2

Describing Data and Distributions

This chapter has three goals. The first goal is to introduce you to the more common summary measures used to describe data. As we explore those measures, we'll key in on two important concepts: central tendency and variability

- *Chapter Summary*
- *Some Other Things You Should Know*
- *Key Terms*
- *Chapter Problems*

This chapter has three goals. The first goal is to introduce you to the more common summary measures used to describe data. As we explore those measures, we'll key in on two important concepts: central tendency and variability or dispersion. The second goal follows from the first—namely, to get you comfortable with some of the symbols and formulas used to describe data. The third goal is a little more far-reaching: getting you to visualize different types of data distributions. The process of data visualization is something that you'll want to call upon throughout your journey. We'll start with some material that should be fairly familiar to you.

Before We Begin

Imagine the following scenario: Let's say that you're reading a report about health care in the United States. As the report unfolds, it reads like a general narrative—outlining the historical changes in leading causes of death, summarizing the general upward trend in the cost of health care, and so forth and so on. You tell yourself that you're doing fine—so far, so good. But before you know it, you're awash in a sea of terms and numbers. Some are terms that you've heard before, but you've never been really comfortable with them. Others are totally new to you. You get the idea of what the report is dealing with, but all the terms and numbers are just too much.

For someone else, it might be a report about crime (e.g., types of crime, length of sentence, characteristics of offenders, etc.), and packed with terms that are unfamiliar. And, just to consider another example, the scenario might involve a report on voter participation, with an emphasis on the last two presidential election cycles.

With any of those topics, it's easy to imagine the scenario. The report begins with a well-crafted narrative, but eventually it turns into a far more quantitative exposé on the subject at hand. What started out as a high level of reading comprehension on your part gives way to a sea of confusion. All too often, it's the reader's lack of solid grounding in basic statistical analysis that makes the report unintelligible.

It is against that background that the next chapter unfolds. You're going to be introduced to quite a few terms. Some of the terms may be very familiar to you, but others will likely take you into new territory. Allow me to throw in a cautionary note at the outset. If some of the terms or concepts are familiar to you, count yourself lucky. On the other hand, don't suspend your concentration on what you're reading. There's likely to be some new material to digest. Accordingly, let me urge you to take whatever time is necessary to develop a thorough understanding of the various concepts. In many ways, they represent essential building blocks in the field of statistics.

Measures of Central Tendency

To a statistician, the mean (or more correctly, the arithmetic mean) is only one of several measures of **central tendency.** The purpose behind any measure of central tendency is to get an idea about the *center,* or typicality, of a distribution. As it turns out, though, the idea of the center of a distribution and what that really reflects depends on several factors. That's why statisticians have several measures of central tendency.

The Mean

The one measure of central tendency that you're probably most familiar with is the one I mentioned earlier—namely, the mean. The **mean** is calculated by

adding all the scores in a distribution and dividing the sum by the number of scores. If you've ever calculated your test average in a class (based on a number of test scores over the semester), you've calculated the mean. I doubt there is anything new to you about this, so let's move along without a lot of commentary.

Now let's have a look at the symbols that make up the formula for the mean. Remember: All that's involved is summing all the scores (or values) and then dividing the total by the number of scores (or values). In terms of statistical symbols, the mean is calculated as follows:

$$\text{Mean} = \frac{\sum X}{N}$$

In this formula, there are only three symbols to consider. The symbol Σ (the Greek uppercase sigma) represents *summation* or *addition*. Whenever you encounter the symbol Σ, expect that summation or addition is involved. As for the symbol *X*, it simply represents the individual scores or values. If you had five test scores, there would be five *X* values in the distribution. Each one is an individual score (something statisticians often refer to as a *raw score*). The *N* in the formula represents the number of test scores (cases or raw scores) that you're considering. We use the lowercase *n* to represent the number of cases in a sample; the uppercase *N* represents the number of cases in a population. If, for example, you were summing five test scores (and treating the five cases as a population), you would say that *N* equals five. Consider the examples in Table 2-1.

As you've no doubt discovered when you have calculated the mean of your test scores in a class, the value of the mean doesn't have to be a value that actually appears in the distribution. For example, let's say you've taken three tests

Table 2-1 Calculation of the Mean

Scores/Values ($N = 5$)	Scores/Values ($N = 7$)	Scores/Values ($N = 10$)
1	2	5
2	4	1
3	6	3
4	7	4
5	8	1
	9	4
$\Sigma X = 15$	13	3
$15/5 = 3$		5
Mean = 3	$\Sigma X = 49$	2
	$49/7 = 7$	2
	Mean = 7	
		$\Sigma X = 30$
		$30/10 = 3$
		Mean = 3

Table 2-2 Calculation of the Mean

Scores/Values (N = 3)	Scores/Values (N = 6)
80	1
84	2
86	3
	4
$\Sigma X = 250$	5
$250/3 = 83.33$	6
Mean = 83.33	
	$\Sigma X = 21$
	$21/6 = 3.50$
	Mean = 3.50

and your scores were 80, 84, and 86. The mean would be 83.33—clearly a value that doesn't appear in the distribution. Similar examples are shown in Table 2-2.

By the same token, consider three incomes: $32,000; $41,500; and $27,200. The mean income would be $33,566.67—a value that isn't found in the distribution.

LEARNING CHECK

Question: What is the mean, and how is it calculated?

Answer: The mean is a measure of central tendency. It is calculated by adding all the scores in a distribution and dividing the sum by the number of cases in the distribution.

Now let's give some thought to what we've been looking at. The formula, at least the way I presented it to you, tells you how to calculate the mean. Now the question is, which mean are we really considering? Since the goal of inferential statistics is to use information from a sample to make statements about a population, it's essential to make it clear when you're referring to the mean of a sample and when you're referring to the mean of a population. Therefore, it shouldn't surprise you to learn that statisticians use different symbols to refer to the mean—one for a sample mean, and the other for a population mean. Just as there's a difference in the way we express the number of cases for a sample (n), as opposed to a population (N), we make a distinction between the mean of a sample and the mean of a population. Here's the difference:

$\overline{X}$ is the symbol for the mean of a sample (and n = number of cases)
μ is the symbol for the mean of a population (and N = number of cases)

So, the symbol for the mean of a sample is $\overline{X}$, and the symbol for the mean of the population is represented by μ. The symbol μ stands for *mu* (the Greek letter, pronounced "mew"). Technically, the term *mean* is used in reference to a sample, and **mu (μ)** is used in reference to a population. It's certainly OK to speak of the population mean, but you should always keep in mind that you are really speaking about mu. The formula essentially is the same for either the mean or mu, so you may be inclined to think this is a minor point—the fact that statisticians have different symbols for the sample mean and the population mean. Later on, you'll develop an appreciation for why the symbols are different. For the moment, just accept the notion that the distinction is an important one—something that you should take to heart. As a matter of fact, it's always a good idea to be clear in your thinking and speech when it comes to statistics. Use expressions such as *sample mean, population mean,* or *mu.* Unless you're making reference to the mean in general, don't just think or speak in terms of a mean without making it clear which mean you have in mind.

