ch. 3

FUNDAMENTAL STATISTICS FOR BEHAVIORAL SCIENCES

FIFTH EDITION

Robert B. McCall
University of Pittsburgh

Under the General Editorship of
Jerome Kagan
Harvard University

Harcourt Brace Jovanovich, Publishers
San Diego New York Chicago Austin Washington, D.C.
London Sydney Tokyo Toronto

COPYRIGHTS AND ACKNOWLEDGMENTS

LONGMAN GROUP Tables B, C, G, and K are taken from Tables III, VII, IV, and XXXIII of Fisher and Yates: *Statistical Tables for Biological, Agricultural and Medical Research* published by Longman Group Ltd., London (previously published by Oliver and Boyd Ltd., Edinburgh) and by permission of the authors and publishers.

ISBN: 0-15-529476-8

Library of Congress Catalog Card Number: 89-85302

Printed in the United States of America

CHAPTER 3

CHARACTERISTICS OF DISTRIBUTIONS

TWO KEY CHARACTERISTICS of a distribution of scores are its *central tendency* (the value of a "typical" score) and its *variability* (the extent to which the scores differ from their central tendency). The purpose of this chapter is to introduce some of the more common numerical indices of central tendency and variability.

The most common measure of central tendency is the mean, or average, score; other measures of central tendency are the median and the mode. Three indices are used to express variability: the range, the variance, and the standard deviation. These measures—and the methods used to calculate their values—are explained in this chapter. A discussion of statistical estimation is also presented to show how values from studies with a limited number of subjects can be used to estimate characteristics of much larger groups. Estimation is a fundamental process of inferential statistics, and will be examined in more detail in Part 2 of this text. The concept is introduced here because descriptive statistics, as presented in this chapter, are often used to estimate corresponding values in larger groups of subjects.

MEASURES OF CENTRAL TENDENCY

The Mean

The most common measure of the central tendency of a group of scores is the average, or mean:

The mean of scores on variable X is symbolized by $\overline{X}$ (read "X bar") and is computed with the formula

$$\overline{X} = \frac{\Sigma X_i}{N}$$

which instructs one to add all the scores (that is, ΣX_i) and divide by N, which is the number of scores in the distribution.

For example:

$$\begin{array}{r} X_i \\ \hline 8 \\ 3 \\ 4 \\ 10 \\ 7 \\ 1 \\ \hline \Sigma X_i = 33 \end{array}$$

$$N = 6 \quad \bar{X} = \frac{\Sigma X_i}{N} = \frac{33}{6} = 5.5$$

The mean is also called the arithmetic average of the scores. Notice that ΣX_i can be written without the limits $i = 1$ and N (see Chapter 1), which are understood to apply. For example, $\sum_{i=1}^{N} X_i = \Sigma X_i = \Sigma X$. This abbreviated notation is common in the remainder of the text.

Different texts use different symbols to represent the mean. A common alternative to the symbol $\bar{X}$ used in this book is the symbol M.

Deviations about the Mean The mean possesses several properties that make it an appropriate measure of central tendency. First,

> **The sum of the deviations of scores about their mean is zero. In other words, if the mean is subtracted from each score in the distribution, the sum of such differences is zero. In symbols**
>
> $$\Sigma(X_i - \bar{X}) = 0$$

Consider the following numerical example:

X_i	$\bar{X}$	$(X_i - \bar{X})$
3	5	$3 - 5 = -2$
6	5	$6 - 5 = 1$
5	5	$5 - 5 = 0$
1	5	$1 - 5 = -4$
10	5	$10 - 5 = 5$
$\Sigma X_i = 25$		$\Sigma(X_i - \bar{X}) = 0$
$N = 5$		
$\bar{X} = 5$		

OPTIONAL TABLE

3–1 PROOF THAT THE SUM OF THE DEVIATIONS ABOUT THE MEAN EQUALS ZERO

Operation	Explanation
To prove $\Sigma(X_i - \bar{X}) = 0$	
1. $\Sigma(X_i - \bar{X}) = \Sigma X_i - \Sigma \bar{X}$	1. The sum of the differences between two quantities equals the difference between their sums.
2. $= \Sigma X_i - N\bar{X}$	2. The sum of a constant $\bar{X}$ added to itself N times (i.e., $\sum_{i=1}^{N} \bar{X}$) is N times the constant (i.e., $N\bar{X}$).
3. $= \Sigma X_i - N\left(\frac{\Sigma X_i}{N}\right)$	3. Substitution of the definition $\frac{\Sigma X_i}{N}$ for $\bar{X}$
4. $\Sigma(X_i - \bar{X}) = \Sigma X_i - \Sigma X_i = 0$	4. Cancellation of N's in second term of #3 above.

This principle can be proven true of all distributions. The algebraic proof is presented in Optional Table 3–1.

While the sum of the deviations of all the scores about the mean is always zero, the sum of the *squared* deviations about the mean is usually not zero. This distinction is important because formulas presented later in the chapter (such as those for the variance) use squared deviations.

Whereas

$$\Sigma(X_i - \bar{X}) = 0$$

the expression

$\Sigma(X_i - \bar{X})^2$ is *not* usually equal to ($\neq$) 0

To illustrate, consider the numerical example given above. If one squares the difference between each score and the mean (the numbers in the extreme right-hand column) and then sums these squared deviations, one obtains

