule may not work well if the distribution is highly skewed or if it is highly discrete, with the variable taking relatively few values. The exact percentages depend on the form of the distribution, as Example 3.10 demonstrates.

Example 3.10 Familiarity with AIDS Victims

The 1993 General Social Survey asked “How many people have you known personally, either living or dead, who came down with AIDS?” Table 3.9 shows part of a computer printout for summarizing the 1598 responses on this variable. It indicates that 76% of the responses are 0, so that the lower quartile (Q_1), median, and upper quartile (Q_3) all equal 0.

The mean and standard deviation are $\bar{Y} = 0.47$ and $s = 1.09$. The values 0 and 1 both fall within one standard deviation of the mean. Now, 88.8% of the distribution falls at these two points, or within $\bar{Y} \pm s$. This is considerably larger than the 68% that the Empirical Rule predicts for bell-shaped distributions. The Empirical Rule does not apply to this frequency distribution, since it is not even approximately bell-shaped. Instead, it is highly skewed to the right, as you can check by sketching a histogram for Table 3.9. The smallest value in the distribution (0) is less than one standard deviation below the mean; the largest value in the distribution (8) is nearly seven standard deviations above the mean. \square

Whenever the smallest or largest observation is less than a standard deviation from the mean, this is evidence of severe skew. For instance, a recent exam one of us gave having scale from 0 to 100 had $\bar{Y} = 86$ and $s = 15$. Since the upper bound of 100 was less than one standard deviation above the mean, we surmised that the distribution of scores was highly skewed to the left.

TABLE 3.9 Frequency Distribution of the Number of People Known Personally With AIDS

AIDS	Frequency	Percent
0	1214	76.0
1	204	12.8
2	85	5.3
3	49	3.1
4	19	1.2
5	13	0.8
6	5	0.3
7	8	0.5
8	1	0.1

Analysis Variable : AIDS

N	1598	Quartiles		Range	8
Mean	0.473	100% Max	8	Q3-Q1	0
Std Dev	1.089	75% Q3	0	Mode	0
		50% Med	0		
		25% Q1	0		
		0% Min	0		

The standard deviation, like the mean, can be greatly affected by an outlier, particularly for small data sets. For instance, the murder rate data in Table 3.1 for the 50 states have $\bar{Y} = 7.33$ and $s = 3.98$. The distribution is somewhat irregular, but you can check that 68% of the states have murder rates within one standard deviation of the mean and 98% within two standard deviations. Now, suppose we include the murder rate for the District of Columbia in the data set, which equals 78.5. Then $\bar{Y} = 8.73$ and $s = 10.72$. The standard deviation more than doubles, and now 96.1% of the murder rates (all except D.C. and Louisiana) fall within one standard deviation of the mean.

Interquartile Range

The *interquartile range*, denoted by IQR, is another range-type statistic for describing variation. It is defined as the difference between the upper and lower quartiles. An advantage of the IQR over the ordinary range or the standard deviation is that it is not sensitive to extreme outlying observations.

To illustrate, we use the U.S. murder rate data shown in the stem and leaf plot in Figure 3.4. The rates range from 1.6 to 20.3, with a lower quartile of 3.9, a median of 6.7, and an upper quartile of 10.3. For these data, $IQR = 10.3 - 3.9 = 6.4$. When we add the observation of 78.5 for D.C. to the data set, the IQR changes only from 6.4 to 6.5. By contrast, the range changes from 18.7 to 76.9 and the standard deviation changes from 4.0 to 10.7.

Like the range and standard deviation, the IQR increases as the variability increases, and it is useful for comparing variation of different groups. To illustrate, we compare variability in U.S. and Canadian murder rates using the data shown in the back-to-back stem and leaf plots of Figure 3.5. The Canadian data has $IQR = 2.6 - 1.2 = 1.4$, showing much less variability than the IQR value of 6.4 for the U.S. data.

For bell-shaped distributions, the distance from the mean to either quartile is roughly $2/3$ rd of a standard deviation, and IQR is very roughly about $(4/3)s$. The insensitivity of the IQR to outliers has recently increased its popularity, though in practice the standard deviation is still much more common.

Box Plots

We conclude this section by presenting a graphical summary of both the central tendency and variation of a data set. This graphic, called a *box plot*, portrays the range and the quartiles of the data, and possibly some outliers.

The *box* contains the central 50% of the distribution, from the lower quartile to the upper quartile. The median is marked by a line drawn within the box. The lines extending from the box are called *whiskers*. These extend to the maximum and minimum values, unless there are outliers.

Figure 3.17 shows the box plot for the U.S. murder rates, in the format of box plots provided with SAS software (with the PLOT option in PROC UNIVARIATE). The upper whisker and upper half of the central box are longer than the lower ones. This indicates that the right tail of the distribution, which corresponds to the relatively large values, is longer than the left tail. The plot reflects the skewness to the right of the distribution of U.S. murder rates.

