AN INTRODUCTION TO STATISTICS

WITH

DATA ANALYSIS

by

SHELLEY RASMUSSEN

Department of Mathematical Sciences
Olney 428T
University of Massachusetts/Lowell
Lowell, MA 01854

Originally published by Brooks/Cole Publishing Company,
Division of Wadsworth, Inc.

ISBN 0-534-13578-1

Shelley_Rasmussen@uml.edu

Minitab is a statistical package, a computer program that performs many statistical procedures. The versions of Minitab now available for use on personal computers are menu-driven and much easier to use than the main-frame version originally discussed in this text. Those sections are not included in this online edition of the text. At this time, the most recent version is Minitab 15, available at very reasonable prices for purchase or rental from:

www.e-academy.com/minitab

System Requirements

Processor:	PC with a 1 GHz 32- or 64-bit processor
Memory:	512 MB or more of available RAM
Disk Space:	125 MB free space available
Operating System:	Microsoft Windows 2000, XP, or Vista.
Display:	A display capable of 1024 X 768 or higher resolution
Software:	Adobe Acrobat Reader 5.0 or higher for Meet Minitab

CHAPTER 3

Studying One Variable at a Time: Descriptive Statistics

IN THIS CHAPTER

Measures of central tendency or location
Mean, trimmed mean, weighted mean
Median
Measures of spread or variation
Range, interquartile range
Standard deviation, for a population and for a sample

How can we describe central tendency of life expectancies in a group of countries? How much variation is there among those life expectancies? We might address these questions in terms of *descriptive statistics,* numbers used to describe or summarize a set of values.

> A **descriptive statistic** is a number used to describe or summarize a set of values.

One class of descriptive statistics that we have already encountered consists of the quantiles, numbers that divide a set of ordered values into equal or approximately equal sized groups. We use quantiles to construct box plots and box graphs.

Now we are going to discuss some descriptive statistics that measure location or central tendency and some that measure spread or variation, for a single variable. The median is a descriptive statistic that we have seen already in our discussion of quantiles. The median, which divides an ordered list of values in half, is a measure of location or central tendency. The range from the minimum to the maximum is a summary measure of variation or spread in a set of values. We begin our discussion with measures of central tendency and then consider measures of variation.

3-1

Measures of Central Tendency

A number describing the location of a set of values is a *measure of central tendency* or *measure of location.* The mean and the median are the most common measures of central tendency. We have already defined the *median* as a number that divides an ordered list of values in half. Now we will look at the mean, then consider variations of the mean and compare them with the median as measures of location of a set of values.

> A **measure of central tendency** or **location** is a descriptive statistic that summarizes central tendency or location of a distribution of values.

The *mean* or *average* of a set of values is the arithmetic average of the values. When calculated from a sample, it is called the *sample mean* or *sample average.* To calculate the mean, add up the values and then divide by the number of values:

$$\textbf{Mean or average} = \frac{\text{Sum of the values}}{\text{Number of values}}$$

In statistics we commonly let n denote the number of data values. We let x_1 denote the first value, x_2 the second value, and so on, to x_n, the nth value. The capital Greek letter sigma, Σ, denotes summation in mathematics. Putting these symbols together, we can write the calculation formula for a mean as

$$\text{Mean of } x_1 \text{ through } x_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

These symbols tell us to add up the values and then divide by the number of values. When calculating an average of sample values, we often denote the resulting sample mean by $\bar{x}$.

The letter representing the values in the formula for the mean is arbitrary. We could let y_1, for example, denote the first value, y_2 denote the second, and so on. (If calculated for a sample, we would then let $\bar{y}$ denote the sample mean.) We can write the calculation formula for the mean of y_1 through y_n as

$$\text{Mean of } y_1 \text{ through } y_n = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Let's calculate the mean life expectancy for several groups of countries.

Life expectancies are shown in Table 3-1 for four high-income oil exporters. We calculate the mean life expectancy for these four countries by adding the four life expectancies and then dividing by 4:

$$\frac{72 + 60 + 62 + 70}{4} = 66 \text{ years}$$

The average is exactly 66. The mean life expectancy for these four high-income oil exporting nations is 66 years. Note that because of the symmetry of these four values of life expectancy, the median (the average of the middle values 62 and 70) also equals 66 years.

Recall that we listed the life expectancies for the industrial market countries in Section 2-5. The average of these 19 life expectancy values is 76.05 years, to two decimal places. If we round to the nearest year, we say the mean life expectancy for the industrial market nations is about 76 years. Because the distribution of these 19 life expectancies is fairly symmetrical, the mean is very close to the median, which we found to be 76 years in Section 2-5.

Thirty-five of the 37 low-income countries have nonmissing values for life expectancy, listed in Section 2-5. The average of these 35 life expectancies is 50.2, or about 50 years. Recall that the median life expectancy for these low-income countries is 49 years: About half the values are greater than or equal to 49 and about half are less than or equal to 49. The mean is larger than the median because the distribution of these 35 life expectancies is somewhat

TABLE 3-1 Life expectancy at birth in 1985 for four high-income oil exporting nations.

Country	1985 life expectancy at birth (years)
Kuwait	72
Libya	60
Saudi Arabia	62
United Arab Emirates	70

positively skewed; three low-income countries have relatively long life expectancies compared with the others (see Figure 3-3). These three values contribute to making the mean larger than the median.

We can interpret the mean or arithmetic average as a balance point or center of gravity. Figure 3-1 is a dot plot of life expectancies for the four high-income oil exporters. Each dot represents a single country and is positioned along the axis at the point corresponding to the country's life expectancy. Think of the axis as a plank of wood for a seesaw and the dots as children of identical size and weight sitting perfectly still on the seesaw. Where should the fulcrum be placed so that the seesaw will remain balanced in the horizontal position? It should be placed at the center of gravity—the mean or arithmetic average of the positions along the seesaw. For the four values plotted in Figure 3-1, the mean of 66 years is the arithmetic average of the positions along the axis. It is the point where we should place the fulcrum to balance the seesaw.