LEARNING CHECK

Question: What is the symbol for the mean of a sample? What is the symbol for the mean of a population? What is another term for the mean of the population?

Answer: The symbol for the mean of a sample is $\overline{X}$; the mean of the population, which is also referred to as mu, is μ.

Let me make one last point about the mean—whether you're talking about a population mean (μ) or a sample mean ($\overline{X}$). One of the properties of the mean is that it is sensitive to extreme scores. In other words, the calculated value of the mean is very much affected by the presence of extreme scores in the distribution. This is something you already know, particularly if you've ever been in a situation in which just one horribly low test score wrecked your overall average.

Imagine, for example, that you have test scores of 80, 90, 80, and 90. So far, so good; everything seems to be going your way. But what if you took a final test, and your score turned out to be 10? You don't even have to calculate the mean to know what a score like that would do to your average. It would pull your average down, and that's just a straightforward way of saying that the mean is sensitive to extreme scores. The 10 would be an extreme score, and the mean would be pulled down accordingly. You shouldn't have to do the calculations; you should be able to feel the effect in your gut, so to speak. If you did take the time to calculate the mean under the two different scenarios, you'd see that it moved from a value of 85 (when you were basing it on the first four tests) to a value of 70 (when you added in the fifth test score of 10). The presence of that one extreme score (the score of 10) reduced the mean by 15 points (see Table 2-3)!

Table 2-3 Effect of an Extreme Score

Test Scores ($N = 4$)	Test Scores ($N = 5$)
80	80
90	90
80	80
90	90
	10
$\Sigma X = 340$	
$340/4 = 85$	$\Sigma X = 350$
Mean = 85	$350/5 = 70$
	Mean = 70

LEARNING CHECK

Question: What does it mean to say that the mean is sensitive to extreme values?

Answer: The mean is sensitive to extreme values in the sense that an extremely high or extremely low score or value in a distribution will pull the value of the mean toward the extreme value.

The Median

Now we turn our attention to a second measure of central tendency—one referred to as the *median*. Unlike the mean, the median is not sensitive to extreme scores. In the simplest of terms, the **median** is the point in a distribution that divides the distribution into halves. It's sometimes said to be the midpoint of a distribution. In other words, one half of the scores in a distribution are going to be equal to or greater than the median, and one half of the scores are going to be equal to or less than the median. Like the mean, the median doesn't have to be a value that actually appears in the distribution.

As I introduce you to the formula for the median, let me emphasize one point. It is a positional formula; that is, it points you to the *position* of the median. Again, the formula yields the position of the median—not the value. Here's the formula for the position of the median:

$$\text{Median} = \frac{N + 1}{2}$$

Note the use of *N*, indicating the number of cases in a population. If we were determining the median for a sample, we would use *n* to represent the number of cases.

Before you apply the formula, there's one thing you should always remember: You have to arrange all the scores in your distribution in ascending or descending *order*. That's a must—otherwise, the formula won't work.

Table 2-4 Calculating the Median

- 13 scores
- Arrange the scores in ascending or descending order
- Formula for the Position of the Median = $\frac{N+1}{2}$

$$\frac{N+1}{2} = \frac{13+1}{2} = \frac{14}{2} = 7\text{th Score}$$

- SCORES

1, 1, 2, 4, 9, 11, 12, 21, 21, 24, 25, 25, 30

↑

7th Score = Position of the Median

Value of the Median = 12

Assuming you've arranged all the scores in ascending or descending order, (see Table 2-4), all you have to know is how many scores you have in the distribution. That's what the *N* in the formula is all about; it's the number of cases, scores, or observations. If there are 13 scores, the formula directs you to add 1 to 13 and then divide by 2. The result would be 14 divided by 2, or 7. The median would be the 7th score—that is, the score in the 7th *position*. Once again, the median would not have the value of 7. Rather, it would be the value of whatever score was in the 7th position (from either the top or the bottom of the distribution). The *value* of the median—of the score in the 7th position—is 12.

The nice thing about the formula for the position of the median is that it will work whether you have an odd or an even number of cases in the distribution. When you have an even number of cases, the formula will direct you to a position that falls halfway between the two middle cases. For example, consider a distribution with the following scores: 1, 2, 3, 12, 20, 24. With 6 scores in the distribution, the formula gives us (6 + 1)/2. The median, then, would be the 3.5th score. The halfway point between the third and fourth scores is found by calculating the mean of the two values: (3 + 12)/2 = 15/2 = 7.5. In other words, the *position* of the median would be the 3.5th score; the *value* would be 7.5. All of this should become more apparent when you look at the examples in Table 2-5.

The other nice thing about the formula for the position of the median is that it works for distributions with a small number or a large number of values. For example, in a distribution with 315 scores, the position of the median would be the 158th score (315 + 1)/2 = 158. In a distribution with 86,204 scores, the position of the median would be the 43,102.5th score (86,204 + 1)/2.

Table 2-5 Locating the Median

Scores/Values		Scores/Values	
1		1	
2		2	
4	← Median = 4	4	← Median = 4
8		8	
12		120	
Scores/Values		**Scores/Values**	
3		10	
5	← Median = 6	15	← Median = 20
7		25	
9		80	
Scores/Values		**Scores/Values**	
10		17	
10		27	
14		34	
23	← Median = 23	34	← Median = 34
23		34	
80		59	
100		62	

Once again, the formula determines the position of the median—not the value of the median. Also, the formula rests on the assumption that the scores in the distribution are in ascending or descending order.

LEARNING CHECK

Question: What is the median, and how is it determined?

Answer: The median is a measure of central tendency; it is the score that cuts a distribution in half. The formula locates the position of the median in a distribution, provided the scores in distribution have been arranged in ascending or descending order.

The Mode

In addition to the mean and the median, there's another measure of central tendency to consider—namely, the mode. The **mode** is generally thought of as the score, value, or response that appears most frequently in a distribution. For example, a distribution containing the values 2, 3, 6, 1, 3, and 7 would produce a mode of 3. The value of 3 appears more frequently than any other value.

A distribution containing the values 2, 3, 6, 1, 3, 7, and 7 would be referred to as a **bimodal distribution** because it has two modes—3 and 7. Both values (3 and 7) appear an equal number of times, and both appear more frequently than any of the other values. A distribution with a single mode is called a **unimodal distribution.** A distribution in which each value appears the same number of times has no mode. Table 2-6 provides a few more examples to illustrate what the mode is all about.

LEARNING CHECK

Question: What is the mode?
Answer: The mode is a measure of central tendency. It's the score or response that appears most frequently in a distribution.