Box plots are particularly useful for comparing two distributions side by side. Figure 3.17 also shows the box plot for the Canadian murder rate data. These side-by-side

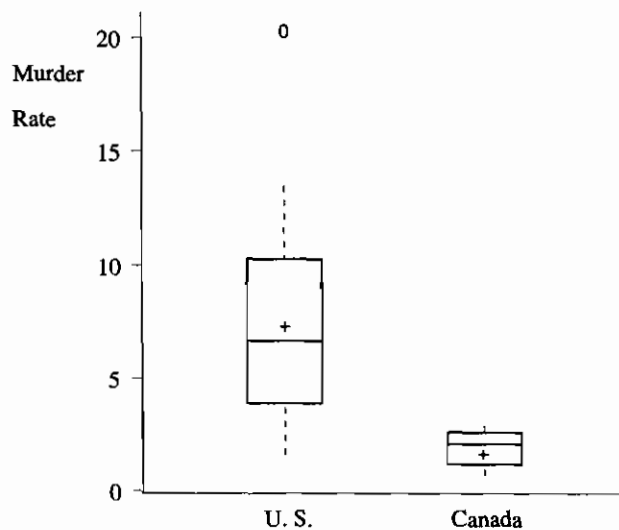


Figure 3.17 Box Plots for U.S. and Canadian Murder Rates

plots reveal that the murder rates in the U.S. tend to be much higher and have much greater variability.

Box plots identify outliers separately. To explain this, we now present a formal definition of an outlier.

Outlier

An observation is an **outlier** if it falls more than 1.5 IQR above the upper quartile or more than 1.5 IQR below the lower quartile.

In box plots, the whiskers extend to the smallest and largest observations only if those values are not outliers; that is, if they are no more than 1.5 IQR beyond the quartiles. Otherwise, the whiskers extend to the most extreme observations within 1.5 IQR, and the outliers are specially marked. For instance, SAS marks by an O (O for outlier) a value between 1.5 and 3.0 IQR from the box and by an asterisk (*) a value even farther away. Figure 3.17 shows one outlier with a very high murder rate, which is the murder rate of 20.3 for Louisiana. The distance of this observation from the upper quartile is $20.3 - 10.3 = 10.0$, which is greater than $1.5 \text{ IQR} = 1.5(6.4) = 9.6$.

The outliers are shown separately because they do not provide much information about the shape of the distribution, particularly for large data sets. SAS also plots the mean on the box plot, representing it by a + sign; these equal 7.3 for the United States and 1.9 for Canada. Comparing the mean to the median, which is the line within the box, helps show any skewness.

3.5 Sample Statistics and Population Parameters

Of the measures introduced in this chapter, the mean \bar{Y} and the standard deviation s are the most commonly reported. We shall refer to them frequently in the rest of the text. The formulas that define \bar{Y} and s refer to sample measurements. Since their values depend on the sample selected, they vary in value from sample to sample. In this sense, they are variables, sometimes called **random variables** to emphasize that their values vary according to the (random) sample selected. Their values are unknown before the sample is chosen. Once the sample is selected and they are computed, they become known sample statistics.

We shall regularly distinguish between sample statistics and the corresponding measures for the population. Section 1.2 introduced the term *parameter* for a summary measure of the population. A statistic describes a sample, while a parameter describes the population from which the sample was taken. In this text, lowercase Greek letters usually denote population parameters and Roman letters denote the sample statistics.

Notation for Parameters

Let μ (Greek mu) and σ (Greek sigma) denote the mean and standard deviation of a variable for the population.

We call μ and σ the *population mean* and *population standard deviation*. The population mean is the average of the population measurements. The population standard deviation describes the variation of the population measurements about the population mean.

Whereas the statistics \bar{Y} and s are variables, with values depending on the sample chosen, the parameters μ and σ are constants. This is because μ and σ refer to just one particular group of measurements, namely, the measurements for the entire population. Of course, the parameter values are usually unknown, which is the reason for sampling and calculating sample statistics as estimates of their values. Much of the rest of this text deals with ways of making inferences about unknown parameters (such as μ) using sample statistics (such as \bar{Y}). Before studying these inferential methods, though, we must introduce some basic ideas of *probability*, which serves as the foundation for the methods. Probability is the subject of Chapter 4.

3.6 Chapter Summary

This chapter introduced *descriptive statistics*—ways of *describing* a sample. Data sets in social science research are often large, and it is imperative to summarize the important characteristics of the information.

Overview of Tabular and Graphical Methods

- A *frequency distribution* of the sample measurements summarizes the counts of responses for a set of intervals of possible values. A *relative frequency* distribution reports this information in the form of percentages or proportions.
- A *histogram* provides a picture of this distribution. It is a bar graph of the relative frequencies. The histogram shows whether the distribution is approximately bell-shaped, U-shaped, skewed to the right (longer tail pointing to the right), or whatever.
- The *stem and leaf plot* is an alternative way of portraying the data, grouping together all observations having the same leading digits (stem), and showing also their final digit (leaf). Turned on its side, it shows the shape of the distribution, like a histogram, but it also presents the individual scores.
- The *box plot* portrays the quartiles, the extreme values, and any outliers. This plot and the stem and leaf plot are useful for back-to-back comparisons of two groups.

Stem and leaf plots and box plots, simple as they are, are relatively recent innovations in statistics (Tukey, 1977). See Cleveland (1985, 1993) and Tufté (1983, 1990) for even more recent and innovative ways to present data graphically.

Overview of Measures of Central Tendency

- **Measures of central tendency** describe the center of the collection of measurements, in terms of the “typical” score.
- The **mean** is the sum of the measurements divided by the sample size. It is the center of gravity of the data.
- The **median** divides the ordered data set into two parts of equal numbers of subjects, half scoring below and half above that point. It is less affected than the mean by outliers or extreme skew.
- The lower quarter of the observations fall below the **lower quartile**, and the upper quarter fall above the **upper quartile**. These are the 25th and 75th percentiles, and the median is the 50th percentile. The quartiles and median split the data into four equal parts.
- The **mode** is the most commonly occurring value. It is valid for any type of data, though usually used with categorical data.