The four points in Figure 3-1 are symmetrically placed about the mean. The mean of 66 years is a good measure of the center of these values. However, 66 is not really close to any of the four values. As we can see from this example, a measure of the center of a set of values need not be particularly close to any of the values.

A dot plot of life expectancies for the 19 industrial market countries is shown in Figure 3-2. The arithmetic mean or balance point is 76.05 years.

Life expectancies for 35 low-income countries are plotted in Figure 3-3. The balance point is at the arithmetic average of 50.2 years.

When we average the life expectancies for a group of countries to calculate the mean, we weight each country equally. This is illustrated in Figures 3-1, 3-2, and 3-3, where dots represent countries and each dot is the same size. Does it make sense to treat countries equally in describing the center of the life expectancies, when some countries have more people than others?

If we consider a country as a single entity with its associated life expectancy, then the mean makes sense. However, if we want to describe life expectancies for the people in a group of countries, then we do not want a simple average across countries. We might instead weight each country's life expectancy by its population size in calculating an average. Then we have a **weighted mean.**

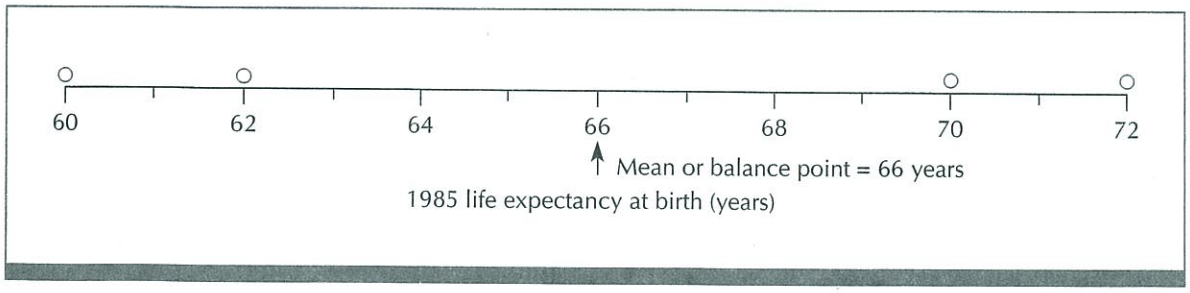

FIGURE 3-1 Dot plot of 1985 life expectancy at birth for four high-income oil exporting nations.

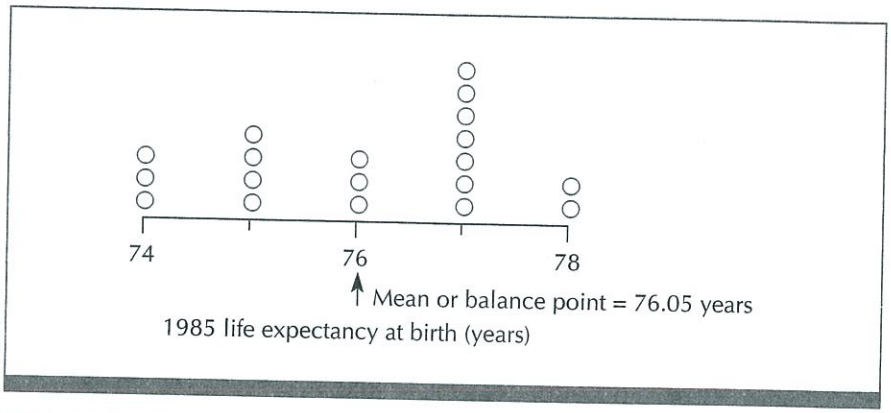

FIGURE 3-2 Dot plot of 1985 life expectancy at birth for 19 industrial market countries.

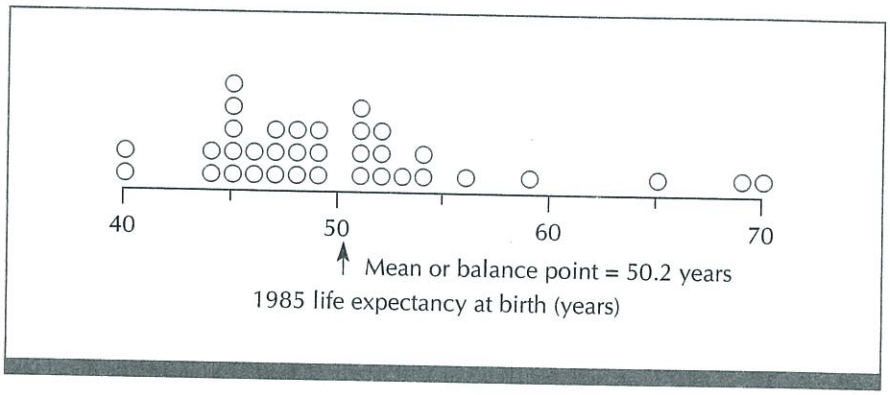

FIGURE 3-3 Dot plot of 1985 life expectancy at birth for 35 low-income countries. Two countries are excluded because of missing values.

To calculate a **weighted mean** or **weighted average,** multiply each value by an appropriate weight, add these products, and then divide by the sum of the weights:

$$\text{Weighted mean} = \frac{\text{Sum of the weighted data values}}{\text{Sum of the weights}}$$

Suppose we let x_1 denote the first data value, x_2 the second value, and so on, through x_n. Let w_1, w_2, through w_n denote a set of weights. Then we can write the corresponding weighted mean as

$$\text{Weighted mean} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{j=1}^{n} w_j}$$

which is the same as the definition above. (Note that this formula for the weighted mean uses different subscript letters in the numerator and denomi-

TABLE 3-2 Population in mid-1985 and life expectancy at birth in 1985 for four high-income oil exporters.