As it turns out, there are some situations in which the mode is the only measure of central tendency that's available. Consider the case of a nominal level variable—for example, political party identification. Imagine that you've collected data from a sample of 100 voters, and they turn out to be distributed as follows: 50 Republicans, 40 Democrats, and 10 Independents. You couldn't calculate a mean or median in a situation like this, but you could report the modal response. In this example, the modal response would be Republican, because that was the

Table 2-6 Identifying the Mode

Scores/Values	Scores/Values
1 2 2 7 9 9 9 21 — 9 is the mode	1 2 2 7 7 7 9 9 9 25 29 — Bimodal modes = 7 and 9
Scores/Values	**Scores/Values**
200 200 200 305 309 318 — 200 is the mode	20 22 28 30 36 38 — No mode

most frequent response. If nothing else, this provides a good example of a point I made earlier about levels of measurement: Which measure of central tendency you use is often a function of the level of measurement that's involved.

In the long run, of course, the most widely used measure of central tendency is the mean, at least in inferential statistics. So, let's return to a brief discussion of the mean as a jumping-off point for our next discussion.

Assume for the moment that you're teaching two classes—Class A and Class B. Further assume that both classes took identical tests and both classes had mean test scores of 70. At first glance, you might be inclined to think the test performances were identical. They were, in terms of the mean scores. But does that indicate the classes really performed the same way? What if the scores in Class A ranged from 68 to 72, but the scores in Class B ranged from 40 to 100? You could hardly say the overall performances were equal, could you? And that brings us to our next topic—variability.

Measures of Variability or Dispersion

The last example carries an important message: If you really want to understand a distribution, you have to look beyond the mean. Indeed, two distributions can share the same mean, but can be very different in terms of the variability of individual scores. In one distribution, the scores may be widely dispersed or spread out (for example, ranging from 40 to 100); in another distribution, the scores may be narrowly dispersed or compact (for example, ranging from 68 to 72). Statisticians are routinely interested in this matter of **dispersion** or **variability**—the extent to which scores are spread out in a distribution.

Statisticians have several measures at their disposal when they want to make statements about the dispersion or variability of scores in a distribution. Even though a couple of the measures aren't of great utility in statistical analysis, you should follow along as we explore each one individually. By paying attention to each one, you're apt to get a better understanding of the big picture.

LEARNING CHECK

Question: What is meant by dispersion?

Answer: Dispersion is another term for variability. It is an expression of the extent to which the scores are spread out in a distribution.

The Range

One of the least sophisticated measures of variability is the **range**—a statement of the lowest score and the highest score in a distribution. For example, a statement that the temperature on a particular day ranged from 65 degrees to

78 degrees would be a statement of the range of a distribution. You could also make a statement that the income distribution of your data ranged from $12,473 to $52,881. To report the range is to report a summary measure of a distribution. Consider the following examples of range:

Test Scores	23–98
Incomes	$15,236–$76,302
Aggression Levels	1.36–7.67
Temperature	62°–81°

The range tells you something about a distribution, but it doesn't tell you much. To have more information, you'd need a more sophisticated measure. We'll eventually explore some of the other measures, but first let's spend a little time on a central concept—the general notion of variability, or deviations from the mean.

LEARNING CHECK

Question: What is the range?

Answer: The range is a measure of dispersion. It is a simple statement of the highest and lowest scores in a distribution.

Deviations From the Mean

Researchers are often interested in questions that have to do with variability. For example, a researcher might want to know why test scores vary, why incomes vary, why attitudes vary, and so forth. In some cases, they want to know whether or not two or more variables vary together—for example, a researcher might want to know if test scores and income levels vary together or not. Before you can even begin to answer a question like that, you first have to understand the concept of variability. To do that, you have to begin with an understanding of the notion of *deviations from the mean*.

The idea of deviation from the mean is fairly basic. It has to do with how far an individual or raw score in a distribution deviates from the mean of the distribution. To calculate the deviation of an individual score from the mean, simply subtract the mean of the distribution from the individual score. When you do this, you're determining how far a given score is from the mean.

For example, imagine a distribution with five values—a distribution with the following income data: $27,000; $32,000; $82,000; $44,000; and $52,000. As it turns out, the mean of that distribution would be $47,400. In terms of deviations from the mean, there will be five of them. An income of $27,000 deviates a certain amount from the mean of $47,400 and so does $32,000. The same is true for the values of $82,000, $44,000, and $52,000.

Table 2-7 Deviations from the Mean

Scores/Values	Deviations	
(N = 5) (X)	(X – Mean)	
1	1 – 3	–2
2	2 – 3	–1
3	3 – 3	0
4	4 – 3	+1
5	5 – 3	+2
		0
Mean = 3		

DEVIATIONS FROM THE MEAN

Note: The same results will occur whether you subtract the mean from each raw score or you subtract each raw score from the mean.

Each value deviates from the mean. To better understand all of this, take a look at the example shown in Table 2-7.

Regardless of how many scores there are in a distribution, there will be a deviation of each score from the mean. Consider the illustrations in Table 2-7. Focus on the relationship between each individual raw score and the mean, and how that translates into the concept of a *deviation* of each score from the mean.

Whether the individual scores in a distribution are widely dispersed or tightly clustered around the mean, the sum of the deviations from the mean will always equal 0 (subject to minor effects due to rounding). This point is important enough that it deserves an illustration. Consider a really simple distribution like the one shown in Table 2-8. Chances are that you can simply look at the first distribution and determine that the mean is equal to 3.

Assuming you've convinced yourself that the mean is equal to 3, take a close look at the distribution. Begin with the score of 1. The score of 1 deviates from the mean by –2 points (1 – 3 = –2). In other words, the score of 1 is 2 points *below* the mean (hence the negative sign). The score of 2 is –1 points from the mean (2 – 3 = –1). The score of 3 has a deviation of 0, because it equals the mean (3 – 3 = 0). Then the pattern reverses as you move to the scores that are above the mean. The score of 4 is 1 point *above* the mean (4 – 3 = 1), and the score of 5 is 2 points above the mean (5 – 3 = 2). If you were to sum all the deviations from the mean, they would equal 0. The sum of the deviations from the mean will always equal 0, because that is how the mean is mathematically defined.

As you learned earlier, the mean doesn't have to be a score that actually appears in a distribution, and the same notion applies in this instance as well.

Table 2-8 Sum of Deviations from the Mean = 0

Scores/Values	Deviations	
(X)	(X – Mean)	
1	1 – 3	–2
2	2 – 3	–1
3	3 – 3	0
4	4 – 3	+1
5	5 – 3	+2
		0
Mean = 3		

Sum of the Deviations Equals 0

Scores/Values	Deviations	
(X)	(X – Mean)	
80	80 – 90	–10
85	85 – 90	–5
90	90 – 90	0
95	95 – 90	+5
100	100 – 90	+10
		0
Mean = 90		

Sum of the Deviations Equals 0

Consider the two examples in Table 2-9. In each case, the calculated value of the mean doesn't really appear in the distribution, but the sum of the deviations from the mean still equals 0.