$$(-2)^2 + (1)^2 + (0)^2 + (-4)^2 + (5)^2 = 46$$

OPTIONAL TABLE

3–2 PROOF THAT THE SUM OF THE SQUARED DEVIATIONS ABOUT THE MEAN IS A MINIMUM

Operation	Explanation
1. $(\bar{X} + c)$, $c \neq 0$ and $c \neq \bar{X}$	1. Assumption.
2. (a) The sum of the squared deviations of scores about $\bar{X}$ equals $\Sigma(X_i - \bar{X})^2$ (b) The sum of the squared deviations of scores about $(\bar{X} + c)$ equals $\Sigma[X_i - (\bar{X} + c)]^2$	2. Definitions.
To prove $\Sigma(X_i - \bar{X})^2 < \Sigma[X_i - (\bar{X} + c)]^2$	To prove that the sum of squared deviations about the mean, $\Sigma(X_i - \bar{X})^2$, is less than the sum of squared deviations about any other value, $\Sigma[X_i - (\bar{X} + c)]^2$.
3. $< \Sigma[(X_i - \bar{X}) - c]^2$	3. Working with the right side of the inequality, removing parentheses and regrouping.
4. $< \Sigma[(X_i - \bar{X})^2 - 2c(X_i - \bar{X}) + c^2]$	4. Binomial expansion of the form: $(a - b)^2 = a^2 - 2ab + b^2$
5. $< \Sigma(X_i - \bar{X})^2 - \underbrace{2c\Sigma(X_i - \bar{X})}_{0} + \underbrace{\Sigma c^2}_{Nc^2}$	5. Distributing the summation sign to all terms within the brackets, the sum (or difference) of several variables is the sum (or difference) of their sums, and the sum of a constant times a variable is the constant times the sum of the variable.
6. $< \Sigma(X_i - \bar{X})^2 - 0 + Nc^2$	6. Since $\Sigma(X_i - \bar{X}) = 0$, the second term is 0, and the sum of Nc^2's equals N times c^2.
7. $\Sigma(X_i - \bar{X})^2 < \Sigma(X_i - \bar{X})^2 + Nc^2$	7. The expression is true because Nc^2 will always be greater than zero.

which is obviously not zero. The reason squared deviations never add to zero (unless all the scores are the same) is that squared numbers can never be negative, and thus positive values will not be balanced by negative ones.

Minimum Variability of Scores about the Mean A second property of the mean concerns the squared deviations of scores about their mean:

> **The sum of the squared deviations of scores about their mean is less than the sum of the squared deviations of the same scores about any other value.**

This fundamental principle will be invoked in the explanation of many subsequent concepts. It states that although the sum of the squared deviations of scores about their mean usually does not equal zero, that sum is nevertheless smaller than if the squared deviations of the same scores were taken about any value other than the mean of their distribution. For example, in the above illustration the sum of the squared deviations about the mean equaled 46. The mean of that distribution was 5.0. The sum of the squared deviations about the number 6.0 equals 51; about the number 4.0 the sum equals 51; and about the number 7.0 it equals 66. The sum of squared deviations about the mean (46) is less than any of these examples, and it can be shown that it always will be less than about any other value. It is in this sense, sometimes called the *least squares sense*, that the mean is an appropriate measure of central tendency: the mean is closer (in terms of squared deviations) to the individual scores over the entire group than is any other single value.

The proof that the sum of the squared deviations about the mean is always less than the sum of the squared deviations about any alternative value is presented in Optional Table 3–2.

The Median

Another measure of central tendency is the median:

> **The median, symbolized by M_d, is the point that divides the distribution into two parts such that equal numbers of scores fall above and below that point.**

The way the median is computed varies, depending first on whether there is an odd or an even number of scores in the distribution and second on whether

there is a duplication of score values near the median point. The phrase "duplication of score values" means that more than one score of the same value exists in the distribution. The distribution (3, 4, 5, 5, 7) has a duplication of score values (the two 5's) while the distribution (2, 3, 5, 6, 8) does not. Duplication of score values is important only when it occurs near the point where the median is located. Otherwise, score duplication can be ignored.

1. **No duplication near the median; odd number of scores.** When there is an odd number of scores and no duplication of scores near the median, the median is the middle score. For example, in the distribution (3, 5, 6, 7, 10), the point that divides the distribution into two equal parts is 6, since two scores fall below and two scores fall above this value.
2. **No duplication near the median; even number of scores.** By custom, when there is an even number of scores in a distribution and no duplication near the median, the average of the middle two scores is taken as the median. Suppose the distribution is (3, 5, 6, 7, 10, 14). The point that divides the distribution in half lies between 6 and 7. The average of these points, 6.5, is taken as the median. Another example illustrates the same custom is followed when the scores near the median are not adjacent values. If the distribution is (3, 3, 4, 8, 14, 16), the median is 6 because $(4 + 8) \div 2 = 6$. Notice that the score duplication (the two 3's) is not considered because it is not near the median point.
3. **Duplication of scores near the median.** When more than one instance of a score value falls near the median, the median is obtained by a procedure called *linear interpolation*, which proceeds in basically the same way regardless of whether the number of scores in the distribution is odd or even. To illustrate, suppose the distribution is (3, 4, 5, 5, 5, 6, 6, 7). Since the median is the point dividing the distribution in such a way that an equal number of scores fall below and above it, the median lies somewhere between the second and third instance of the score 5. Presumably, the scores 3, 4, 5, 5 are below the median point and 5, 6, 6, 7 are above it. A single numerical value that expresses this situation can be determined by observing that the scores 3, 4, and two of the three scores of 5—that is two-thirds of the 5s—must be below the median. The score value of 5 occupies the score interval bordered by the real limits of 5, namely 4.5 to 5.5. Its width is 1. If the three scores of 5 are assumed to be equally spaced within the score interval, then two of the three scores will occupy two-thirds of the interval of size 1, which is $\frac{2}{3}(1) = .67$. Adding $\frac{2}{3} = .67$ to the lower real limit of this score interval, (that is, 4.5) gives

$$4.5 + .67 = 5.17$$

as the median. This same process is graphically illustrated in Figure 3–1, in which scores are placed on the scale within the real limits of their score values. It can be seen that a total of four of the eight frequencies

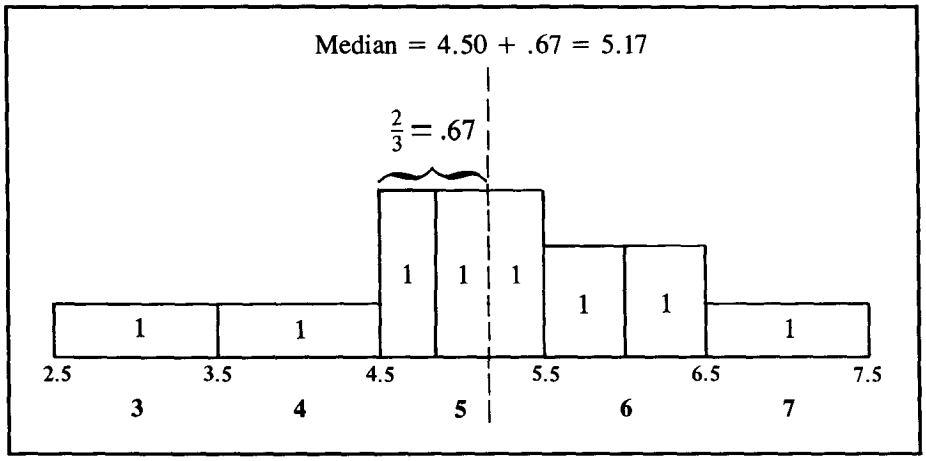

Figure 3–1 Computing the median when there is duplication of scores (even number of frequencies).

