

CHAPTER 5

Measures of Variation

Key Concepts

Deviation from Mean
Range
Interquartile Range

Semi-Interquartile Range
Variance
Standard Deviation

Explaining Variance

Introduction

Measures of variation refer to a group of statistics that is intended to provide us with information on how a set of scores are distributed. An examination of measures of variation is a logical extension of any description of a data set using the measures of central tendency that we examined in the previous chapter. Consider a case where there are two sections of a course in statistics, and you are told that each section is taught by the same professor, each section has an enrollment of 15 students, and that the mean, and median score on a recent examination is 80 in both sections of the course. Without any additional information you would be tempted to conclude that the performance of the students in the two sections of the course is reasonably similar. As a matter of fact, all of the information up to this point would suggest that the performance of the students in the two sections is identical.

Now suppose that you are shown the actual performance of each student on the examination in both sections of the course (see Figure 5.1).

Clearly the performance of the students in the two sections is radically different. The score of 80 is not only the mean of section 1, but also a score that seems to be more representative of the performance of the entire class. While not all of the students scored 80, more were at that score

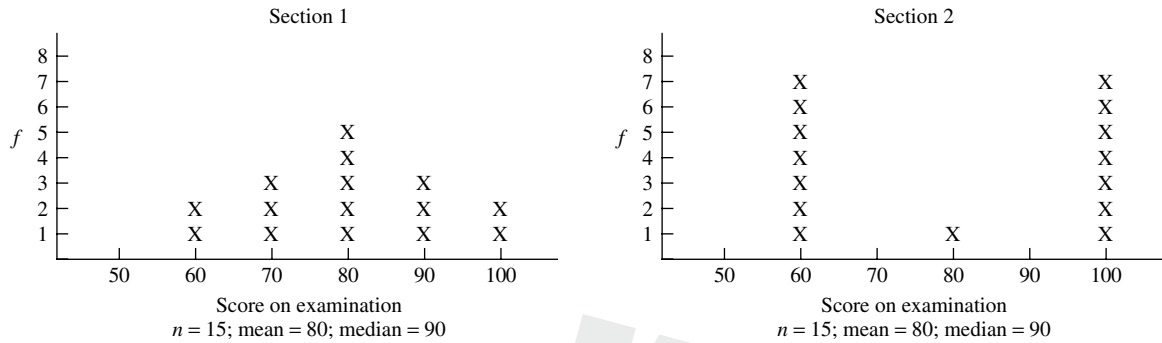


Figure 5.1 Graph showing distributions for the two sections

than any other, and the number of students scoring higher or lower than 80 falls off the further the scores deviate from 80. The performance of the students in section 2 is far different even though the distribution has the same mean and median score as the first distribution. In section 2 the mean of 80 is not at all representative of the typical performance. In fact, only one student earned a score at the mean. Seven students earned perfect scores of 100, while the remaining seven students earned a very low score of only 60.

Just as the mean is a single number that is designed to tell us where the central point of a distribution of scores is located, measures of variation are single numbers that are designed to tell us how the individual scores are distributed. By examining both a measure of central tendency such as the mean, and an appropriate measure of variation, we will be able to know not only where the central point of a distribution is located, but also if it tends to look more like the distribution of scores in section 1 or if it looks more like the distribution of scores in section 2.

The measures of variation examined in this chapter can be divided into two groups. The first group of statistics measures variation in a distribution in terms of the distance from the smaller scores to the higher scores. Included in this group of measures of variation is the range, which is a simple measure of the variation in a distribution computed by examining the distance from the smallest score to the largest score. Also included in this group of statistics are the interquartile range (IQR), and the semi-interquartile range (SIQR). These latter two measures of variation are often used in educational research. The second group of statistics measures variation in terms of a summary measure of each score's **deviation from the mean**. The two statistics of this type that we will examine are the variance, and the standard deviation. The measures of variation based on deviation from the mean tend to be more useful, and are fundamental concepts in behavioral science research.

The Range, Interquartile Range, and Semi-interquartile Range

Range

The simplest measure of variation in a distribution of data is the **range**. The range is defined as the distance from the smallest observed score to the largest observed score in a set of data. For raw data, the range may be computed by subtracting the lower limit of the smallest observed score from the upper limit of the largest observed score. Consider the set of raw data below:

X : 2 4 5 5 7 8 10 12 15 18 21

The range is $(21.5 - 1.5) = 20$.