The principle that the sum of the deviations equals 0 holds so steadfastly that you can assure yourself of one thing: If you ever add the deviations from the mean and the total doesn't equal 0, you've made a mistake somewhere along the way. You've either calculated the deviations incorrectly, or you've calculated the mean incorrectly. As mentioned before, the only exception would be a case in which a value other than 0 resulted because of rounding procedures.

LEARNING CHECK

Question: What are the deviations from the mean? What does the sum of the deviations from the mean always equal?

Answer: The deviations from the mean are the values obtained if the mean is subtracted from each score in a distribution. The sum of the deviations from the mean will always equal 0.

Table 2-9 Sum of Deviations from the Mean = 0

Scores/Values (X)	Deviations (X – Mean)	
2	2 – 5	–3
4	4 – 5	–1
6	6 – 5	+1
8	8 – 5	+3
		0
Mean = 5		

Sum of the Deviations Equals 0 (even when the mean is a value that doesn't appear in the original distribution)

Scores/Values (X)	Deviations (X – Mean)	
28	28 – 33	–5
30	30 – 33	–3
32	32 – 33	–1
34	34 – 33	+1
36	36 – 33	+3
38	38 – 33	+5
		0
Mean = 33		

Sum of the Deviations Equals 0 (even when the mean is a value that doesn't appear in the original distribution)

Assuming our goal is to get a summary measure that produces an overall picture of the deviation from or about the mean, we're obviously facing a bit of a problem. If we don't take some sort of corrective action, so to speak, we'll always end up with the same sum of deviations (a value of 0), regardless of the underlying distribution—and that tells us nothing.

The Mean Deviation

One way out of the problem would be simply to ignore the positive and negative signs we get when calculating the difference between individual scores and the mean. Indeed, that's what the *mean deviation* is all about. Before introducing you to the formula, however, let me explain the logic. I suspect it will strike you as very straightforward and remarkably similar to the calculation of the mean.

To calculate the mean deviation, here's what you do:

1. Determine the mean of the distribution.
2. Find the difference between each raw score and the mean; these are the deviations.

3. Ignore the positive or negative signs of the deviations; treat them all as though they were positive. This means you are considering only the absolute values.
4. Calculate the sum of the deviations (that is, the absolute values of the deviations).
5. Divide the sum by the number of cases or scores in the distribution.

The result is the **mean deviation.** The result gives you a nice statement of the average deviation. Indeed, the measure is sometimes referred to as the **average deviation.** The mean deviation (or average deviation) will tell you, on average, how far each score deviates from the mean. Here's the formula for the mean deviation for a set of sample scores:

$$\text{Mean Deviation} = \frac{\sum |X - \bar{X}|}{n}$$

Remember: The bars indicate that you are to take the absolute values; ignore positive and negative signs.

To understand how similar the mean deviation formula is to the formula for the mean, just give it a close look and think about what the formula instructs you to do. It tells you to sum something and then divide by the number of cases (the same thing that the formula for the mean instructs you to do). In the case of the formula for the mean deviation, what you are summing are absolute deviations from the mean. Take a look at the illustration in Table 2-10.

The mean deviation would be a wonderfully useful measure, were it not for one important consideration. It's based on absolute values, and absolute values are difficult to manipulate in more complex formulas. For that reason, statisticians turn elsewhere when they want a summary characteristic of the variability of a distribution. One of their choices is to look at the *variance* of a distribution.

Table 2-10 Calculation of the Mean Deviation

Scores/Values	Deviations		Absolute Values
(N = 5)			
(X)	(X – Mean)		
1	1 – 3	–2	2
2	2 – 3	–1	1
3	3 – 3	0	0
4	4 – 3	+1	1
5	5 – 3	+2	2
		0	Σ = 6
Mean = 3			

Step 1 Calculate mean.
Step 2 Calculate deviations from the mean.
Step 3 Convert deviations to absolute values.
Step 4 Sum the absolute values.
Step 5 Divide the sum by the number of cases. 6/5 = 1.20; Mean Deviation = 1.20

LEARNING CHECK

Question: What is the mean or average deviation? How does it get around the problem that the sum of the deviations from the mean always equals 0? What is its major drawback?

Answer: The mean or average deviation is a measure of dispersion. It solves the problem by using absolute values (ignoring the positive and negative signs) of the deviations from the mean. The use of the absolute values, however, makes it difficult to use in more complex mathematical operations. As a result, it is rarely used.

The Variance

Variance, as a statistical measure, attacks the problem of deviations' summing to 0 in a head-on fashion. As you know from basic math, one way to get rid of a mix of positive and negative numbers in a distribution is to square all the numbers. The result will always be a string of positive numbers. It's from that point that the calculation of distribution's variance begins. As before, we'll start with the logic.

Think back to the original goal. The idea is to get some notion of the overall variability in the scores in a distribution. We already know what to expect if we look at the extent to which individual scores deviate from the mean. We could calculate all the deviations, but they would sum to 0. If we squared the deviations, though, we would eliminate the sum-to-zero problem. Once we squared all the deviations, we could then divide by the number of cases, and we'd have a measure of the extent to which the scores vary about the mean. And that's what the **variance** is. It's the result you'd get if you calculated all the deviations from a mean, squared the deviations, summed the squared deviations, and divided by the number of cases in the distribution.

That sounds like something that's rather complicated, but it really isn't, provided you take on the problem in a step-by-step fashion. Let's consider a fairly simple distribution (see Table 2-11) and have a look at the calculation of the variance both mathematically and conceptually. Here's the step-by-step approach that we'll use:

1. Calculate each deviation and square it. Remember that you're squaring the deviations because the sum of the deviations would equal 0 if you didn't.
2. Sum all the squared deviations.
3. Divide the sum of the squared deviations by the number of cases.

Applying this approach to the scores shown in Table 2-11, you can move through the process step by step.

Table 2-11 Calculation of the Variance of a Population

Scores/Values	Deviations		Squared Deviations
(N = 5)			
(X)	(X – Mean)		
1	1 – 3	–2	4
2	2 – 3	–1	1
3	3 – 3	0	0
4	4 – 3	+1	1
5	5 – 3	+2	4
		0	Σ = 10
Mean = 3			

Sum of the squared deviations equals 10

N = 5 (treating the 5 scores as a population)

10/5 = 2

Variance = 2

Table 2-12 Calculating the Variance of a Population (showing how values explode when they are squared)

Scores/Values	Deviations		Squared Deviations
(N = 6)			
(X)	(X – Mean)		
$21,800	21,800 – 41,300	–19,500	380,250,000
$35,600	35,600 – 41,300	–5,700	32,490,000
$52,150	52,150 – 41,300	+10,850	117,722,500
$64,250	64,250 – 41,300	+22,950	526,702,500
$32,000	32,000 – 41,300	–9,300	86,490,000
$42,000	42,000 – 41,300	+700	490,000
			Σ = 1,144,145,000
Mean = $41,300			
			1,144,145,000/6
			Variance = $190,690,833.33

I assure you the same approach will work whether your distribution has small values (for example, from 1 to 10) or much larger values (for example, a distribution of incomes in the thousands of dollars). The example in Table 2-12 illustrates that the same approach works just the same when you're dealing with larger values.