Overview of Measures of Variation

- **Measures of variation** describe the variability of the measurements.
- The **range** is the difference between the largest and smallest measurements. The **interquartile range** is the difference between the upper and lower quartiles; it is less affected by extreme outliers.

TABLE 3.10 Measures of Central Tendency and Variation

	Measure	Definition	Interpretation
Central Tendency	Mean	$\bar{Y} = \Sigma Y_i / n$	Center of gravity
	Median	Middle measurement of ordered sample	50th percentile
	Mode	Most frequently occurring value	Most likely outcome, valid for all types of data
Variability	Variance	$s^2 = \Sigma (Y_i - \bar{Y})^2 / (n - 1)$	Greater with more variability, average squared distance from mean
	Standard deviation	$s = \sqrt{\Sigma (Y_i - \bar{Y})^2 / (n - 1)}$	Empirical Rule: If bell-shaped, 68%, 95% within s , $2s$ of \bar{Y}
	Range	Difference between largest and smallest measurement	Greater with more variability
	Interquartile range	Difference between upper and lower quartiles	Encompasses middle half of data

- The **variance** averages the squared deviations about the mean. Its square root, the **standard deviation**, is easier to interpret. The Empirical Rule states that for a sample with a bell-shaped distribution, about 68% of the measurements fall within one standard deviation of the mean and about 95% of the measurements fall within two standard deviations. Nearly all, if not all, the measurements fall within three standard deviations of the mean.

Table 3.10 summarizes the measures of central tendency and variation. A **statistic** summarizes a sample. A **parameter** summarizes a population. It is usually more relevant than the particular value of the statistic, which depends on the sample chosen. **Statistical inference** uses statistics to make predictions about parameters.

PROBLEMS

Practicing the Basics

1. According to the Bureau of the Census (*Current Population Reports*), in 1994 in the United States there were 23.6 million households with one person, 31.2 million with two persons, 16.9 million with three persons, 15.1 million with four persons, 6.7 million with five persons, 2.2 million with six persons, and 1.4 million with seven or more persons.
 - a) Construct a relative frequency distribution.
 - b) Construct a histogram. What is its shape?
 - c) Using a score of 8 for the final category, find the mean number of persons per household.
 - d) Report and interpret the median and mode of household size.
2. According to News America Syndicate, in 1986 the number of followers of the world's major religions were 835 million for Christianity, 420 million for Islam, 322 million for Hinduism, 300 million for Confucianism, 210 million for Buddhism, 79 million for Shinto, 50 million for Taoism, and 12 million for Judaism.
 - a) Construct a relative frequency distribution for these data.
 - b) Construct a bar graph for these data.
 - c) Can you calculate a mean, median, or mode for these data? If so, do so and interpret.
3. Refer to Table 3.1. Use software to construct a histogram for these data, using its default method of forming intervals. Describe the shape of the distribution, and construct the corresponding relative frequency distribution.
4. Table 3.11 shows the number (in millions) of the foreign-born population of the United States in 1990, by place of birth.
 - a) Construct a relative frequency distribution.
 - b) Plot the data in a bar graph.
 - c) Is "Place of birth" quantitative or qualitative? How, if at all, can you describe these data using numerical measures?
5. A researcher in an alcoholism treatment center, interested in summarizing the length of stay in the center for first-time patients, randomly selects ten records of individuals institutionalized within the previous two years. The lengths of stay in the center, in days, are as follows: 11, 6, 20, 9, 13, 4, 39, 13, 44, and 7.
 - a) Construct a stem and leaf plot.

TABLE 3.11

Place of Birth	Number
Europe	4.0
Former Soviet Union	.3
Asia	5.0
Canada	.7
Mexico	4.3
Caribbean	1.9
Central America	1.1
South America	1.0
Africa	.4
Oceania	.1
Total	18.8

Source: Statistical Abstract of the United States, 1994

- b) Find the mean, and interpret.
 - c) Find the median, and interpret.
 - d) Find the standard deviation, and interpret.
 - e) For a similar study 25 years ago at the same institution, lengths of stay for ten sampled individuals were 32, 18, 55, 17, 24, 31, 20, 40, 24, and 15 days. Compare results to those in the new study using (i) a back-to-back stem and leaf plot, (ii) the mean, (iii) the median, (iv) the standard deviation. Interpret any differences you find.
 - f) Actually, the new study also selected one other record. That patient is still institutionalized after 40 days. Thus, that patient's length of stay is at least 40 days, but the actual value is unknown. Can you calculate the mean or median for the complete sample of size 11 including this partial observation? Explain. (An observation such as this is said to be *censored*, meaning that the measured value is "cut short" of its true, unknown value.)
6. The 1994 General Social Survey asked respondents "How often do you read the newspaper?" The possible responses were (every day, a few times a week, once a week, less than once a week, never), and the counts in those categories were (969, 452, 261, 196, 76).
 - a) Identify the median response.
 - b) Identify the mode.
 - c) Consider the variable, Y = number of times reading the newspaper in a week, measured as described above. Can you calculate \bar{Y} ? Why? What would you need to do to approximate its value?
 7. Table 3.12 summarizes responses of 1250 subjects in the 1991 General Social Survey to the question, "About how often did you have sex during the last 12 months?"
 - a) Construct a bar graph, and interpret.
 - b) Report the median and the mode. Interpret.
 - c) Treat this scale in a quantitative manner by assigning the scores 0, .1, 1.0, 2.5, 4.3, 10.8, and 17 to the categories, representing approximate monthly frequency. Calculate the sample mean, and interpret.
 8. The 1991 General Social Survey asked respondents, "How many sex partners have you had in the last 12 months?" Table 3.13 shows results for 637 respondents.
 - a) Calculate and interpret the median and the mode.