Country	Population (millions) in mid-1985	1985 life expectancy at birth (years)	Product of population and life expectancy
Kuwait	1.7	72	122.4
Libya	3.8	60	228.0
Saudi Arabia	11.5	62	713.0
United Arab Emirates	1.4	70	98.0
Sum or Total	18.4	264	1,161.4

Mean life expectancy: $\frac{264}{4} = 66$ years

Weighted mean life expectancy, life expectancy weighted by population size: $\frac{1{,}161.4}{18.4} = 63.12$, or 63 years

nator. This is to remind us that these sums are calculated separately, and then we find the quotient.)

Table 3-2 shows population size and life expectancy for the four high-income oil exporters. Let's calculate a weighted average life expectancy, with weights equal to population sizes. We multiply the life expectancy for each country by the population size for that country. The resulting products are shown in the last column of Table 3-2. We add these products, getting a sum of 1,161.4. We also add the weights. The sum of the weights is 18.4, the total population (in millions) for the four high-income oil exporters. We then divide 1,161.4 by 18.4 to get a weighted average life expectancy of about 63 years.

The weighted average life expectancy is 3 years less than the unweighted average. Examining Table 3-2, we see why. The two countries with the largest populations, Saudi Arabia and Libya, have the shorter life expectancies. Kuwait and United Arab Emirates have longer life expectancies, but relatively small population sizes. The two countries with the larger populations (and shorter life expectancies) contribute more to the weighted mean than to the unweighted mean. The weighted mean life expectancy of 63 years is a measure of the middle of life expectancies for the combined populations of these four high-income oil exporting nations. Exercises 3-2 and 3-3 ask you to find weighted mean life expectancy for the industrial market and low-income countries, respectively.

Trimmed Means as Measures of Central Tendency

The problem with the mean as a measure of central tendency is that it can be greatly influenced by a few extreme values. One very large value can make the mean much larger than it would be if that value were excluded (see Exercise 3-1), and similarly for an extremely small value. In such cases, the mean or average may not be a good measure of center.

TABLE 3-3 Population in mid-1985 for 35 low-income countries (values missing for Afghanistan and Kampuchea).

Country	Population (millions) in mid-1985	Country	Population (millions) in mid-1985
Bhutan	1.2	Madagascar	10.2
Central African Republic	2.6	Ghana	12.7
Togo	3.0	Mozambique	13.8
Lao PDR	3.6	Uganda	14.7
Sierra Leone	3.7	Sri Lanka	15.8
Benin	4.0	Nepal	16.5
Burundi	4.7	Kenya	20.4
Chad	5.0	Sudan	21.9
Somalia	5.4	Tanzania	22.2
Haiti	5.9	Zaire	30.6
Rwanda	6.0	Burma	36.9
Guinea	6.2	Ethiopia	42.3
Niger	6.4	Viet Nam	61.7
Senegal	6.6	Pakistan	96.2
Zambia	6.7	Bangladesh	100.6
Malawi	7.0	India	765.1
Mali	7.5	China	1,040.3
Burkina Faso	7.9		

Measures of central tendency based on 35 nonmissing values (in millions of people):
Mean: 69.0
5% *trimmed mean*: 19.6
15% *trimmed mean*: 13.5
Median: 7.9

Consider, for example, the population sizes shown in Table 3-3 for 35 low-income countries. The mean is 69.0 million people. But only 4 of the 35 countries have populations greater than 69.0 million; most have populations much smaller than 69.0 million. The mean is so large mainly because two countries have very large populations: China (1,040.3 million) and India (765.1 million). The mean does not provide a very satisfactory measure of the center of these 35 population sizes.

One way around this problem is to exclude very large and very small values before calculating a mean. The resulting measure of central tendency is called a trimmed mean. To calculate a 5% *trimmed mean,* we exclude the largest 5% and the smallest 5% of the values and calculate the mean of the remaining values. For a 15% *trimmed mean,* we calculate the mean after excluding the largest 15% and the smallest 15% of the values. In general, for an x% *trimmed mean* we exclude the smallest x% and the largest x% of the values and calculate the mean of the remaining $(100 - 2x)$% of the values. Typical values for the percent x excluded from each end are integers from 1 to 15. It is easiest to find a trimmed mean from an ordered list of values (such as we might have in a stem-and-leaf plot).

A **trimmed mean** of a set of values is a mean with a specified percentage of the largest and smallest values excluded from the calculation. An ***x*% trimmed mean** is a mean calculated after the largest x% and the smallest x% of the values have been excluded.

Let's find the 5% trimmed mean and the 15% trimmed mean for the 35 population sizes in Table 3-3. Five percent of 35 is 1.75, or about 2. For the 5% trimmed mean we exclude the two largest population sizes (for China and India) and the two smallest population sizes (for Bhutan and Central African Republic). The mean of the remaining 31 values is 19.6 million, the 5% trimmed mean. Eleven of the 35 countries have populations larger than 19.6 million, and 24 have populations smaller than 19.6 million.

Fifteen percent of 35 is 5.25, or about 5. For the 15% trimmed mean we exclude the five largest population sizes (for China, India, Bangladesh, Pakistan, and Viet Nam) and the five smallest population sizes (for Bhutan, Central African Republic, Togo, Lao PDR, and Sierra Leone). The mean of the remaining 25 population sizes is 13.5 million, the 15% trimmed mean. Fifteen of the 35 countries have populations greater than 13.5 million, and 20 have populations smaller than 13.5 million.

Only 4 of the 35 countries in Table 3-3 have population sizes greater than the mean value of 69.0 million people. Eleven countries have populations greater than the 5% trimmed mean of 19.6 million. Fifteen countries have populations greater than the 15% trimmed mean of 13.5 million. The 5% trimmed mean seems closer to the center than the mean. The 15% trimmed mean seems closer to the center than the 5% trimmed mean. Following this line of thought, we might say that the center of the population sizes is a number that divides the set of values in half—that is, the *median.*

The median population size for these low-income countries is 7.9 million, the middle of the 35 ordered values in Table 3-3. Eighteen of the 35 population sizes are greater than or equal to 7.9 million and 18 are less than or equal to 7.9 million. As a measure of central tendency, the median has two satisfying characteristics. It really is the center of a set of values in the sense that it divides the data set in half. Also, the median is not influenced by extremely small or large values. We can think of the median as the most extreme instance of a trimmed mean. If the number of values is odd, we trim all but the middle value, which is the median. If the number of values is even, we trim all but the middle two values and average them to obtain the median.