To develop a solid understanding of what the variance tells us, consider the four distributions shown in Table 2-13. In the top two distributions, the variances are the same, but the means are very different. In the bottom two distributions, the means are equal, but the variances are very different.

By now you should be developing some appreciation for the concept of variance, particularly in terms of how it can be used to compare one distribution to another. But there's still one problem with the variance as a statistical measure.

Table 2-13 Comparison of Distributions: Equal Variances and Different Means, Equal Means and Different Variances

Scores	Deviations (X – Mean)		Squared Deviations	
1	1 – 3	–2	4	
2	2 – 3	–1	1	
3	3 – 3	0	0	
4	4 – 3	+1	1	
5	5 – 3	+2	4	
			10	10/5 = 2
Mean = 3				Variance = 2

Scores	Deviations (X – Mean)		Squared Deviations	
51	51 – 53	–2	4	
52	52 – 53	–1	1	
53	53 – 53	0	0	
54	54 – 53	+1	1	
55	55 – 53	+2	4	
			10	10/5 = 2
Mean = 53				Variance = 2

Scores	Deviations (X – Mean)		Squared Deviations	
30	30 – 50	–20	400	
40	40 – 50	–10	100	
50	50 – 50	0	0	
60	60 – 50	+10	100	
70	70 – 50	+20	400	
			1000	1000/5 = 200
Mean = 50				Variance = 200

Scores	Deviations (X – Mean)		Squared Deviations	
46	46 – 50	–4	16	
48	48 – 50	–2	4	
50	50 – 50	0	0	
52	52 – 50	+2	4	
54	54 – 50	+4	16	
			40	40/5 = 8
Mean = 50				Variance = 8

The act of squaring the deviations has a way of markedly changing the magnitude of the numbers we're dealing with—something that happens anytime you square a number.

The illustration you encountered in Table 2-12 is a good example. That illustration was based on a distribution of income data, and income data necessarily involve some fairly large numbers. Inspection of that illustration reveals how quickly the original values explode in magnitude when deviations are squared.

In truth, all values have a way of exploding in magnitude when they're squared. Whether you're dealing with single-digit numbers or values in the thousands, the same process is at work. The mere act of squaring numbers can radically alter the values. In the process, you're apt to lose sight of the original scale of measurement you were working with.

Fortunately, there is a fairly easy way to bring everything back in line, so to speak. All you have to do is calculate the variance and then turn right around and take the square root. Indeed, statisticians make a habit of doing just that. Moreover, they have a specific name for the result. It is referred to as the **standard deviation.**

LEARNING CHECK

Question: What is the variance? How does it deal with the problem that the sum of the deviations from the mean always equals 0? What is a major limitation of the variance?

Answer: The variance is a measure of dispersion. To avoid the problem of the deviations from the mean always summing to 0, the variance is based on squaring the deviations before they are summed. A major limitation of the variance is that squaring the deviations inflates the magnitude of the values in the distribution, which can cause you to lose sight of the original units of measurement.

The Standard Deviation

Before we get to the business of calculating the standard deviation, let me point out an important distinction. When we're referring to the standard deviation of a sample, we use the symbol *s*. When we're referring to the standard deviation of a population, however, we use the symbol σ (the lowercase symbol for the Greek letter *sigma*). Let me underscore that again. Here's the difference:

s is the standard deviation of a sample.

σ is the standard deviation of a population.

As I mentioned previously, the whole idea behind the standard deviation is to bring the squared deviations back in line, so to speak, and we do that by taking the square root of variance. In other words, the standard deviation is the square root of the variance. Looked at the other way around, the variance is simply the standard deviation squared.

Variance = Standard Deviation Squared

Square Root of the Variance = Standard Deviation

Variance is what is under the radical (square root symbol) before you take the square root when calculating the standard deviation.

LEARNING CHECK

Question: What is the relationship between the standard deviation and the variance?

Answer: The standard deviation is the square root of the variance. The variance is the standard deviation squared.

You may have noticed that I didn't use any sort of symbol to refer to the variance when I first introduced you to the concept. As a matter of fact, I didn't even give you any sort of formula for the variance. I simply explained it to you—telling you that variance, as a measure of dispersion, is nothing more than the sum of the squared deviations from the mean, divided by the number of cases.

I avoided the use of any formula or symbols for variance for a specific reason. It had to do with how the standard deviation and variance are related to each other. As you now know, the standard deviation is simply the square root of the variance. By the same token, the standard deviation squared is equal to the variance.

Recall for a moment how we symbolize the standard deviation: s = the standard deviation of a sample, and σ = the standard deviation of a population. Since the variance is equal to the standard deviation squared, we symbolize the variance as follows:

s^2 is the variance of a sample.

σ^2 is the variance of a population.

No doubt it would have caused great confusion had I used those symbols (s^2 and σ^2) when I first introduced you to the concept of variance. Had I given you a formula for the variance, my guess is that you would have expected to see a symbol such as V—certainly not s^2 or σ^2. Recall that we had not yet mentioned

the standard deviation (*s* and σ). Just to make certain that you now understand the link between the two—standard deviation and variance—let's summarize:

s = Standard Deviation of a Sample
σ = Standard Deviation of a Population

s^2 = Variance of a Sample
σ^2 = Variance of a Population

LEARNING CHECK

Question: What are the symbols for the standard deviation and the variance of a sample? What are the symbols for the standard deviation and variance of a population?

Answer: For a sample, the symbol for the standard deviation is *s*, and the symbol for the variance is s^2. For a population, the symbol for the standard deviation is σ, and the symbol for variance is σ^2.

Presumably you now know what the symbols *s* and σ refer to—the standard deviation of a sample and a population, respectively—so we can get back to our discussion. As I mentioned before, the standard deviation is a particularly useful measure of dispersion because it has the effect of bringing squared values back into line, so to speak. You'll see the standard deviation often in the field of statistics, so you'll want to become very familiar with the concept. To help you along, consider a simple example.

Let's assume that you want to calculate the standard deviation of some data for a small class (let's say five students). Assume that you're looking at the number of times each student has been absent throughout the semester. Since you're only interested in the results for this class, the class constitutes a population. In this example, then, you're calculating the standard deviation of a population (or σ). You'll want to start by having a close look at the formula. Then you'll want to follow the example through in a step-by-step fashion.