TABLE 3.12

How Often Had Sex	Frequency
Not at all	292
Once or twice	99
About once a month	108
2 or 3 times a month	181
About once a week	233
2 or 3 times a week	265
More than 3 times a week	72

b) For the highest 11 values, we know only an interval within which the observation fell. To approximate these values, we could use midpoint scores. For instance, for interval 5–10, use $(5+10)/2 = 7.5$. We must choose an arbitrary score over 100 for the final interval. Using 120 for that observation, calculate the mean response. Compare to the median, and interpret.

c) Suppose the highest two observations were misrecorded, and the actual values were 4 for each. Recompute the mean and median, and use this example to describe potential effects of outliers on these measures.

TABLE 3.13

Number of Sex Partners	Frequency
0	146
1	418
2	39
3	15
4	8
5–10	7
11–20	2
21–100	1
More than 100	1

9. For 1992, the statewide number of abortions per 1000 women 15 to 44 years of age, for states in the Pacific region of the United States, were: Washington, 33; Oregon, 16; California, 304; Alaska, 2; and Hawaii, 11 (*Statistical Abstract of the United States, 1994*).
 - a) Calculate the mean.
 - b) Calculate the median. Why is it so different from the mean?
10. For 1993, Table 9.1 in Chapter 9 shows data on the statewide violent crime rate per 100,000 population. In this exercise, do not use the observation for D.C.
 - a) Using the intervals 0–100, 100–200, 200–300, and so forth, tally the frequencies and construct a frequency distribution.
 - b) Find the relative frequencies.
 - c) Sketch a histogram. How would you describe the shape of the distribution?
 - d) Drop the final digit of each crime rate. Then, construct a stem and leaf plot on this set of modified values. How does this plot compare to the histogram in (c)?

11. Refer to the preceding problem. Table 3.14 shows part of a computer printout for analyzing the data; the first column refers to the entire data set, and the second column deletes the observation for D.C.
- a) Report and interpret the mean and median of the first set of crime rates. Explain what their relative values suggest about the shape of the distribution.
- b) For each statistic reported, evaluate the effect of including the outlying observation for D.C.

TABLE 3.14

Variable=VIOLENT

N	51	N	50
Mean	612.84	Mean	566.66
Std Dev	441.10	Std Dev	295.88
Quartiles		Quartiles	
100% Max	2922	100% Max	1206
75% Q3	780	75% Q3	766
50% Med	515	50% Med	509.5
25% Q1	326	25% Q1	326
0% Min	82	0% Min	82
Range	2840	Range	1124
Q3-Q1	454	Q3-Q1	440
Mode	208	Mode	208

12. In 1992 in the United States, the median family income was \$38,909 for white families, \$21,161 for black families, and \$23,901 for Hispanic families (U.S. Bureau of the Census, *Current Population Reports*, P-60-184). In constant 1992 dollars, the median family incomes in 1975 were \$35,619 for white families, \$21,916 for black families, and \$23,844 for Hispanic families. Interpret the medians in 1992 and the changes in their values between 1975 and 1992.
13. Table 3.15 shows 1994 female economic activity for countries in South America.
- a) Construct a back-to-back stem and leaf plot of these values contrasted with those from Eastern Europe in Table 3.6. What is your interpretation?
- b) Compare the means for the two sets of nations, and interpret.
- c) Compare the medians, and interpret.
14. According to the U.S. Bureau of the Census, *Current Population Reports*, in 1994 the median household income was \$32,368 for whites and \$18,660 for blacks, whereas the mean household income was \$40,708 for whites and \$25,409 for blacks. Does this sug-

TABLE 3.15

Country	Activity	Country	Activity	Country	Activity
Argentina	38	Ecuador	24	Colombia	28
Uruguay	44	Paraguay	26	Peru	32
Chile	39	Bolivia	31	Brazil	38
Venezuela	39	Guyana	34		

Source: *Human Development Report, 1995*, United Nations Development Programme.

gest that the distribution of income is symmetric, or skewed to the right, or skewed to the left? Explain.

15. Refer to the previous exercise. The results refer to 57.9 million white households and 8.0 million black households.
 - a) Find the overall mean income.
 - b) If the mean income equals \$30,291 for 5.9 million Hispanic families, find the overall mean for the three groups combined.
16. For towns with population size 2500 to 4599 in the U.S. Northeast in 1994, the mean salary of chiefs of police was \$37,527, and the median was \$30,500 (*The Municipal Year Book 1995*. Washington, D.C.: International City/County Management Association, 1995). Does this suggest that the distribution of salary was skewed to the left, symmetric, or skewed to the right? Explain.
17. According to the National Association of Home Builders, the U.S. nationwide median selling price of homes sold in 1995 was \$118,000.
 - a) Would you expect the mean to be larger, smaller, or equal to \$118,000? Explain.
 - b) Which of the following is the most plausible value for the standard deviation: (i) -15,000, (ii) 1,000, (iii) 45,000, (iv) 1,000,000? Why?
18. The 1990 General Social Survey asked respondents, "During the past 12 months, how many people have you known personally that were victims of homicide." Table 3.16 shows a computer printout from analyzing responses for 1370 subjects.