(scores) lie above and four lie below the median of 5.17. In short, the median divides the boxed area of Figure 3–1 into two equal portions.

The logic is the same if an odd number of frequencies are in the distribution. Suppose the distribution consists of (3, 4, 5, 5, 5, 6, 6, 7, 7), which case is presented in Figure 3–2. There are 9 scores, and thus the median must be the point such that $4\frac{1}{2}$ frequencies fall below and $4\frac{1}{2}$ fall above it. Counting from the low end upward, the scores 3 and 4 plus $2\frac{1}{2}$ of the 3 scores of 5 will be below the median. Therefore, $2\frac{1}{2}$ of 3, or

$$\frac{2\frac{1}{2}}{3} = \frac{\frac{5}{2}}{3} = \frac{5}{6} = .83$$

of the score interval having a size of 1, that is, .83(1) = .83, must be added to the lower real limit of the score interval of 5 (that is, 4.5), which gives

$$4.5 + .83 = 5.33$$

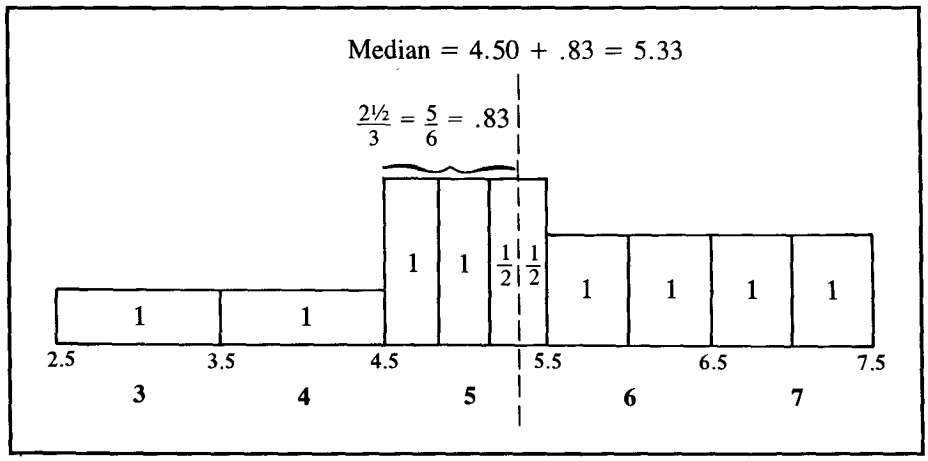

Figure 3–2 Computing the median when there is duplication of scores (odd number of frequencies).

as the median. Again, if the frequencies in Figure 3–2 are added, it can be seen that the median point 5.33 indeed has $4\frac{1}{2}$ frequencies below and $4\frac{1}{2}$ frequencies above it.

The following formula represents these linear interpolation steps and can be used whenever there is duplication of scores near the median. It is assumed that the data are not grouped into class intervals.

The median is computed by

$$M_d = L + \left(\frac{N/2 - n_b}{n_w}\right) i$$

in which

M_d = the median

L = the lower real limit of the score interval containing the median

N = the number of scores in the total distribution

n_b = the number of scores falling below the lower real limit of the score interval containing the median

n_w = the number of cases within the score interval containing the median

i = the size of the score interval ($i = 1$ if the data are in whole numbers; $i = .1$ if the data are in tenths; and so on)

In terms of the last example,

$L = 4.5$

$N = 9$

$n_b = 2$

$n_w = 3$

$i = 1$

$$M_d = L + \left[\frac{N/2 - n_b}{n_w}\right] i$$

$$= 4.5 + \left[\frac{\frac{9}{2} - 2}{3}\right] 1$$

$$= 4.5 + \frac{\frac{5}{2}}{3}$$

$$= 4.5 + .83$$

$$M_d = 5.33$$

The Mode

A third measure of central tendency is the mode:

> **The mode, symbolized M_0, is the most frequently occurring score.**

If the distribution is (3, 4, 4, 5, 5, 5, 6, 8), the mode is 5. Sometimes a distribution will have two modes, such as the distribution (3, 4, 4, 4, 5, 6, 6, 7, 7, 7, 8). In this case, the modes are 4 and 7 and this distribution is called **bimodal**. A distribution that contains more than two modes is called **multimodal**.

Comparison of the Mean, Median, and Mode

The essential difference between the mean and the median is that the mean reflects the value of each score in the distribution, whereas the median is based largely on where the midpoint of the distribution falls, without regard for the particular value of many of the scores. For example, consider the following illustration:

Scores	Mean	Median
1, 2, 3, 4, 5	3	3
1, 2, 3, 4, 50	12	3
1, 2, 3, 4, 100	22	3

Only the last number differs from one distribution to the other. The mean reflects these differences, but the median does not. This is because the median is the midpoint of the distribution such that an equal *number* of scores fall above and below it. The particular *value* of the extreme scores does not matter since only the fact that those scores are above the midpoint is considered. In contrast, the mean takes into account the value of every score. Thus, changing any score value will likely change the value of the mean.

The mode reflects only the most frequently occurring score. It is not used much in the behavioral sciences, except to describe a bimodal or highly skewed distribution.