Measures of Central Tendency: An Example

Let's now consider measures of central tendency for the percentage of married women of childbearing age using contraception in 1984. We will look at three groups of countries: low-income countries, industrial market countries, and lower-middle-income countries.

A dot plot of contraception use is shown in Figure 3-4 for the 28 low-income countries with nonmissing values. This distribution is positively skewed: Most of the values are concentrated between 0 and 8%, while two countries

(Sri Lanka and China) have much higher levels of contraception use (57% and 69%, respectively). Four measures of central tendency are shown in Figure 3-4. The mean is greater than the median, inflated by the few large values. This is true in general for a positively skewed distribution. Eliminating extreme values from the calculation of the mean dampens this effect; the trimmed means are closer to the median. Note that trimming does not affect the median at all.

Figure 3-5 illustrates levels of contraception use for 13 industrial market countries with nonmissing values. This distribution is negatively skewed: Two countries have relatively low values compared with the others. The mean is less than the median for these 13 values; this is in general true for a negatively

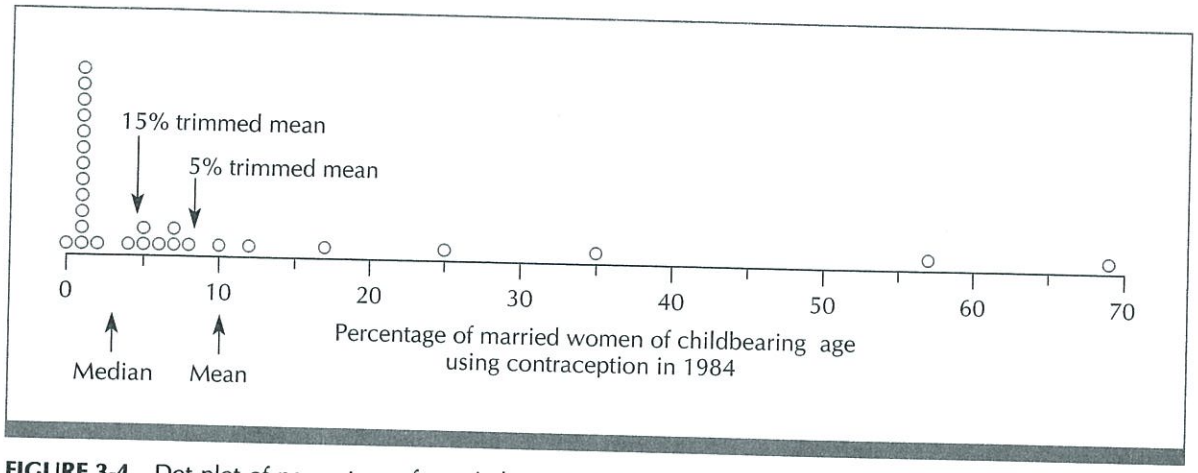

FIGURE 3-4 Dot plot of percentage of married women of childbearing age using contraception in 1984, for 28 low-income countries. Nine countries are excluded because of missing values.

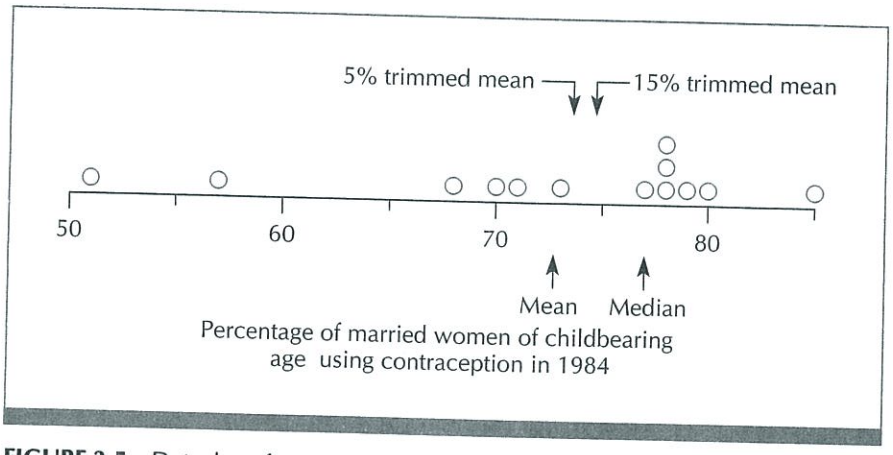

FIGURE 3-5 Dot plot of percentage of married women of childbearing age using contraception in 1984, for 13 industrial market countries. Six countries are excluded because of missing values.

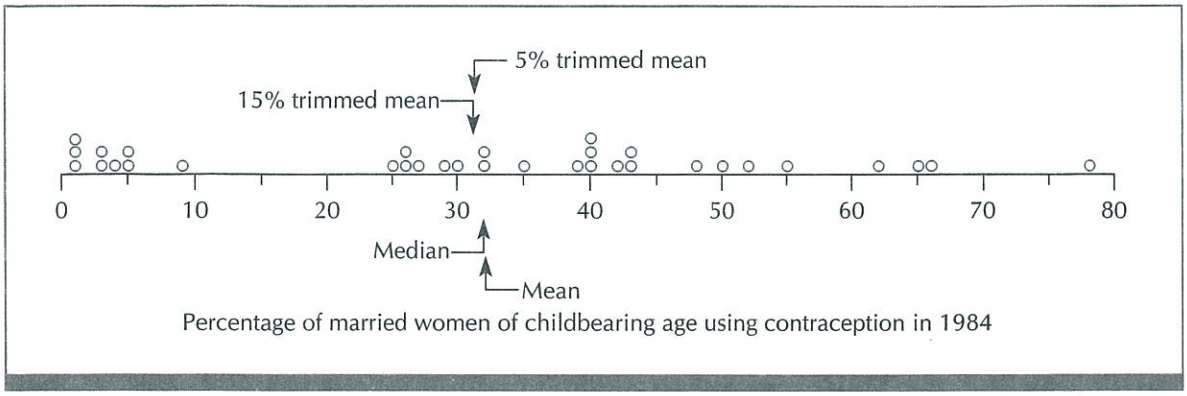

FIGURE 3-6 Dot plot of percentage of married women of childbearing age using contraception in 1984, for 33 lower-middle-income countries. Three countries are excluded because of missing values.

skewed distribution. The trimmed means lie between the median and the mean. Again, the median is the same for the trimmed and untrimmed sets of values.