TABLE 3.16

VICTIMS	Frequency	Percent								
0	1244	90.8								
1	81	5.9								
2	27	2.0								
3	11	0.8								
4	4	0.3								
5	2	0.1								
6	1	0.1								
N	Mean	Std Dev	100% Max	75% Q3	50% Med	25% Q1	0% Min			
1370	0.146	0.546	6	0	0	0	0			

- a) Report the relative frequency distribution.
- b) Sketch a histogram. Is the distribution bell-shaped, skewed to the right, or skewed to the left?
- c) Calculate the mean, median, and mode, and interpret their values.
- d) Report and interpret the standard deviation. Does the Empirical Rule apply to this distribution. Why or why not?
19. The Human Development Index (HDI) has three components: life expectancy at birth, educational attainment, and income. It ranges from 0 to 1, with higher values representing greater development. In 1992, the HDI ratings for eight Central American countries were .884 for Belize, .884 for Costa Rica, .579 for El Salvador, .591 for Guatemala, .578 for Honduras, .842 for Mexico, .611 for Nicaragua, and .856 for Panama.
 - a) Construct a stem and leaf plot. Drop the final digit, and split the values into two parts;

that is, have two lines for responses with first digit 8, putting entries with second digit 0 to 4 on one line and 5 to 9 on the second, have two lines for 7, two lines for 6, and two lines for 5. What is the shape of the distribution?

b) Calculate and interpret the mean, median, and range.

20. According to *Statistical Abstract of the United States, 1995*, average salary (in dollars) of secondary school classroom teachers in 1994 in the United States varied among states with a five-number summary of:

100% Max	51,700 (Connecticut)
75% Q3	38,500
50% Med	33,900
25% Q1	29,800
0% Min	25,300 (South Dakota)

a) Find and interpret the range and the interquartile range.

b) Construct a box plot.

c) Based on (b), predict the direction of skew for this distribution. Explain.

d) If the distribution, though skewed, is approximately bell-shaped, which of the following values would you expect for the standard deviation:

(i) 100, (ii) 1000, (iii) 6000, (iv) 15,000? Explain.

21. Consider the data in Table 3.8 on the number of people you know who have committed suicide. The mean equals .145, and the standard deviation equals .457. From the results reported in the table, what percentage of measurements fall within one standard deviation of the mean? Is the Empirical Rule appropriate for this distribution? Why or why not?
22. Why is the median sometimes preferred over the mean as a measure of central tendency? Give an example to illustrate your answer.
23. Why is the mean sometimes preferred over the median? Give an example to illustrate your answer.
24. Give an example of a variable for which the mode applies, but not the mean or median.
25. A group of high school students takes an exam. The mean score for the boys is 65, and the median is 75. Both the mean and the median score for the girls is 70. How can you explain the large difference between the two summary measures for the boys?
26. During the spring semester of 1995 at the University of Florida, computer usage of students having accounts on a mainframe computer at the university was summarized by a mean of 1921 and a standard deviation of 11,495 kilobytes of drive usage.
- a) Does the Empirical Rule apply to this distribution? Why?
- b) Would you expect this distribution to be symmetric, skewed to the right, or skewed to the left? Explain.
- c) What could cause the standard deviation to be so large compared to the mean? (Data supplied by Dr. Michael Conlon, University of Florida.)
27. Refer to Problem 3.26. The five-number summary of these data was minimum = 4, $Q1 = 256$, median = 530, $Q3 = 1105$, and maximum = 320,000. What does this suggest about the shape of the distribution? Why?
28. Residential electrical consumption in March 1994 in Gainesville, Florida, had a mean of 780 and a standard deviation of 506 kilowatt-hours (Kwh). The minimum usage was 3 Kwh and the maximum was 9390 Kwh. (Data supplied by N. Todd Kamhoot, Gainesville Regional Utilities.)
- a) What shape do you expect this distribution to have? Why?
- b) Do you expect this distribution to have any outliers? Explain.