Because the three different measures of central tendency are sensitive to different aspects of the group of scores, they are usually not the same value in a given distribution. If the distribution is symmetrical and unmimodal (having one mode), then the mean, median, and mode are indeed identical. This

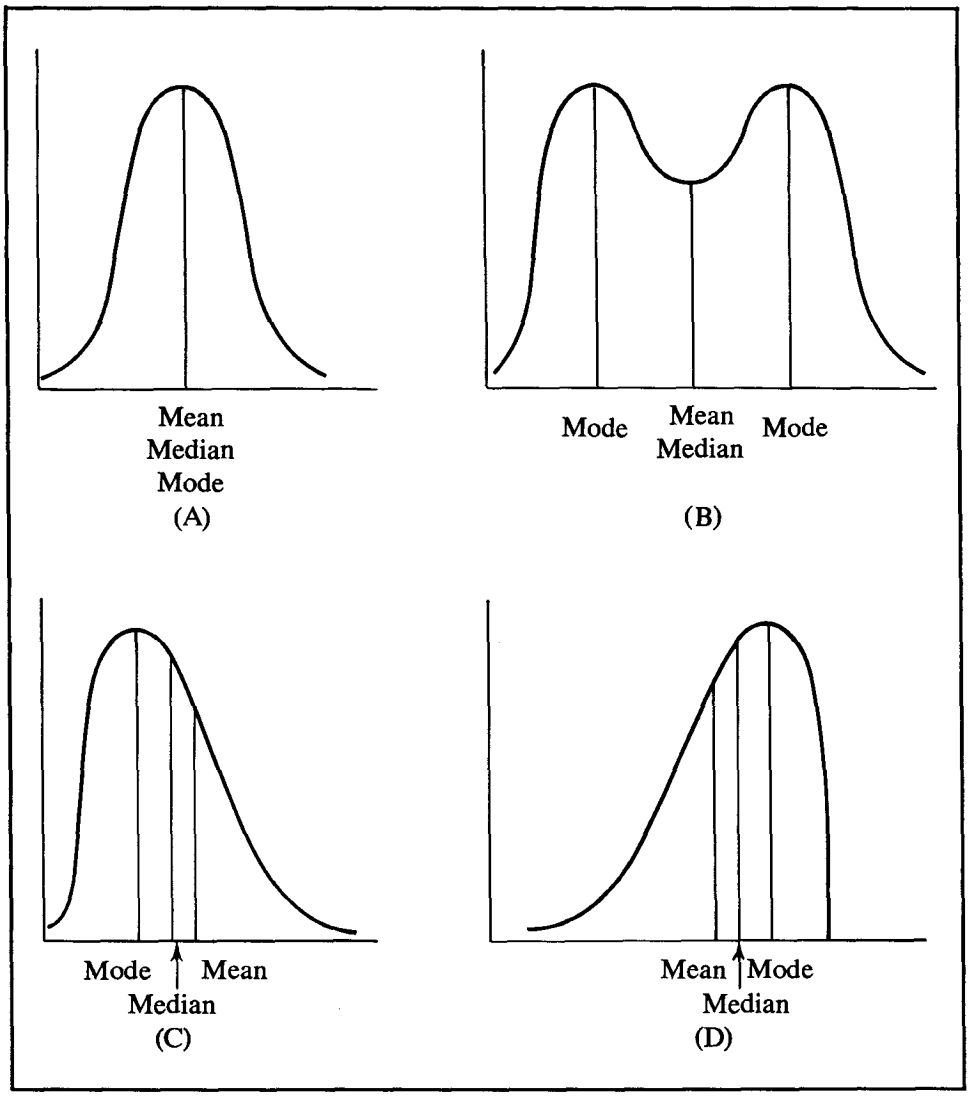

Figure 3 – 3 Mean, median, and mode in different distributions.

condition is graphed in part A of Figure 3–3. If the distribution is symmetrical but has two modes as in part B, the mean and median are the same but the modes are different (the distribution is bimodal). In Chapter 2, a skewed distribution was defined as a distribution that is not symmetrical, having scores bunched at one end. Parts C and D of Figure 3–3 show two skewed distributions and the relative positions of the three measures of central tendency. The distribution in part C illustrates a condition in which most scores have moderate values but a few are very high. In this case, the mean, being sensitive to those extreme values, is somewhat larger than the median, which divides the area under the curve (that is, the total number of cases) into two equal parts. A common instance of this situation occurs in the reporting of typical family income. The mean family income is usually higher than the

median income because the relatively few really high incomes push the mean upward without influencing the median. Part D of Figure 3–3 illustrates the relative positions of the measures of central tendency when the skewness of the distribution is in the other direction.

Ordinarily, behavioral science researchers select the mean as the measure of central tendency. There are several reasons for preferring the mean, but one major consideration is that the mean is required by so many other statistical procedures. However, sometimes the circumstances are such that the median would reflect the central tendency of the distribution more accurately than would the mean. When the distribution is very skewed, the mean may not be a value that coincides with one's intuitive impression of the typical score. For example, the distribution (1, 2, 3, 4, 100) has a mean of 22 and a median of 3. In this case the median seems to characterize the central tendency more faithfully than does the mean. As noted above, annual incomes are often skewed to the right—very few people have very high incomes. The median, then, is often a better measure of central tendency of annual incomes than the mean. Thus, in the case of a markedly skewed distribution, the median may be preferred. The mode is rarely used by itself to express central tendency. It is most often reported as a supplement to the mean or median, especially for distributions that are skewed or bimodal.

MEASURES OF VARIABILITY

In addition to an index of central tendency, a measure of variability is needed to characterize a distribution more fully.

Variability refers to the extent to which the scores in a distribution differ from their central tendency.

For example, suppose two groups of scores, A and B, are defined to be

$$A = (5, 7, 9)$$

$$B = (3, 7, 11)$$

Although they both have the same mean of 7, set B has more variability because the scores differ from that mean more than do the scores in A. The purpose of this section is to present numerical measures of the variability of scores in a distribution.

The Range

One measure of variability is the range. The range may be determined by taking the largest score minus the smallest score in the distribution.[1] In the distribution (3, 5, 6, 6, 8, 9), the range is $9 - 3 = 6$.