Estimated levels of contraception use are plotted in Figure 3-6 for 33 lower-middle-income countries with nonmissing values. This dot plot is more symmetrical than the ones in Figures 3-4 and 3-5. When a distribution is nearly symmetric, measures of central tendency are similar. For these 33 levels of contraception use, the measures of central tendency illustrated in Figure 3-6 all equal either 31% or 32%.

The median contraception use is 3% for the low-income countries, 32% for the lower-middle-income countries, and 77% for the industrial market countries with available information. The lower-middle-income countries fall between the low-income countries and the industrial market countries with respect to contraception use, but there is a fair amount of overlap in the three distributions. Nine of the lower-middle-income countries, for example, have levels of contraception use less than 10%, typical of the low-income group. One lower-middle-income country (Mauritius) has a recorded contraception use of 78%, a typical value for the industrial market group.

We turn now to measures of variation.

3-2 Measures of Variation

As we know, measures of central tendency help describe the location of a set of values along a number line. But location alone provides a very incomplete description of a distribution. Compare Figures 3-2 and 3-3, for instance. We know that the center of the life expectancies for the industrial market countries (Figure 3-2) is different from the center for the low-income countries (Figure 3-3). But the spread in the values is also very different for the two groups of countries. The distribution of life expectancies for the industrial

market countries is very compact; the life expectancies are not far from each other. On the other hand, the distribution for the low-income countries is spread out; there is a great deal of variation among these values. This comparison suggests that we will have a more complete summary of a distribution if we describe variation as well as location.

One way to study the variation in a set of values is to look at the extremes, the *maximum* (or largest value) and the *minimum* (or smallest value). These two extremes define the *range*. Sometimes we think of the range as the difference between the maximum and minimum values. For the life expectancies of low-income countries, the range is from a minimum of 40 years to a maximum of 70 years; we say these life expectancies span a range of 30 years. On the other hand, the life expectancies for the industrial market countries span a range of 4 years, from a minimum of 74 to a maximum of 78 years. Certainly the range helps summarize the difference in variation between these two sets of values.

> The **range** is defined by the minimum (smallest) and the maximum (largest) values. The difference between these two extremes is a measure of spread or variation in a set of values.

The range depends only on the extreme values of a distribution; for this reason it may not provide a good summary of the variation among the bulk of the values. Instead of comparing the minimum and maximum values, we might compare the first and third quartiles. The first and third quartiles define the *interquartile range*. We know that about one-fourth of the values are below the first quartile and about one-fourth are above the third quartile. Therefore, the interquartile range from the first to the third quartile contains the middle 50% of the ordered data values. The interquartile range describes variation in the middle half of the distribution, whereas the range describes variation only between extremes of the distribution.

> The **interquartile range** is defined by the first and third quartiles. The difference between the third and first quartiles describes spread or variation in the middle half of a distribution of values.

We found the quartiles for life expectancies of low-income countries and for industrial market countries in Section 2-5. The middle 50% of the nonmissing life expectancies for low-income countries lie in the interquartile range from 45 to 52 years. For comparison, the middle 50% of the life expectancies for industrial market countries are in the interquartile range from 75 to 77 years. We might say the interquartile range is 7 years for the low-income countries, compared with 2 years for the industrial market countries. Note that the interquartile range is the length of the box portion of a box plot or box graph (see Figure 2-18, for instance).

The Standard Deviation as a Measure of Variation

The most commonly used measure of spread or variation is the *standard deviation,* often abbreviated SD. (We use the sample standard deviation in a num-

ber of formal statistical procedures for making inferences about population means and variances, as we will see in Part III.) In contrast to the range and interquartile range, we use all the data values to calculate the standard deviation, which measures variation of those values about the mean.

When the standard deviation measures variation in a population, we call it the *population standard deviation* and denote it by the Greek symbol σ; when it measures variation in a sample, we call it the *sample standard deviation* and denote it by the letter s. The computing formulas for the population standard deviation and the sample standard deviation differ slightly, as shown below. The sample standard deviation is used extensively in formal statistical inference. For purposes of data analysis we will follow common practice and use the calculation formula for the sample standard deviation; this is the formula used in computer statistical packages for calculating the standard deviation as a descriptive statistic.

To calculate a **population standard deviation,** we take the difference between each value and the mean, and square it. We add up the squared differences, and then divide by the number of data values. Then we take the square root:

$$\sigma = \text{Population standard deviation} = \sqrt{\frac{\text{Sum of (data value} - \text{mean)}^2}{\text{Number of values}}}$$

To calculate the **sample standard deviation,** we follow the same steps, except that we divide by 1 less than the number of data values:

$$s = \text{Sample standard deviation} = \sqrt{\frac{\text{Sum of (data value} - \text{mean)}^2}{\text{Number of values} - 1}}$$

We can write the calculation formula for the sample standard deviation, s, this way:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}{n - 1}}$$

where n denotes the number of values; x_1, x_2, through x_n represent the data values; and $\bar{x}$ denotes the sample mean. The second formula given above for the sample standard deviation is algebraically equivalent to the first, and is easier to use for hand calculations.