29. Residential water consumption in March 1994 in Gainesville, Florida, had a mean of 7.1 and a standard deviation of 6.2 (thousand gallons). What shape do you expect this distribution to have? Why? (Data supplied by N. Todd Kamhoo, Gainesville Regional Utilities.)
30. For each of the following, sketch roughly what you expect a histogram to look like, and explain whether the mean or the median would be greater. Also sketch box plots for cases (a) and (c) that are consistent with the histograms.
 - a) The selling price of new homes in 1997
 - b) The number of children ever born per woman age 40 or over
 - c) The score on an easy exam (mean = 88, standard deviation = 10, maximum possible = 100)
 - d) The number of cars owned per family
 - e) Number of months in which subject drove a car last year
31. For each of the following variables, indicate whether you would expect its relative frequency histogram to be bell-shaped, U-shaped, skewed to the right, or skewed to the left. For parts (a), (b), and (g), sketch a box plot that would be plausible for that variable.
 - a) Exam score (scores fall between 0 and 100, with a mean of 90 and a standard deviation of 10)
 - b) IQ
 - c) Number of times arrested in past year
 - d) Time needed to complete difficult exam (maximum time is 1 hour)
 - e) Assessed value of home
 - f) Age at death
 - g) Weekly church contribution (median is \$10 and mean is \$17)
 - h) Number of years lived in present home (mode = 0 to 1 year)
 - i) Attitude toward legalization of abortion
32. Give examples of social science variables having a distribution that you would expect to be
 - a) Approximately symmetric
 - b) Skewed to the right
 - c) Skewed to the left
 - d) Bimodal
 - e) Skewed to the right, with a mode and median of 0 but a positive mean
33. A recent Roper organization survey asked: "How far have environmental protection laws and regulations gone?" For the possible responses not far enough, about right, and too far, the percentages of responses were 51%, 33%, and 16%.
 - a) Which response is the mode?
 - b) Can you compute a mean or a median for these data? If so, do so; if not, explain why not.
34. A company conducts a study of the number of miles traveled using public transportation by its employees during a typical day. A random sample of ten employees yields the following values (in miles):

0, 0, 4, 0, 0, 0, 10, 0, 6, 0

 - a) Calculate and interpret the mean, median, mode, range, variance, and standard deviation of these measurements.
 - b) The next person sampled lives in a different city and travels 90 miles a day on public transport. Recompute the mean, median, and standard deviation, and note the effect of this outlying observation.

35. To measure variation:
 - a) Why is the standard deviation s usually preferred over the range?
 - b) The IQR is sometimes preferred to s when there are some extreme outliers. Why?
36. In the mid-1980s, the General Social Survey asked respondents how many close friends they had. For a sample of size 1467, the mean was 7.4 and the standard deviation was 11.0. The distribution had a median of 5 and a mode of 4. Based on these statistics, what would you surmise about the shape of the distribution? Why?
37. In 1994 the General Social Survey asked, "On the average day, about how many hours do you personally watch television?" Of 1964 responses, the mode was 2, the median was 2, the mean was 2.8, and the standard deviation was 2.4. Based on these statistics, what would you surmise about the shape of the distribution?
38. For an exam given to a class, the students' scores ranged from 35 to 98, with a mean of 74. Which of the following is the most realistic value for the standard deviation? 1, 12, 60, -10? Why?
39. The sample mean for a data set equals 80. Which of the following is an impossible value for the standard deviation? 200, 0, -20.
40. According to a recent report from the U.S. National Center for Health Statistics, females with age between 25 and 34 years have a bell-shaped distribution on height, with mean of 65 inches and standard deviation of 3.5 inches.
 - a) Give an interval within which about 95% of the heights fall.
 - b) What is the height for a female who is three standard deviations below the mean in height. Would this be a rather unusual height? Why?
41. In a large northern city, monthly payments to people on welfare last year were observed to have approximately a bell shape with mean \$700 and standard deviation \$100. Give a range of values within which all or nearly all the payments fell.
42. For the WWW data on number of times a week reading a newspaper, referred to in Problem 1.7, Figure 3.18 shows a computer printout of the stem and leaf plot and the box plot.
 - a) From the box plot, identify the minimum, lower quartile, median, upper quartile, and maximum.
 - b) Identify these five numbers using the stem and leaf plot.
 - c) Do the data appear to contain any outliers? If so, identify.
 - d) Based on the box plot, indicate the approximate value of the mean. The standard deviation is one of the following values—.5, 3, 10, 20. Which do you think it is, and why?
43. Suppose the distribution of the prices of new homes built in the United States in 1996 was approximately bell-shaped, with a mean of \$120,000 and a standard deviation of \$40,000.
 - a) Describe the distribution using properties of the standard deviation.
 - b) If your new house was priced half a standard deviation above the mean in 1996, how much did it cost?
 - c) If the distribution is not actually bell-shaped, what shape would you expect it to have? Why?
44. In 1993, the five-number summary for the statewide percentage of people without health insurance had a minimum of 8.7% (Wisconsin), $Q1 = 11.9$, $Med = 13.4$, $Q3 = 17.8$, and maximum of 23.9% (Louisiana) (*Statistical Abstract of the United States, 1995*).
 - a) Construct a box plot for these data.
 - b) Do you think that the distribution is symmetric, skewed to the right, or skewed to the left? Why?