However, the range is limited in its ability to reflect the variability of a distribution. It is not sensitive to the variability of *all* the scores, only to the difference between the two most extreme values. For example,

$$C = (5, 10, 11, 12, 13, 18)$$

$$D = (5, 6, 7, 8, 16, 17, 18)$$

have the same range, but D has more variability than C. Therefore, although the range is easily computed, it is usually employed only as a crude approximation of variability.

The Variance and the Standard Deviation

Two numerical indices reflect the variability of scores in a distribution but do not suffer the limitations of the range. They are the variance and its square root, the standard deviation.

The Variance, s^2 An index that reflects the degree of variability in a group of scores but which does not have the limitations of the range is the variance:

The variance, symbolized by s^2, is defined to be

$$s^2 = \frac{\Sigma(X_i - \bar{X})^2}{N - 1}$$

Notice the numerator of this formula, $\Sigma(X_i - \bar{X})^2$. This quantity is the sum of the squared deviations of the scores about their mean, or, as the phrase is

[1]Technically, the range should probably be defined as the difference between the upper real limit of the largest score minus the lower real limit of the smallest score. Since the range is an approximate index of variability at best, it does not seem appropriate to insist upon this level of precision.

often shortened, the **sum of squares.**[2] (This term is also used in Chapters 10 and 13 on the analysis of variance.) Except that the denominator is $N - 1$ instead of just N (which is explained under "Good Estimates" near the end of the chapter), the variance is essentially the "average sum of squared deviations of scores about their mean," or "the average sum of squares."

In the example below, the variance for distribution $A = (5, 7, 9)$ is computed using the definitional formula. Notice that the mean is computed first ($\bar{X} = 7.0$). Then the mean is subtracted from each score ($X_i - \bar{X}$) and this difference is squared [$(X_i - \bar{X})^2$]. The sum of the squared deviations about the mean is divided by $N - 1$ to obtain the variance. Thus, the variance of distribution A is 4.00. Distribution B consists of (3, 7, 11). It has greater variability, and its variance is 16.

X_i	$\bar{X}$	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
5	7	−2	4
7	7	0	0
9	7	+2	4
$\Sigma X_i = 21$		0	$\Sigma(X_i - \bar{X})^2 = 8$
$N = 3$			
$\bar{X} = 7$			

$$s^2 = \frac{\Sigma(X_i - \bar{X})^2}{N - 1}$$

$$= \frac{8}{2}$$

$$s^2 = 4.00$$

The Standard Deviation, s The variance measures variability in squared units. If a researcher recorded how long it took animals to find their way to a goal box at the end of a maze, the mean time would be in seconds but the variance would be in "squared seconds." This results from the fact that the

[2] Other texts often use a special notation for this quantity. Specifically, the deviation of a single score from its mean is called a *deviation score* and it is symbolized by a lowercase italicized letter corresponding to the letter used for that variable, such as x for X. The squared deviation score would be symbolized by x^2, and the sum of squared deviation scores—or simply the "sum of squared deviations"—would be represented by Σx^2. The sum of squared deviations, then, is also abbreviated to the phrase *sum of squares*, the same concept as described above. In short:

$$x = (X_i - \bar{X})$$

$$x^2 = (X_i - \bar{X})^2$$

and

$$\Sigma x^2 = \Sigma(X_i - \bar{X})^2 = \text{sum of squared deviations}$$

$$\Sigma x^2 = \text{sum of squares}$$

formula for the mean uses the scores as they are (ΣX_i) but the formula for the variance squares the deviations [$\Sigma(X_i - \bar{X})^2$]. However, it is also useful to have a measure of variability in terms of the original units of measurement, not squared units.

The standard deviation, symbolized by s, is defined to be the positive square root of the variance:

$$s = \sqrt{s^2}$$

or

$$s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N - 1}}$$

Since the variance is in squared units, taking its square root accomplishes a return to the original units of measurement.

Just as some other texts use M rather than $\bar{X}$ to symbolize the mean, they sometimes use the symbol SD rather than s for the standard deviation.

Computational Formulas for s^2 and s Formulas like those given above are called *definitional* formulas because they define and reflect the logic behind the concepts they express. However, definitional formulas are frequently inconvenient to use in making calculations, especially when dealing with large amounts of data. An expression for a statistic that is mathematically equivalent to the definitional formula but that is more convenient for calculating is called a *computational* formula.

The computational formula for the variance is

$$s^2 = \frac{N\Sigma X_i^2 - (\Sigma X_i)^2}{N(N - 1)}$$

The computational formula for the standard deviation is

$$s = \sqrt{\frac{N\Sigma X_i^2 - (\Sigma X_i)^2}{N(N - 1)}} \quad \text{or} \quad s = \sqrt{s^2}$$

OPTIONAL TABLE

3–3 PROOF THAT THE DEFINITIONAL AND COMPUTATIONAL FORMULAS FOR THE VARIANCE ARE EQUIVALENT

Operation	Explanation
To prove $\dfrac{\Sigma(X_i - \bar{X})^2}{N-1} = \dfrac{N\Sigma X_i^2 - (\Sigma X_i)^2}{N(N-1)}$	
1. $s^2 = \dfrac{\Sigma(X_i - \bar{X})^2}{N-1}$	1. Definition.
2. $= \dfrac{\Sigma(X_i^2 - 2X_i\bar{X} + \bar{X}^2)}{N-1}$	2. Binomial expansion of the form: $(a-b)^2 = a^2 - 2ab + b^2$
3. $= \dfrac{\Sigma X_i^2 - \Sigma 2X_i\bar{X} + \Sigma\bar{X}^2}{N-1}$	3. The sum of several terms is the sum of the separate terms: $\Sigma(X + Y + Z) = \Sigma X + \Sigma Y + \Sigma Z$
4. $= \dfrac{\Sigma X_i^2 - 2\bar{X}\Sigma X_i + \Sigma\bar{X}^2}{N-1}$	4. The sum of a constant times a variable equals the constant times the sum of the variable ($\bar{X}$ is a constant): $\Sigma kX = k\Sigma X$
5. $= \dfrac{\Sigma X_i^2 - 2\bar{X}N\bar{X} + N\bar{X}^2}{N-1}$	5. The sum of a constant taken N times