The standard deviation is always greater than or equal to 0. It tends to be large when values are very spread out about the mean, and to be small when values are close to the mean. If all the values are the same, then the standard deviation is 0. However, if the values are not all equal, the standard deviation will always be greater than 0.

Let's calculate the standard deviation of the life expectancies for the four high-income oil exporters, plotted in Figure 3-1. Using the formula for the sample standard deviation, Table 3-4 shows the steps in the calculation. The standard deviation is 5.9, or about 6 years.

TABLE 3-4 Calculating the standard deviation of 1985 life expectancy at birth for four high-income oil exporters, using the formula for the sample standard deviation for purposes of data analysis.

	Value Life expectancy (years)	Value − Mean Life expectancy − 66	(Value − Mean)^2 (Life expectancy − 66)^2
	60	−6	36
	62	−4	16
	70	4	16
	72	6	36
Sum	264	0	104

Mean: $\frac{264}{4} = 66$ years

Standard deviation: $\sqrt{\frac{104}{4 - 1}} = 5.9$ years

The second column of Table 3-4 contains some interesting information—distances of the individual values from the mean. These distances are examples of residuals. In general, a *residual* is the difference between an observation and a summary value. (Later, when we discuss modeling experimental results as part of statistical inference, we will see that this summary value is an estimate of the mean, or a predicted value, of the observation based on a probability model we assume for the experiment.)

> A **residual** is the difference between an observation and a summary value for the observation.

The summary value we use in this simple example is the mean of the four observations. A residual is then the difference between an observation and this mean. The residuals in column 2 of Table 3-4 show us that the life expectancies range from 6 years below the mean to 6 years above the mean. If we plot these residuals, we get a dot plot that has the same appearance as the dot plot of life expectancies in Figure 3-1, but the values are shifted, with a mean of 0. These shifted values (or residuals) give us a feel for the variation about the mean. Statisticians often look at differences such as these; we find residuals very useful for many analyses in formal statistical inference.

The standard deviation of the 35 life expectancies for low-income countries (Figure 3-3) is 6.9 years; recall that the range is 30 years and the interquartile range is 7 years. The standard deviation of life expectancies for the 19 industrial market countries (Figure 3-2) is 1.3 years; the range is 4 years and the interquartile range is 2 years. For simple data analysis, we may prefer to use the interquartile range and the range as measures of variation; the variation they represent is easier to visualize than for the standard deviation.

Recall that we used box graphs to compare population growth for low-income and industrial market countries in Figure 2-20. We noticed that the variation in the two sets of values appeared to be about the same. This observation is supported when we compare measures of variation. For the low-

income countries, the range in population growth values is 2.9, the interquartile range is .8, and the standard deviation is .6. The range in population growth values for the industrial market countries is 1.6, the interquartile range is .8, and the standard deviation is .4.

In Chapters 2 and 3, we have considered ways to look at a single variable. In Chapters 4 and 5, we look at ways to study relationships between two or more variables.

Summary of Chapter 3

Measures of central tendency describe the location of a set of values. The mean or average can be greatly influenced by extreme values. Trimmed means are less affected by extremes, whereas the median is not affected at all. Sometimes we find it useful to calculate a weighted mean, in which a value's contribution to the calculation is determined by a weight assigned to that value. (When we weight the values equally, we get the mean.)

We consider three measures of spread or variation. The range is defined by the extremes, while the interquartile range is defined by the middle 50% of the ordered values. The standard deviation is a measure of variation about the mean calculated from all the data values. The calculation formulas differ slightly for the population standard deviation and the sample standard deviation. When calculating the standard deviation as a descriptive statistic, we follow common practice and use the calculation formula for the sample standard deviation. We use the sample standard deviation extensively in classical statistical inference. For simple data analysis, the interquartile range and range are easier to interpret than the standard deviation.

Minitab Appendix for Chapter 3

Calculating Descriptive Statistics

The DESCRIBE command will provide all of the descriptive statistics in Chapter 3 except the weighted mean. Referring to the data for Exercise 5-35, the command

```
MTB> describe 'calorie'
```

produces the output in Figure M3-1.

The N column gives the number of nonmissing observations; N*, the

	N	N*	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
calorie	37	3	110.00	114.00	110.58	22.38	3.68

	MIN	MAX	Q1	Q3
calorie	68.00	143.00	91.50	130.00

FIGURE M3-1 Descriptive statistics for the variable CALORIE in Exercise 5-35.

Exercises for Chapter 3

In the exercises, answer the following questions: What would you need to know about the sample to be willing to use it to make inferences about a larger population? What is that larger population (if any)? What limitations do you see in the sample?

For each figure and table, include a legend that completely describes its contents. Note the number of cases included and the number excluded if there are missing values.

EXERCISE 3-1

Consider this simple (hypothetical) example of how different a mean and median can be. Suppose that while looking for a job, you discover that the average salary of employees at Company A last year was $135,000. Intrigued, you do some investigation. You discover that Company A has a president, plus seven other employees. You are able to determine last year's salaries for these other seven: $8,000; $8,000; $9,000; $9,000; $11,000; $15,000; $20,000. You are surprised at how low these salaries are, until you find last year's salary for the president: $1,000,000. Do you want to work for Company A? Compare the mean with the median and a trimmed mean as measures of central tendency for these eight salaries. Which measure, if any, conveys a more correct impression of a typical salary?

EXERCISE 3-2

Population in mid-1985 (millions), life expectancy at birth in 1985 (years) and their product are shown below for 19 industrial market countries.