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

There's no reason to let the formula throw you. It's really just a statement that tells you to calculate the variance and then take the square root of your

answer. Even if you forget everything you already know about the variance, you should be able to go through the formula step by step. Think of it this way:

1. Forget the radical or square root sign for a moment, or simply think of it as a correction factor. You have to square some numbers (to get rid of the negative signs), so you're eventually going to turn around and take the square root.
2. Look at each deviation (the difference between the mean and each raw score) and square it. Once again, you're squaring the deviations because the sum of the deviations would equal 0 if you didn't.
3. Sum all the squared deviations.
4. Divide the sum of the squared deviations by the number of cases.
5. Take the square root to get back to the original scale of measurement.

Most of those steps should be familiar—after all, most of them are the same steps you used in calculating the variance.

Table 2-14 shows you the step-by-step calculations. Remember that we're calculating the standard deviation of a population. It may be very small populations (only five cases), but we're treating it as a population nonetheless. Later on, we'll deal with the standard deviation for samples.

Now let's give some thought to what the standard deviation tells us. Like the variance, the standard deviation gives us an idea of the dispersion of a distribution. It gives us an idea as to how far, in general, individual scores deviate from the mean. It gives us an overall notion as to the variability in the distribution. Moreover, it does so in a way that is free of the problems associated with the variance. Remember: The big problem with the variance is that values are magnified as a result of the squaring process.

So, what does the standard deviation really tell you about a distribution? Suppose you were told that the standard deviation for a distribution has a value of 15.5. This value of 15.5 may mean 15.5 dollars or 15.5 pounds or 15.5 test points, depending on what variable you're looking at and the nature

Table 2-14 Calculating the Standard Deviation of a Population

Scores/Values	Deviations		Squared Deviations	
(N = 5)				
(X)	(X – Mean)			
5	5 – 12	–7	49	Sum of Squared Deviations = 106
10	10 – 12	–2	4	
12	12 – 12	0	0	106/5 = 21.2 (5 is Number of Cases)
14	14 – 12	+2	4	
19	19 – 12	+7	49	Square Root of 21.2 = 4.60
			106	
Mean = 12				**Standard Deviation = 4.60**

of the data you've collected. But still it's reasonable to ask: So what? So what does the standard deviation (or variance, for that matter) really tell us? From my perspective, there are at least three answers to that question.

First, you can think of the standard deviation as a measure that tells you (sort of) how far scores or values (in general) deviate from the mean. In short, the standard deviation tracks along with the overall variability in a distribution. When there is more variability in a distribution, the standard deviation increases.

It's that last point—the notion that the value of the standard deviation increases when there is more variability in a distribution—that leads to a second interpretation or interpretative guideline regarding the standard deviation. I would also add that I believe that it's the best way to think of the standard deviation, at least at this point in your education. Simply put, you should think of the standard deviation as a relative or comparative sort of measure. In other words, it's probably best to think in terms of one standard deviation compared to another. For example, you might want to compare the standard deviation of incomes in two cities or you might want to compare the standard deviation of test scores in two classes. When you think of the standard deviation (or the variance, for that matter) as a relative or comparative measure, you begin to view it as a measure that may be very useful in situations involving more than one distribution. For example, your ultimate concern may boil down to which of the two or three or four distributions (let's say which of several sets of test scores) has the largest amount of variability. In a case like that, the standard deviation is likely to be a very useful measure.

Finally, as a third answer to the question of what the standard deviation tells us, I would tell you that it is a critical element in understanding the concept of a normal distribution or normal curve. I don't expect you to get the connection right now, particularly since there's an entire chapter devoted to the topic of the normal curve, and you've yet to encounter that chapter. For the moment, let me simply encourage you to do whatever is necessary to understand where the standard deviation fits in relationship to the variance (it is the square root of the variance—remember?). Let me also urge you to get a firm foundation in how to calculate the standard deviation. You will eventually discover that the notion of the standard deviation is a central concept.

Returning to the formula for the standard deviation (and the variance, for that matter), I should point out that different texts may present the formula in a different format—something that's quite common in the world of statistics. Sometimes it's a matter of personal preference; sometimes it's an effort to provide a formula that is more calculator-friendly. For example, consider the following two formulas for the calculation of the standard deviation for a population:

Formula for σ in This Text	**A Common Alternative Formula (more suited for use with a calculator)**
$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$	$\sigma = \sqrt{\frac{\sum X^2}{N} - \mu^2}$

My preference for the formula presented in this text is tied to what I call its *intuitive appeal:* It strikes me as more closely representing what's apt to be going on in your mind as you think through what the concept means.

So much for abstract examples and discussions of formulas. I suspect you're getting anxious to see some direct application of all of this, so let's head in that direction. Imagine for a moment that you were in a class of 200 students taking four 100 points tests—a Math Test, a Verbal Test, a Science Test, and a Logic Test. Then assume that you received the following information about the tests: your score, the class average for each test, and the standard deviation for each test. Suppose the information came to you in a form like this:

Test	Mean	Standard Deviation	Your Score
Math	82	6	80
Verbal	75	3	75
Science	60	5	70
Logic	70	7	77

Just so you'll get in the habit of keeping matters straight in your mind, the example we're dealing with involves four populations of test scores. In each case, you have a population mean or mu (μ) and a population standard deviation (σ). Now here are the questions:

What was your best performance, relative to your classmates?

What was your worst performance? Why?

Let me suggest that you give the questions a little bit of thought before you arrive at the answers.

Assuming you've thought about it, you now have some answers in mind. But rather than just giving you the answers, let me walk you through the logic involved in deriving them.

A good place to begin is with a comparison of your individual test scores to the means. In the case of the Math Test, the mean was 82, and your score was 80. In other words, your score was actually below the mean (so that's not too good). In the case of the Verbal Test, you had a score of 75, but the mean was 75. You didn't score above or below the mean—an OK performance, but not really that great.

Now have a look at your performance on the Science Test. In that case, you had a score of 70, but the mean was 60. In other words, you scored 10 points above the mean—not bad! As a matter of fact, the standard deviation on that test was 5, so your 10 points above the mean really equates to a score that was two standard deviation units above the mean. Here's the reasoning: Each standard deviation equals 5 points; your score was 10 points above the mean; therefore, your score was two standard deviation units above the mean.

Now let's take a look at your performance on the Logic Test. The mean on that test was 70, and you had a score of 77. In other words, you scored

7 points above the mean. As it happens, the standard deviation on the test was 7 points, so your score was only one standard deviation unit above the mean.

If you want to know just how poorly you did on the Math Test, the same logic will apply. The mean on that test was 82, and your score of 80 was two points below that. Since the standard deviation on the Math Test was 6 points, your score was 2/6 or 1/3 of a standard deviation unit below the mean.