6.	$= \dfrac{\Sigma X_i^2 - 2\left(\dfrac{\Sigma X_i}{N}\right)\Sigma X_i + N\left(\dfrac{\Sigma X_i}{N}\right)\left(\dfrac{\Sigma X_i}{N}\right)}{N-1}$	is N times the constant: $\sum_{i=1}^{N} \bar{X} = N\bar{X}$ 6. Substitution: $\bar{X} = \dfrac{\Sigma X_i}{N}$
7	$= \dfrac{\Sigma X_i^2 - 2\left(\dfrac{\Sigma X_i}{N}\right)\Sigma X_i + (\Sigma X_i)\left(\dfrac{\Sigma X_i}{N}\right)}{N-1}$	7. Cancellation in the third term.
8	$= \dfrac{\Sigma X_i^2 - \left(\dfrac{\Sigma X_i}{N}\right)\Sigma X_i}{N-1}$	8. Subtraction involving the last two terms of the numerator.
9	$= \dfrac{N\Sigma X_i^2 - N\left(\dfrac{\Sigma X_i}{N}\right)\Sigma X_i}{N(N-1)}$	9. Multiplication of numerator and denominator by N.
10.	$s^2 = \dfrac{N\Sigma X_i^2 - (\Sigma X_i)^2}{N(N-1)}$	10. Cancellation.

It is important to realize that these computational formulas yield results that are equivalent (within rounding error) to those calculated by the definitional formulas. Optional Table 3–3 demonstrates the algebraic equivalence of the two formulas for the variance.

The computational formula has the advantages of requiring only one division, not requiring that the mean be calculated first, and being easy to compute on a standard hand calculator. Only three quantities are needed: ΣX_i, ΣX_i^2, and N. Consider the variance and standard deviation of the following distribution:

X_i	X_i^2
3	9
4	16
7	49
8	64
8	64
9	81
10	100
$\Sigma X_i = 49$	$\Sigma X_i^2 = 383$
$N = 7$	

Variance	Standard Deviation
$s^2 = \dfrac{N\Sigma X_i^2 - (\Sigma X_i)^2}{N(N-1)}$	$s = \sqrt{s^2}$
$= \dfrac{7(383) - (49)^2}{7(7-1)} = \dfrac{2681 - 2401}{42}$	$= \sqrt{6.67}$
$= \dfrac{280}{42}$	$s = 2.58$
$s^2 = 6.67$	

It is very important to distinguish between two quantities used in the computational formulas: ΣX_i^2 and $(\Sigma X_i)^2$. Recall that the first, ΣX_i^2, represents the sum of all the squared scores—*first* square each score, *then* add. The second, $(\Sigma X_i)^2$, represents the square of the sum of the scores—*first* add all the scores, *then* square this sum. Confusion between these two operations is often the source of computational error.

Properties of s^2 and s as Measures of Variability The variance is difficult to explain because it cannot be diagrammed or pointed at. Rather, the variance is an abstract numerical index that increases with the amount of variability in the group of scores. But despite its abstractness, the variance does have a number of properties that make it (and its square root, the

standard deviation) an appropriate measure of variability:

1. Since the mean is the central value of the distribution, it seems natural to base a measure of variability upon the extent to which the scores deviate from their mean—that is, on $\Sigma(X_i - \bar{X})^2$. In addition, recall from the discussion of the mean that the sum of the squared deviations about the mean is less than about any other value. This fact adds to the logic of selecting squared deviations about the mean (as opposed to some other value) as an index of variability.
2. Squared numbers are always positive because squaring a negative number results in a positive value. Therefore s^2 and s are always positive values. If the deviations were not squared, the negative deviations would cancel out the positive and their sum would be zero because $\Sigma(X_i - \bar{X}) = 0$.
3. Large deviations, when squared, contribute disproportionately to the total. A deviation of 4 units becomes 16 when squared, but a deviation of twice that size, that is of 8 units, contributes 64 to the total sum of squared deviations. Thus, the variance is especially sensitive to departures from the mean because the deviations, when squared, become disproportionately large.
4. The variance, which is approximately the average squared deviation of the scores from their mean, is proportional to the average squared deviation of each score from every other score. The concept of variability refers, in its most general sense, to the extent to which the scores differ from one another. One might express such variability by calculating the difference between each score and every other score, squaring those differences (to make them all positive), adding those squared deviations, and computing their average by dividing the sum by the number of such pairs of scores (that is, there will be $N(N - 1)/2$ such pairs). This average squared deviation of the scores from one another is proportional to the variance, which is approximately the average squared deviation of each score from its mean. Therefore, the variance is a good measure of the extent to which scores in a distribution deviate from one another.
5. As the variability of the scores increases, the statistical variance also increases. This can be seen in the few examples listed below:

Scores	s^2
10, 10, 10	0
8, 10, 12	4
6, 10, 14	16
4, 10, 16	36
2, 10, 18	64

As the scores show more and more variability, the value of s^2 increases, reflecting the extent to which the scores deviate from the mean and from one another. Similar arguments can be made for the standard deviation.

6. If there is no variability among the scores, that is, if all the scores in the distribution are the same value, the variance and standard deviation are both zero. This is so because if all scores are the same value (for example, 5, 5, 5, 5, 5), the mean will also be that value (that is, 5). Then all quantities $(X_i - \bar{X})^2$ and their sum will be zero because X_i and $\bar{X}$ will always be identical. Therefore, when there is no variability among the scores of a distribution (if all scores have the same value), $s^2 = 0$.
7. Under certain conditions the variance can be partitioned and its portions attributed to different sources. This capability of being partitioned permits statisticians to ask questions such as, What portion of the variability in a group of scores can be attributed to cause A as opposed to cause B? This aspect of the variance is taken up in greater detail in Chapters 10 and 13.

POPULATIONS AND SAMPLES

Frequently, a scientist performs an experiment on a relatively small group of subjects. At the conclusion of the research, however, the results are generalized to a much larger group of subjects. For example, in Chapter 1 a hypothetical experiment was described in which 40 children with attention-deficit disorder with hyperactivity (ADD/H) were randomly assigned to one of two groups, one that received drug medication and one that was given a placebo that had no real effect on the disorder. While only 40 children were actually studied, the results were intended to be generalized to all children with ADD/H. The 40 children are said to constitute a *sample* from the *population* of all ADD/H children.