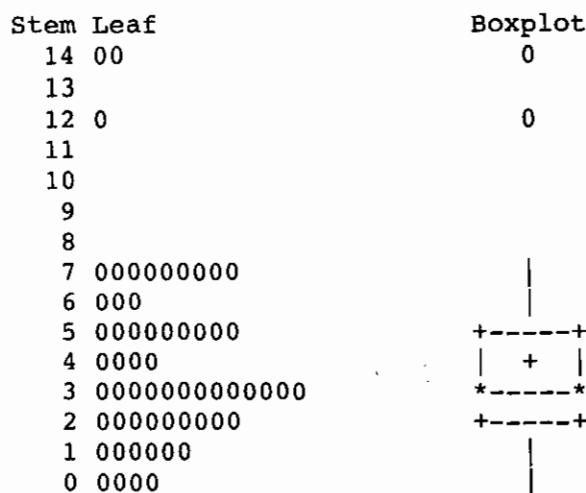


Figure 3.18

- c) Which of the following is the most plausible value for the standard deviation of this distribution: 0, 4, 13, 22? Why?
45. Refer to Problem 3.20. Construct a box plot for these data. Are there any apparent outliers?
46. The distribution of high school graduation rates in the United States in 1993 had a minimum value of 64.3 (Mississippi), lower quartile of 73.9, median of 76.75, upper quartile of 80.1, and maximum value of 86.6 (Alaska) (*Statistical Abstract of the United States, 1995*).
- Report and interpret the 50th percentile.
 - Report the range and the interquartile range.
 - Sketch a box plot. Are there any outliers?
 - Provide a guess for the standard deviation. Justify.
47. In your library, find the percentage of the vote that Bill Clinton received in each state in the 1996 presidential election.
- Prepare a stem and leaf plot. Are there any apparent outliers?
 - Construct a box plot. Are there any outliers?
 - Construct back-to-back stem and leaf plots or side-by-side box plots for Northeastern and West Coast states versus other states. Interpret.
48. Refer to Problem 3.10.
- Using the data set without D.C., find the quartiles and the interquartile range.
 - According to the definition of an outlier in terms of the IQR, are any of the observations outliers?
 - Construct a box plot for the distribution.
 - Repeat the analyses, including the D.C. observation, and compare results.
49. What is the difference between the descriptive measures symbolized by
- \bar{Y} and μ ?
 - s and σ ?

Concepts and Applications

50. For the WWW data file (Problem 1.7), use computer software to conduct graphical and numerical summaries for **a)** distance from home town, **b)** weekly hours of TV watching, **c)** weekly number of times reading a newspaper, and **d)** number of HIV-AIDS victims known. Describe the shapes of the distributions, and summarize your findings.
51. Refer to the data file you created in Problem 1.7. For variables chosen by your instructor, conduct descriptive statistical analyses. Prepare a report, interpreting and summarizing your findings.
52. Refer to the data in Table 9.1 on poverty rates. Using methods of this chapter, summarize these data. Prepare a report, graphically displaying the data and summarizing the central tendency and variation. In your report, discuss whether there any outliers, and if there are, analyze their influence on the results.
53. The number of therapeutic abortions in 1988 in Canada, per 100 live births, is shown in Table 3.17. Using methods of this chapter, present a descriptive statistical analysis of these data, interpreting your results.

TABLE 3.17

Alberta	15.0	British Columbia	25.5
Manitoba	16.6	New Brunswick	4.9
Newfoundland	6.3	Nova Scotia	14.2
Ontario	20.9	Prince Edward Island	3.5
Quebec	14.7	Saskatchewan	7.7
Yukon	22.6	Northwest Territories	17.9

Source: *Canada Year Book*, 1991.

54. Refer to Problem 3.19. Table 3.18 shows the HDI ratings for African countries. Using graphical and numerical methods of this chapter, summarize HDI for these countries, and compare to the distribution of HDI for Central American countries.
55. Obtain data on statewide murder rates from the latest edition of *Statistical Abstract of the United States*.
- a)** Analyze the data using the graphical and numerical methods of this chapter.
- b)** Use graphical and numerical methods to compare the murder rate distribution to the one for the data in Table 3.1.

TABLE 3.18

Libya	.768	Tunisia	.763	Algeria	.732	S. Africa	.705
Egypt	.613	Namibia	.611	Gabon	.579	Morocco	.554
Zimbabwe	.539	Congo	.538	Swaziland	.522	Cameroon	.503
Ghana	.482	Kenya	.481	Zambia	.425	Nigeria	.406
Zaire	.384	Sudan	.379	Ivory Coast	.369	Tanzania	.364
Cen. Af. Rep.	.361	Senegal	.340	Malawi	.330	Uganda	.329
Liberia	.325	Gambia	.299	Chad	.296	Angola	.291
Burundi	.286	Somalia	.246	Mozambique	.246	Ethiopia	.227
Mali	.222						

Source: *U.N. Human Development Report*, 1995.

56. During the strike of professional baseball players in 1994, two quite different numbers were reported for the central tendency of players' annual salaries. One was \$1.2 million and the other was \$500,000. One of these was the median and one was the mean. Which value do you think was the mean? Why?
57. In 1986, the U.S. Federal Reserve sampled about 4000 households to estimate overall net worth of a family. Excluding some outliers of extremely wealthy individuals, they reported the summaries \$44,000 and \$145,000. One of these was the mean, and one was the median. Which do you think was the median? Why?
58. According to a recent report from the U.S. National Center for Health Statistics, for males with age 25–34 years, 2% of their heights are 64 inches or less, 8% are 66 inches or less, 27% are 68 inches or less, 39% are 69 inches or less, 54% are 70 inches or less, 68% are 71 inches or less, 80% are 72 inches or less, 93% are 74 inches or less, and 98% are 76 inches or less. These are called *cumulative percentages*.
 - a) Find the median height.
 - b) Nearly all the heights fall between 60 and 80 inches, with less than 1% falling outside that range. If the heights are approximately bell-shaped, give a rough approximation for the standard deviation of the heights. Explain your reasoning.
59. Grade point averages of graduating seniors at the University of Rochester are approximately bell-shaped in distribution, ranging from 2.0 to 4.0 with a mean of about 3.0. Using the fact that all or nearly all measurements for this form of distribution occur within three standard deviations of their mean, give an approximation for the value of the standard deviation.