So now you have all the answers. First, your best performance (in a relative sense) was on the Science Test, even though that was your worst absolute score. Second, your worst performance turned out to be on the Math Test, even though that was your highest absolute score.

Just to demonstrate that point more completely, consider one final example. Assume for the moment that there was a fifth test thrown into the mix—let's say it's a Foreign Language Ability Test. But let's also say that unlike the other tests that were 100 point tests, let's say that the Foreign Language Ability Test was a 250 point test. In other words, scores on the Foreign Language Ability Test could range from 0 to 250. Let's also assume that the mean score on the Foreign Language Ability Test was 120 with a standard deviation of 15. Now what if your test score was a score 90—what sort of performance would that be?

If you use the same logic that you used in the other situations, you'd quickly discover that you have a new "worst performance." Your performance on the Foreign Language Ability Test equated to a score that was two standard deviations below the mean (i.e., you were 30 points below the mean; a standard deviation equals 15 points; therefore, you were two standard deviations below the mean). Before the Foreign Language Ability Test was thrown into the mix, your worst performance was on the Math Test (you were 1/3rd of a standard deviation below the mean). On the Foreign Language Ability Test, though, you were two standard deviations below the mean. Thus, your score on the Foreign Language Ability Test becomes your worst performance.

The point of the Foreign Language Ability Test example is to demonstrate something very important—it doesn't make any difference whether you're comparing tests that have the same underlying scale (e.g., test scores that can range from, let's say 0 to 100), or you're comparing all sorts of test scores—scores on a 100 point test, scores on a 250 point test, or scores on a 1500 point test, for that matter. The underlying goal is the same: Determine where a given score (in this case, your score) falls, in relationship to the mean, and express the difference in standard deviation units. To fully grasp this point, just remember what was going on in your mind as you worked through the questions—just think back to the calculations.

If you really think back to the calculations, you'll eventually arrive at a very important point—namely that all the mental calculations you went through amounted to calculating ratios. In each case, you calculated a ratio: the difference between an individual score and the mean of the distribution, expressed in terms of standard deviation units. This very important point will come up again later on, so let me urge you to take the time to really comprehend what it means to say you were calculating ratios. Once again, you were calculating a

ratio that reflected the difference between the individual score and the mean of the distribution, expressed in terms of standard deviation units.

LEARNING CHECK

Question: If you determine the difference between an individual score and the mean of a distribution, and then you divide the difference by the standard deviation of the distribution, what does the result tell you?

Answer: The answer is a statement of the distance the score is from the mean, expressed in standard deviation units.

Before we move on to the next chapter, I need to explain one last matter concerning the standard deviation—a matter I alluded to earlier. Since the standard deviation is so widely used in inferential statistics, and the business of inferential statistics involves moving from a sample to a population, it's time I introduced you to a slight difference between samples and populations when it comes to the standard deviation.

n Versus *n* – 1

We'll start with the notion that we generally deal with a sample in an effort to make some statement about a population (a point you encountered when we first discussed the idea of inferential statistics). It would be great if a sample standard deviation gave us a perfect reflection of the population standard deviation, but it doesn't. In fact, the accuracy of a sample standard deviation (as a reflection of the population standard deviation) is somewhat affected by the number of cases in the sample.

Here's the logic behind that last statement. Start by imagining a population distribution that has substantial variability in it—let's say the distribution of 23,000 students' ages at a large university. No doubt there would be some unusually young students in the population, just as there would be some unusually old students in the population. In other words, there would probably be a substantial amount of variability in the population. But if you selected a sample of students, there's a good chance that you wouldn't pick up all the variability that actually exists in the population. Most of your sample cases would likely come from the portion of the population that has most of the cases to begin with. In other words, it's unlikely that you'd get a lot of cases from the outer edges of the population.

If, for example, most of the students were between 20 and 25 years of age, most of the students in your sample would likely be within that age range. What you're not likely to get in your sample would be a lot of really young or really old students. You might get some, but probably not many. In other words, your sample probably wouldn't reflect all of the variability that really exists in the population. As a result, the standard deviation of your sample would likely be slightly less than the true standard deviation of the population.

Since the idea is to get a sample standard deviation that's an accurate reflection of the population standard deviation—one that can provide you with an unbiased estimate of the population's standard deviation—some adjustment is necessary. Remember: The idea in inferential statistics is to use sample statistics to estimate population parameters. If you're going to use a sample standard deviation to estimate the standard deviation of a population, you'll want a sample standard deviation that more closely reflects the true variability or spread of the population distribution.

Statisticians deal with this situation by using a small correction factor. When calculating the standard deviation of a sample (or the variance of a sample, for that matter), they change the n in the denominator to $n - 1$. This slight reduction in the denominator results in a larger standard deviation—one that better reflects the true standard deviation of the population.

LEARNING CHECK

Question: What is the effect of using $n - 1$, as opposed to n, in the formula for calculating the standard deviation of a sample?

Answer: The effect of using $n - 1$ (as opposed to n) in the denominator is to yield a slightly larger result—one that will be a better reflection of the population standard deviation.

To better understand the reason for making this change in the formula, think about the effect of sample size: The larger your sample is, the greater the likelihood that you've picked up all the variability that is really present in the population. Imagine that first you select a sample of, let's say, 30 students. Then you select another sample, but this time you include 50 students. Each time you increase the sample size—as you work up to a larger and larger sample—you get closer and closer to having a sample standard deviation that equals the population standard deviation. What would happen if you gradually increased your sample size until you were working with a sample that was actually the entire population? You'd have the actual population standard deviation in front of you.

When you use $n - 1$ in the denominator of the formula for the standard deviation of a sample, you not only slightly increase the final answer (or value of the standard deviation), you do so in a way that is sensitive to sample size. The smaller the size of the sample, the more of an impact the adjustment will make. For example, dividing something by 2 instead of 3 will have a much greater impact than dividing something by 999 instead of 1000. In other words, the adjustment factor wouldn't have a lot of impact if you were working with a really large sample, but it would have a major impact if you were working with a really small sample.

At this point, I should tell you that different statisticians have different approaches to the use of the correction factor ($n - 1$, as opposed to just n, in the denominator). Some statisticians quit correcting when a sample size is 30 or greater; that is, they use n when the sample size reaches 30. Others require

a larger sample size before they're willing to rely on n (as opposed to $n - 1$) in the denominator. The approach in this text is to always use $n - 1$ when calculating a sample standard deviation.

The issue at this point isn't when different statisticians invoke the correction factor and when they don't; the issue is why. Because the answer to that question is one that usually takes some serious thought, let me suggest that you take time out for one of those dark room moments I mentioned earlier.