A population is a collection of subjects, events, or scores that have some common characteristic of interest.[3]

A sample is a subgroup of a population.

It is important to note that *sample* and *population* are relative terms. All students enrolled at State University might be the population from which a sample of 100 students is drawn for a given experiment, or all students at State University might function instead as a sample of the larger population of all college students.

[3]A population is sometimes considered to be composed of an infinite number of cases. As such, the population is a theoretical concept because it can never actually be observed or assessed. The sample, in contrast, is an empirical concept because it can be observed and assessed. The concept of theoretically infinite populations is crucial to mathematical statistics, but it is of less use to students at this level of study.

One obvious reason for using samples rather than populations in research is that populations are usually too large to be studied efficiently. In addition, research results that must be limited to the specific subjects studied are less interesting and less useful than those applicable to a much larger group. Therefore, the scientist designs the experiment so that generalizations from the sample to the population may be made.

Parameters and Statistics

Frequently, the results that are generalized from sample to population are the statistical quantities, such as the mean and variance, that are calculated on the sample of scores. That is, the mean and variance computed on the sample are used as **estimators** of the mean and variance in the population. Therefore, it is necessary to be able to distinguish between statistical quantities associated with a sample and those associated with the population.

A quantitative characteristic of a sample is called a statistic and is symbolized with a Roman letter.

A quantitative characteristic of a population is called a parameter and is symbolized with a Greek letter.

Earlier in this chapter, indices of central tendency and variability were presented. These were calculated on samples of subjects, so the mean and standard deviation, for example, were statistics symbolized by the Roman letters, $\bar{X}$ and s, respectively. Most of the data that researchers analyze are from samples rather than populations. That is why this course is called statistics instead of parameters, and why this book presents all quantitative characteristics of distributions as statistics and not as parameters.

Almost all data sets are considered samples because scientists use statistics to estimate population parameters. For example, if the English department at State University wanted to estimate the ability of all freshmen at State to guide them in creating an appropriate English curriculum, they might sample 100 freshmen and test them on basic English skills. If they found the average score of the sample to be 82 and the standard deviation to be 7, they could regard these sample statistics as estimates of their respective population parameters—that is, of the mean and standard deviation for all freshmen at State. Of course, sometimes it is possible to assess all the members of a population. One could, for example, test all incoming freshmen at State, rather than just a sample of 100. But even then, the results might be used to estimate the abilities of the *next* freshman class, which would make this year's freshman class a sample of a larger population.

In research practice, then, it is rare to measure an entire population. Instead, sample statistics are typically used to estimate population parameters, so it is necessary to know what the population parameters corresponding to each sample statistic are called and how they are symbolized. This information is summarized in Table 3–4 for the mean, variance, and standard deviation. Of course, all other statistics—including the median, mode and range—also have corresponding parameters symbolized with Greek letters, but we will not have occasion to use them in this text.

"Good" Estimators

Since a main purpose of calculating statistics is to use them to estimate population parameters, statisticians are concerned that a statistic be a "good" estimator of its corresponding population parameter. The criteria that make an estimator "good" are discussed in Chapter 7 but we have already seen the consequence of this concern in the formula for the sample variance. Recall that

$$s^2 = \frac{\Sigma(X_i - \bar{X})^2}{N - 1}$$

But notice in Table 3–4 that the formula for the variance of a population is

$$\sigma^2 = \frac{\Sigma(X_i - \bar{X})^2}{N}$$

In particular, the sum of the squared deviations is divided by N in the formula for the *population parameter* σ^2 but it is divided by $N - 1$ in the formula for

3–4 SUMMARY OF THE NAMES, SYMBOLS, AND FORMULAS FOR COMMON STATISTICS AND PARAMETERS

	Sample Statistic			Population Parameter		
Quantity	**Symbol**	**Read As**	**Formula**	**Symbol**	**Read As**	**Formula[1]**
Mean	$\bar{X}$	"X bar"	$\frac{\Sigma X}{N}$	μ	"mew"	$\frac{\Sigma X}{N}$
Variance	s^2	"s squared"	$\frac{\Sigma(X - \bar{X})^2}{N - 1}$	σ^2	"sigma squared"	$\frac{\Sigma(X - \mu)^2}{N}$
Standard Deviation	s	"s"	$\sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}}$	σ	"sigma"	$\sqrt{\frac{\Sigma(X - \mu)^2}{N}}$

[1] When populations are of uncountable size or are theoretical, other mathematical procedures are used to define these quantities.

the *sample statistic* s^2. It turns out that dividing by $N - 1$ makes s^2 a "better" estimator of σ^2 than it would be if the sum of squared deviations were divided by N. In contrast, notice in Table 3–4 that the formula for the sample mean is the same as the formula for the population mean. Therefore, for some statistics the formulas are the same as for their corresponding parameters, whereas for others the formulas are slightly different than for their corresponding parameters. This difference often consists of dividing by $N - 1$ (or even by $N - 2$) instead of just N, and this is done to make the statistic a "better" estimator of its corresponding parameter (the criteria for making a statistic a "better" estimator are presented in Chapter 7). Because we almost never have an entire population available, only the formulas for statistics are presented in the remainder of this text.

A NOTE ON CALCULATORS AND COMPUTERS

A great variety of hand calculators and computer programs are now available to perform the calculations needed in this text. Some of them simply require the student to enter the numbers in a distribution, for example, and push one or two buttons or enter a simple instruction and the machine will automatically create one or more different frequency or cumulative frequency distributions; calculate the mean, median, mode, range, variance, and standard deviation; and perhaps even provide additional statistics, such as numerical indices for skewness and kurtosis.

Unfortunately, each machine is different, so it is impossible to write a textbook keyed to all of these machines or programs. Therefore, you must read the directions for the particular machine or program you have available and learn how to use it with this text. You will, however, likely run into some problems—your machine or program may not do things the same way as described in the text.