For the following two multiple-choice items, select the correct response(s).

60. In Canada in 1981, for the categories Catholic, Protestant, Eastern Orthodox, Jewish, None, Other for religious affiliation, the relative frequencies were 47.3%, 41.2%, 1.5%, 1.2%, 7.3%, 1.5% (*Canada Year Book, 1992*).
 - a) The median religion is Protestant.
 - b) The distribution is bimodal.
 - c) Only 2.7% of the subjects fall within one standard deviation of the mean.
 - d) The mode is Catholic.
 - e) The "Other" response is an outlier.
61. The 1991 General Social Survey asked whether having sex before marriage is always wrong, almost always wrong, wrong only sometimes, not wrong at all. The response counts in these four categories were 274, 98, 186, 435. This distribution is
 - a) Skewed to the right.
 - b) Approximately bell-shaped.
 - c) Bimodal.
 - d) Shape does not make sense, since the variable is nominal.
62. Ten families are randomly selected in Florida and another ten families are randomly selected in Alabama. Table 3.19 provides summary information on mean family income. The mean is higher in Alabama both in rural areas and in urban areas. Which state has the larger overall mean income? (The reason for this apparent paradox is that mean urban incomes are larger than mean rural incomes for both states and the Florida sample has a higher proportion of urban residents than the Alabama sample.)

TABLE 3.19

State	Rural		Urban	
Florida	\$26,000	($n = 3$)	\$39,000	($n = 7$)
Alabama	\$27,000	($n = 8$)	\$40,000	($n = 2$)

63. Refer to Problem 3.10. Explain why the mean of these 50 measurements is not necessarily the same as the violent crime rate for the entire U.S. population.
64. The mean and standard deviation of a sample may change if data are rescaled. For a sample with mean \bar{Y} , adding a constant c to each observation changes the mean to $\bar{Y} + c$, and the standard deviation s is unchanged. Multiplying each observation by c changes the mean to $c\bar{Y}$ and the standard deviation to $|c|s$.
- a) Scores on a difficult exam have a mean of 57 and a standard deviation of 20. The teacher boosts all the scores by 20 points before awarding grades. Report the mean and standard deviation of the boosted scores.
- b) Suppose that annual income has a mean of \$39,000 and a standard deviation of \$15,000. Values are converted to British pounds for presentation to a British audience. If one British pound equals \$1.50, report the mean and standard deviation in British currency.
65. The results of the study described in Problem 3.34 are to be reported in a French newspaper. The ten measurements are converted to kilometer units (1 mile = 1.6 kilometers). Report the mean and standard deviation of the converted measurements. (*Hint:* You do not need to convert the original scores; you just need to convert the original mean and standard deviation.)
66. ^{*1} The crude death rate is the number of deaths in a year, per size of the population, multiplied by 1000. According to the U.S. Bureau of the Census, in 1995 Mexico had a crude death rate of 4.6 (i.e., 4.6 deaths per 1000 population) while the United States had a crude death rate of 8.4. Explain how this overall death rate could be higher in the United States even if the United States had a lower death rate than Mexico for people of each specific age.
67. ^{*} The sample means for k sets of data with sample sizes n_1, n_2, \dots, n_k are $\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_k$. Show that the overall sample mean for the combined data set is

$$\bar{Y} = \frac{n_1 \bar{Y}_1 + n_2 \bar{Y}_2 + \dots + n_k \bar{Y}_k}{n_1 + n_2 + \dots + n_k}$$

Interpret \bar{Y} as a weighted average of $\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_k$.

68. ^{*} Show that $\Sigma(Y_i - \bar{Y})$ must equal 0 for any collection of measurements Y_1, Y_2, \dots, Y_n .
69. The Russian mathematician Tchebysheff proved that for any real number $k > 1$, the proportion of the measurements that fall more than k standard deviations from the mean can be no greater than $1/k^2$. Moreover, this holds for *any* distribution, not just bell-shaped ones.
- a) Find the upper bound for the proportion of measurements falling (i) more than two standard deviations from the mean, (ii) more than three standard deviations from the mean, (iii) more than ten standard deviations from the mean.
- b) Compare the upper bound for $k = 2$ to the approximate proportion falling more than

¹Problems marked with an asterisk are of slightly greater difficulty or else introduce new and optional material

two standard deviations from the mean in a bell-shaped distribution. Why is there a difference?

70. * The *least squares property* of the mean states that the data fall closer to \bar{Y} than to any other real number c , in the sense that the sum of squares of deviations of the data about their mean is smaller than the sum of squares of their deviations about c . That is,

$$\sum (Y_i - \bar{Y})^2 < \sum (Y_i - c)^2$$

If you have studied calculus, prove this property by treating $f(c) = \sum (Y_i - c)^2$ as a function of c and deriving the value of c that provides a minimum. (*Hint*: Take the derivative of $f(c)$ with respect to c and set it equal to zero.)

Bibliography

- Cleveland, W. S. (1985). *The Elements of Graphing Data*. Belmont, CA: Wadsworth.
 Cleveland, W. S. (1993). *Visualizing Data*. Summit, NJ: Hobart Press.
 Huff, D. (1954). *How to Lie with Statistics*. New York: W. W. Norton
 Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
 Tufte, E. R. (1990). *Envisioning Information*. Cheshire, CT: Graphics Press.
 Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA.: Addison-Wesley.