Country	Population	Life expectancy	Product
Australia	15.8	78	1,232.4
Austria	7.6	74	562.4
Belgium	9.9	75	742.5
Canada	25.4	76	1,930.4
Denmark	5.1	75	382.5
Finland	4.9	76	372.4
France	55.2	78	4,305.6
Germany, West	61.0	75	4,575.0
Ireland	3.6	74	266.4
Italy	57.1	77	4,396.7
Japan	120.8	77	9,301.6
Netherlands	14.5	77	1,116.5
New Zealand	3.3	74	244.2
Norway	4.2	77	323.4
Spain	38.6	77	2,972.2
Sweden	8.4	77	646.8
Switzerland	6.5	77	500.5
United Kingdom	56.5	75	4,237.5
United States	239.3	76	18,186.8
Sum	737.7	1,445	56,295.8

a. Find the mean life expectancy for these 19 countries.

b. Find the 10% trimmed mean life expectancy for these countries.

c. Find the weighted mean life expectancy, with life expectancies weighted by population size.

d. Compare the three measures of central tendency from parts (a), (b), and (c).

e. How do these three measures compare with the life expectancy of the United States, the industrial market country with the largest population?

EXERCISE 3-3

Population in mid-1985 (millions), life expectancy at birth in 1985 (years), and their product are shown below for 35 low-income nations.

Country	Population	Life expectancy	Product
Bangladesh	100.6	51	5,130.6
Benin	4.0	49	196.0
Bhutan	1.2	44	52.8
Burkina Faso	7.9	45	355.5
Burma	36.9	59	2,177.1
Burundi	4.7	48	225.6
Central African Republic	2.6	49	127.4
Chad	5.0	45	225.0
China	1,040.3	69	71,780.7
Ethiopia	42.3	45	1,903.5
Ghana	12.7	53	673.1
Guinea	6.2	40	248.0
Haiti	5.9	54	318.6
India	765.1	56	42,845.6
Kenya	20.4	54	1,101.6
Lao PDR	3.6	45	162.0
Madagascar	10.2	52	530.4
Malawi	7.0	45	315.0
Mali	7.5	46	345.0
Mozambique	13.8	47	648.6
Nepal	16.5	47	775.5
Niger	6.4	44	281.6
Pakistan	96.2	51	4,906.2
Rwanda	6.0	48	288.0
Senegal	6.6	47	310.2
Sierra Leone	3.7	40	148.0
Somalia	5.4	46	248.4
Sri Lanka	15.8	70	1,106.0
Sudan	21.9	48	1,051.2
Tanzania	22.2	52	1,154.4
Togo	3.0	51	153.0
Uganda	14.7	49	720.3
Viet Nam	61.7	65	4,010.5
Zaire	30.6	51	1,560.6
Zambia	6.7	52	348.4
Sum	2,415.3	1,757	146,424.4

a. Find the mean life expectancy for these 35 low-income countries.

b. Find the 15% trimmed mean life expectancy for these countries.

c. Find the weighted mean life expectancy, with life expectancy weighted by population size.

d. Compare the three measures of central tendency in parts (a), (b), and (c).

e. Now consider just China and India. Find the mean life expectancy for these two countries. Find the weighted mean life expectancy, with life expectancy weighted by population size. Compare these two measures of central tendency.

f. Consider the 33 low-income countries that remain after you exclude China and India. Find the mean life expectancy for these 33 countries. Find the weighted mean life expectancy, with life expectancy weighted by population size. Compare these two measures of central tendency.

g. Why are the mean and weighted mean life expectancy you calculated for the 33 countries in part (f) closer in value than the mean and weighted mean life expectancy for all 35 countries from parts (a) and (c)?

EXERCISE 3-4

The percentage of married women of childbearing age using contraception in 1984 is shown below for each of 28 low-income countries with information available.

Country	Contraception use	Country	Contraception use
Somalia	0	Sierra Leone	4
Burkina Faso	1	Burma	5
Burundi	1	Sudan	5
Chad	1	Benin	6
Guinea	1	Haiti	7
Malawi	1	Nepal	7
Mali	1	Pakistan	8
Niger	1	Ghana	10
Rwanda	1	Senegal	12
Tanzania	1	Kenya	17
Uganda	1	Bangladesh	25
Zaire	1	India	35
Zambia	1	Sri Lanka	57
Ethiopia	2	China	69

a. Find the median, mean, 5% trimmed mean, and 15% trimmed mean of contraception use for these 28 low-income countries.

b. Find the weighted mean contraception use, with contraception use weighted by population size. Population sizes for these countries are given in Exercise 3-3.

c. Compare the measures of central tendency you found in parts (a) and (b).

EXERCISE 3-5

As part of a study of vegetation damage, researchers measured the oxidant content of dew water in 12 samples collected at Port Burwell, Ontario, from August 25 to August 30, 1960. The 12 measurements in parts per million (ppm) ozone are (Hollander and Wolfe, 1973, page 57; from Cole and Katz, 1966):

.08 .11 .15 .17 .17 .20 .21 .22 .28 .31 .32 .35

a. Construct a dot plot of these observations.

b. Find the mean, 15% trimmed mean, and median. Locate these three measures of central tendency on the plot. How do they compare?

c. Find the range, interquartile range, and standard deviation. What does each of these descriptive statistics measure?

d. Describe the distribution of these ozone measurements.

EXERCISE 3-6

Operating hours until first failure of air-conditioning equipment are shown below for 13 Boeing 720 airplanes (Hollander and Proschan, 1984, page 176; from Proschan, 1963).

23 50 50 55 74 90 97 102 130 194 359 413 487

a. Construct a dot plot of these observations.

b. Is the distribution unimodal, bimodal, or multimodal?

c. Would you describe this distribution as symmetrical, negatively skewed, positively skewed, or none of these?

d. Calculate the mean, 15% trimmed mean, and median of these values. Locate these three measures of central tendency on your plot and compare them.

e. Find and discuss these three measures of variation: range, interquartile range, standard deviation.