First, take the time to give some serious thought to the ideas of variability and the standard deviation in general. Then take some time to think about how the standard deviation of a population is related to the standard deviation of a sample. Develop a mental picture of a population and a sample from that population. Mentally focus on why you would expect the standard deviation of the population to be slightly larger than the standard deviation of the sample. You should think about the relationship between the two long enough to fully appreciate why the correction factor is used. It all goes back to the point that the variability of a sample is going to be smaller than the variability of a population, and that's why a correction factor has to be used.

Finally, in an effort to make certain that you fully understand how to calculate the standard deviation of a sample, and the point about $n - 1$ in the denominator, let me suggest that you take a close look at Table 2-15. It's an illustration of the calculation of the standard deviation for a sample. My suggestion is that you repeat each of the calculations shown in the illustration, working each step on your own, while also paying particular attention to the next to the last step (i.e., dividing by $n - 1$ before you take the square root).

Assuming you feel comfortable about the different measures of central tendency and measures of variability (and the standard deviation, in particular), we

Table 2-15 Calculating the Standard Deviation of a Sample

Scores/Values	Deviations		Squared Deviations	
(N = 9)				
(X)	(X – Mean)			
7	(7 – 4)	3	9	Sum of Squared Deviations = 54
1	(1 – 4)	–3	9	
3	(3 – 4)	–1	1	54/8 = 6.75
5	(5 – 4)	1	1	Note that $n - 1$ or 8 is used
6	(6 – 4)	2	4	
2	(2 – 4)	–2	4	
8	(8 – 4)	4	16	
1	(1 – 4)	–3	9	
3	(3 – 4)	–1	1	Square Root of 6.75 = 2.598
			54	
Mean = 4				**Standard Deviation = 2.598 or round to 2.60**

can move forward. Next we turn our attention to the graphic representation of data distributions—the world of graphs and curves. That's where we'll go in the next chapter.

Chapter Summary

In learning about measures of central tendency and dispersion, you've learned some of the fundamentals of data description. Moreover, you've had a brief introduction to the business of statistical notation—why, for example, different symbols are used when referring to a sample, as opposed to a population. Ideally the connection to the previous chapter hasn't been lost in the process, and you've begun to understand that it's essential to make clear whether you're discussing a sample statistic or a population parameter.

As to what you've learned about measures of central tendency, you should have digested several points. First, several measures of central tendency are available, and each one has its strength and weakness. One measure might be appropriate in one instance but illsuited for another situation. Second, you've likely picked up on the importance of the mean as measure of central tendency—a measure that finds its way into a variety of statistical procedures. For example, the mean is an essential element in calculating both the variance and the standard deviation.

On the variability or dispersion side of the ledger, you have been introduced to several different measures. Working from the simplest to the more complex, you've learned that some measures have more utility than others. You've also learned how the variance and the standard deviation are related to each other, and (ideally) you've developed a solid understanding of why both measures are in the statistical toolbox.

Finally, you have learned that there's some room for judgment and personal preference in the matter of statistical analysis. For example, you've encountered different formulas for calculating the standard deviation—one that's ideally suited for use with a calculator, and another that better reflects the logic behind the procedure. You've also learned that different statisticians have different preferences when it comes to using n versus $n - 1$ in the denominator of the formula for the sample standard deviation. These are small matters, perhaps, but they help explain why different texts present different formulas for the same statistical procedure.

Some Other Things You Should Know

At this point, you deserve to know that data and data distributions can be presented in a variety of ways. Indeed, the art of data presentation is a field in itself. The data distributions we've considered so far have been presented as

ungrouped data, meaning that scores or values have been presented individually. If three 22s were present in a distribution, for example, each 22 was listed separately in the distribution. Frequently, however, statisticians find themselves working with grouped data—data presented in terms of intervals, or groups of values.

For example, a data distribution of income might be presented in terms of income intervals, showing how many people in a study had incomes between $25,000 and $29,999, how many had incomes between $30,000 and $39,999, and so on. As you might expect, statisticians have procedures to deal with such situations. An excellent treatment of the topic can be found in Moore (2000).

You should also be reminded of a point I made earlier in reference to the standard deviation (and variance, for that matter). Different formulas abound, not just for standard deviation or variance, but with respect to many other measures and procedures. It's not uncommon for two texts to approach the same topic in different ways. If a formula jumps out at you, and it's not quite the same as the presentation you've encountered here or somewhere else, don't be disheartened, threatened, or confused. Think conceptually. Think about the elements in the formula. Think about the formula in terms of its component parts, recognizing that there may be more than one way to approach some of those component parts. Sometimes the difference in presentation reflects the author's personal preference. Sometimes it's oriented toward a particular tool, such as a calculator. Whatever the reason, the fact that such differences exist is something you'll want to keep in mind, should you find yourself consulting different sources for one reason or another. The rule of thumb in this text is to focus on the formula or approach that seems to have the most intuitive appeal.

Key Terms

average deviation
bimodal distribution
central tendency
dispersion (variability)
mean
mean deviation
median
mode
mu (μ)
range
standard deviation
unimodal distribution
variance

Chapter Problems

Fill in the blanks, calculate the requested values, or otherwise supply the correct answer.

1. Three measures of central tendency are the _____, _____, and _____.

2. The measure of central tendency that is sensitive to extreme scores is the _____.
3. The most frequently represented score or response in a distribution is the _____.
4. The _____ is a measure of central tendency that represents the midpoint of a distribution.
5. The _____ is a measure of _____ that is based on a statement of the highest and lowest scores in a distribution.
6. A distribution has 14 scores. Each score is represented only once in the distribution, with two exceptions. The score of 78, appears three times, and the score of 82 appears four times. What is the mode of the distribution?
7. A distribution has 32 scores. Each score appears once, with the following exceptions: The score of 18 appears twice, and the score of 21 appears twice. How would you state the mode of the distribution?
8. The measure of dispersion that is based upon the absolute values of the deviations from the mean is the _____.
9. The sum of the deviations from the mean is always equal to _____.
10. Because the sum of the deviation from the mean always equals _____, the variance gets around the problem by _____ the deviations before they are summed.
11. The standard deviation is the _____ of the variance.
12. In order to obtain a more accurate reflection of the standard deviation of a population, the standard deviation for a sample can be calculated by using _____ in the denominator of the formula, as opposed to using _____ in the formula.

Application Questions/Problems

1. Consider the following data from a sample of five cases:

   7 6 3 1 4

   a. What is the mean?
   b. What is the position of the median?
   c. What is the value of the median?
   d. What is the mean or average deviation?
   e. What is the variance?
   f. What is the standard deviation?

2. Consider the following data from a sample of eight cases:

   20 21 18 16 12 15 12 13

   a. What is the mean?
   b. What is the position of the median?

8. The mean score for a verbal exam is 65, with a standard deviation of 4. You are told that your score is two standard deviations above the mean. What is your score?

9. The mean score for a mathematics exam is 125 (on a 200 point exam), with a standard deviation of 30. You are told that your score is 1.5 standard deviations above the mean. What is your score?