For one thing, the symbols might be different. Whereas $\bar{X}$ is used to symbolize the mean in the text, M might be used in the manual or on the key pad of your machine. Also, whereas s^2 and s are used here for the variance and standard deviation, *Var* and *SD* might be used by your machine.

Second, you may not always get the same answer as in the text or even as a classmate who uses another machine or program. There may be several reasons for this. Perhaps the rules for rounding numbers are not the same. For example, numbers ending in 5 may always be rounded up (or down), regardless of the odd or even nature of the previous digit. Or perhaps the same number of digits is not used during the calculations, causing one answer to be slightly (and sometimes substantially) different from another.

Third, and more serious, perhaps the formulas being used are different. Does the calculator or program use the formula for the statistic or for the parameter? Often it is not obvious which is being used, so consult the manual for the calculator or computer program to determine whether the formula

being used to calculate the variance, for example, is that for the sample variance (with a denominator of $N - 1$) or that for the population variance (with a denominator of N). This issue also pertains to the standard deviation, of course, and a few other statistics discussed later, but not to the mean, which is calculated by the same formula in either case. If you cannot locate the manual, test out the button or program on the computational examples in the text to determine which formula is being used. Do not rely on the manual or the label on the calculator button to follow the tradition of using Roman letters for statistics and Greek letters for parameters. Just because the symbol s^2 is used, it is no guarantee that the formula for the sample variance is used.

Finally, while calculators and computer programs speed calculations and —with the exceptions noted above—produce more accurate answers, they do not teach you much about statistics or the statistical quantities they calculate. You will learn more if you work at least some of the problems "by hand." This is why the numbers in the examples and in the problems are kept simple and do not generally require elaborate calculations. Use the calculator or computer primarily to check your answers to the first few problems and to do more complicated calculations.

SUMMARY

Two important characteristics of distributions are central tendency and variability. The mean, or average, is the most common index of central tendency, partly because the sum of the deviations about the mean equals zero and the sum of the squared deviations of each score about the mean is less than about any other value. The median (the point that divides the distribution into two equal parts) and the mode (the most frequently occurring score) are also used as measures of central tendency. The median is often preferred when the distribution is skewed, while the mode is useful when a distribution is skewed or bimodal. The range, variance, and standard deviation are measures of variability. The variance and standard deviation are frequently used in other statistical formulas.

A population is a collection of subjects, events, or scores that have some common characteristic of interest, and a sample is a subgroup of a population. Quantitative characteristics of a population are called parameters, whereas such characteristics of a sample are called statistics. Statistics are often used to estimate parameters.

FORMULAS

1. **Mean**

$$\bar{X} = \frac{\Sigma X_i}{N}$$

The population mean is symbolized by μ.

2. **Median**
   a. **No duplication near the median, odd number of scores:** M_d is the middle score.
   b. **No duplication near the median, even number of scores:** M_d is the average of the two middle scores.
   c. **Duplication of scores near the median:**

$$M_d = L + \left[\frac{N/2 - n_b}{n_w}\right] i$$

where L = lower real limit of the score interval containing the median
N = number of scores in the distribution
n_b = number of scores falling below the lower real limit of the interval containing the median
n_w = number of cases within the score interval containing the median
i = the size of the score interval ($i = 1$ if the data are in whole numbers)

3. **Mode**
   The mode, M_0, is the most frequently occurring score.
4. **Range**
   The range is estimated by taking the largest minus the smallest score.
5. **Variance**

$$s^2 = \frac{\Sigma(X_i - \bar{X})^2}{N - 1} \qquad \text{(definitional)}$$

$$s^2 = \frac{N\Sigma X_i^2 - (\Sigma X_i)^2}{N(N - 1)} \qquad \text{(computational)}$$

The population variance is symbolized by σ^2.

6. **Standard deviation**

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N - 1}} \qquad \text{(definitional)}$$

$$s = \sqrt{\frac{N\Sigma X_i^2 - (\Sigma X_i)^2}{N(N - 1)}} \qquad \text{(computational)}$$

The population standard deviation is symbolized by σ.

EXERCISES

1. Compute the mean, median, and mode for each of the distributions below.

A	B	C	D
1	1	2	2
2	1	2	2
2	1	2	4
4	3	3	4
7	4	4	4
8	5	5	5
11	6	8	7
	7	9	7
		10	

2. Show that the sum of the deviations about the mean of distribution C in problem 1 is zero. Compute the sum of the squared deviations about the mean and median of distribution C. Which is less, the sum of the squared deviations about the mean or the sum of the squared deviations about the median?
3. Indicate which measure of central tendency would be preferred for each of the following distributions, and explain why.
   a. family incomes in the United States
   b. heights of seniors in public high school
   c. IQ scores of third graders in a public school
4. Find the median for the following distributions:
   a. 2, 5, 6, 8, 9
   b. 0, 2, 3, 6, 8, 10
   c. 1, 2, 3, 3, 4, 5
   d. 5, 6, 6, 6, 6, 20, 21
   e. 0, 2, 3, 3, 7,
   f. 3.1, 3.2, 3.3, 3.4, 3.6 (note that the score interval is .1)
   g. 3.1, 3.2, 3.3, 3.3, 3.3, 3.8
   h. 3.1, 3.2, 3.3, 3.3, 3.3, 3.8, 3.9
5. Draw the curves for distributions in which
   a. the mean, median, and mode are identical.
   b. the mean and median are identical but the mode is different.
   c. the mean is greater than the median.
   d. the median is greater than the mean.
6. Compose the score values for three distributions which have the same mean and median but which differ in their amount of variability.
7. Calculate the variance and the standard deviation for each distribution in problem 1.
8. Discuss the limitations of the range as a measure of variability and present some numerical examples to illustrate your point.
9. Compute with both the definitional and computational formulas the variance and standard deviation for each of the following distributions. Also compare the means of the distributions.
   a. 2, 4, 4, 6
   b. 0, 2, 4, 10
   c. $-4, -3, -2, 0, 0, 5, 8, 8, 12, 16$
10. Discuss the properties, characteristics, and advantages of s as a measure of variability. Does s have any potential advantages over s^2?
11. Why does the formula for the variance have $N - 1$ in the denominator and not N?