EXERCISE 3-7

The following are errors (in inches) made by a robot in applying an adhesive (Milton and Arnold, 1986, page 250; based on Hegland, 1983):

.001 .002 .003 .004 .006 .001 .002 .003 .004 .006 .001
.003 .003 .006 .001 .003 .003 .005 .007 .002 .003 .004
.005 .008 .004

These values range from .001 to .008 inch. We often find it helpful in calculations to use coded or transformed values. If we change our units to thousandths of an inch, then we multiply each number above by 1,000. The transformed values then range from 1 to 8. You may wish to use these new units in your calculations below.

a. Construct a dot plot. Describe the distribution in terms of location, spread, peaks, and symmetry or skewness.

b. Calculate the mean, 15% trimmed mean, and median of these values. Locate these three measures of central tendency on your plot and compare them.

c. Calculate and discuss these three measures of variation: range, interquartile range, standard deviation.

EXERCISE 3-8

A new chip can be reprogrammed without removing it from the microcomputer. Times (in seconds) to reprogram a byte of memory on this chip are shown below (Milton and Arnold, 1986, page 258; from *Design News,* April 1983, page 26).

11.6	12.3	12.5	12.9	13.0	13.1	13.2	13.3
13.3	13.4	13.8	14.2	14.7	15.1	15.3	

a. Construct a dot plot of these programming times.

b. Calculate the mean, 15% trimmed mean, and median of these values. Locate these three measures of central tendency on your plot and compare them.

c. Calculate the range, interquartile range, and standard deviation. What do these three descriptive statistics measure?

d. Describe this distribution in terms of location, spread, peaks, and symmetry or skewness.

e. A company has claimed that a byte of memory on this chip can be reprogrammed in less than 14 seconds. Does this claim seem reasonable?

EXERCISE 3-9

Consider the following measurements of leaf protein (mg/g fresh weight) from six plants of a variety of soybean (Devore, 1982, page 23; from *Science,* volume 199, page 974):

4.9	5.1	6.1	11.7	14.0	16.1

a. Construct a dot plot of these six measurements.

b. Describe this distribution in terms of location, spread, peaks, and symmetry or skewness.

c. Calculate the mean, 15% trimmed mean, and median of these values. Locate these three measures of central tendency on the plot and compare them.

d. Find the range, interquartile range, and standard deviation. Discuss these three measures of variation.

EXERCISE 3-10

Researchers measured pulmonary compliance ($cm^3/cm\ H_2O$) for each of 16 construction workers who had been exposed over a long period to asbestos. Pulmonary compliance is a measure of how well the lungs expand and contract. These measurements were taken 8 months after asbestos exposure (Devore, 1987, page 275; from "Acute Effects of Chrysotile Asbestos Exposure on Lung Function," *Environ. Research,* 1978, pages 360–372).

167.9	180.8	184.8	189.8	194.8	200.2	201.9	206.9
207.2	208.4	226.3	227.7	228.5	232.4	239.8	258.6

a. Construct a dot plot of these measurements.

b. Calculate the mean, 5% trimmed mean, and median of these values. Locate these three descriptive statistics on your plot and compare them.

c. Find the range and interquartile range.

d. Describe this distribution in terms of location, variation, peaks, and symmetry or skewness.

EXERCISE 3-11

Researchers measured levels of the amino acid alanine (in mg/100 ml) for six normal baby boys on an isoleucine-free diet (Devore, 1987, page 275; from "The Essential Amino Acid Requirements of Infants," *Amer. J. Nutrition,* 1964, pages 322–330):

1.44 2.70 2.80 2.84 2.94 3.54

a. Construct a dot plot of these measurements.

b. Is this distribution symmetrical, negatively skewed, or positively skewed?

c. Calculate the mean, 15% trimmed mean, and median of these values. Locate these three measures of location on your plot and compare them.

d. Find the range, interquartile range, and standard deviation. Discuss these three measures of variation.

EXERCISE 3-12

Investigators measured radiation levels (in milliroentgens per hour) in the television display areas of 10 department stores (Devore, 1987, page 301; from "Many Set Color TV Lounges Show Highest Radiation," *J. Environmental Health,* 1969, pages 359–360):

.15 .16 .36 .40 .48 .50 .50 .60 .80 .89

You may wish to change units by multiplying each value by 100 before doing any calculations.

a. Construct a dot plot of these radiation levels.

b. Would you describe this distribution as symmetrical, negatively skewed, positively skewed, or none of these?

c. Calculate the mean, 10% trimmed mean, and median of these values. Locate these three descriptive statistics on your plot and compare them.

d. Find the range, interquartile range, and standard deviation. Discuss these three measures of variation.

e. The limit recommended for such radiation exposure is .50 milliroentgen per hour. Based on these sample measurements, discuss the safety of department store television display areas at the time this study was done.

EXERCISE 3-13

A softball player wanted to compare his hitting distance with two types of bats (wooden and aluminum) and two brands of balls. He hit four balls pitched from a pitching machine for each combination of bat and ball type. The distances he hit the softballs are shown below (Shaughnessy, 1988).

Bat, ball type	Distances hit (feet)			
Wooden bat, Dudley Thunder	230	242	242	250
Wooden bat, Worth Red Dot	258	264	265	275
Aluminum bat, Dudley Thunder	265	270	277	282
Aluminum bat, Worth Red Dot	290	302	310	318

a. Draw a dot plot for each of the four sets of distances. Use the same scale for each plot.

b. Find the mean, standard deviation, and range for each of the four sets of distances.

c. Using the wooden bat, on average what was the player's hitting advantage using the Worth Red Dot ball rather than the Dudley Thunder?

d. Using the aluminum bat, on average what was the player's hitting advantage using the Worth Red Dot rather than the Dudley Thunder?

e. Looking at your answers to parts (c) and (d), is the hitting advantage of the Worth Red Dot softball the same for the wooden and aluminum bats?

f. Using the Worth Red Dot softball, on average what was the player's hitting advantage using the aluminum rather than the wooden bat?

g. Using the Dudley Thunder softball, on average what was the player's hitting advantage using the aluminum rather than the wooden bat?

h. Looking at your answers to parts (f) and (g), is the hitting advantage of the aluminum bat the same for the two brands of softball?

i. Compare the range and standard deviation across the four sets of distances. What can you say about the variation in distances hit under the four sets of conditions?