A simple alternative is to subtract the smallest observed score from the largest observed score, and then add 1 additional unit of measurement to compensate for the upper and lower limits. In this case, the range is

$$(21 - 2) + 1 = 20$$

In either case, the set of data ranges over 20 units, just as if you began reading a book on page 2 and continued through page 21 you would have read a total of 20 pages.

Keep in mind when using the alternative method of subtracting the smallest observed score from the largest observed score that we add 1 *unit of measurement*, not just 1. For example, if we had the following data on income:

Income: 12,000 14,000 18,000 35,000 46,000 58,000

The range would be computed as:

$$(58,000 - 12,000) + 1,000 = 47,000$$

It would not be correct to compute the range as:

$$(58,000 - 12,000) + 1 = 46,001$$

The computation is similar for data organized in a simple or full frequency distribution as illustrated below.

X	f	Cf
12	7	20
10	3	13
8	3	10
7	5	7
6	2	2
$\Sigma f = 20$		

In a simple frequency distribution where the interval size $i = 1$, we do not lose any precision in measurement. In this case the range would be computed as

$$(12.5 - 5.5) = 7$$

or, alternatively:

$$(12 - 6) + 1 = 7.$$

In a frequency distribution where the interval size $i > 1$, we follow the same general procedure, but use the lower limit of the smallest interval, and the upper limit of the largest interval as our parameters for the computation of the range. For the data below:

X	f	Cf
25–29	9	75
20–24	16	66
15–19	20	50
10–14	16	30
5–9	10	14
0–4	4	4
$\Sigma f = 75$		

the range would be computed as: $[29.5 - (-0.5)] = 30$
 or, alternatively $(29 - 0) + 1 = 30$.

Interquartile range

The **interquartile range** (IQR) is defined as the distance from the 75th percentile to the 25th percentile in a set of data. In the previous chapter on central tendency we saw that a few extreme scores at one end of the distribution can bias a measure such as the mean. The same situation is true for a measure of variation such as the range. A few extreme scores at one or the other end of a distribution will affect the size of the range. The IQR is an alternative measure of variation that eliminates the effect of the extreme scores in a distribution by reporting the range between the 75th and 25th percentiles. In effect, the IQR represents the range of the middle 50% of the distribution, and ignores the top 25% and bottom 25% of the data that may be subject to extreme scores.

Computing the IQR is as simple as subtracting the 25th percentile from the 75th percentile:

$$\text{IQR} = P75 - P25$$

The formula for finding a particular percentile from Chapter 3 is provided below, along with the frequency distribution previously used to illustrate the procedure. To compute the IQR we will need to find both the 25th percentile and the 75th percentile.

X	f	Cf
25–29	20	125
20–24	22	105
15–19	28	83
10–14	20	55
5–9	25	35
0–4	10	10
	$\Sigma f = 125$	

The general formula for finding a given percentile is given by the equation below.

$$P_x = LL + \left[\left(\frac{F_p - CF_{\text{below}}}{f_{\text{int}}} \right) \times i \right]$$

where

- P_x = the desired percentile
- F_p = the number of frequencies below the desired percentile
- CF_{below} = the value from the cumulative frequency column for the interval just below F_p
- LL = the lower limit of the interval containing F_p
- f_{int} = the number of frequencies in the F_p interval
- i = the interval size.

For the 25th percentile:

$$F_p = (.25) \times (125) = 31.25$$

For the 75th percentile:

$$F_p = (.75) \times (125) = 93.75$$

The intervals containing the 25th and 75th percentiles are indicated below:

X	f	Cf	
25–29	20	125	
20–24	22	105	<--($F_p = 93.75$ is in this interval)
15–19	28	83	<--($CF_{\text{below}} = 83$)
10–14	20	55	
5–9	25	35	<--($F_p = 31.25$ is in this interval)
0–4	10	10	<--($CF_{\text{below}} = 10$)
	$\Sigma f = 125$		

To find the 25th percentile we simply substitute appropriate values into the formula as follows:

$$P_{25} = 4.5 + \left[\left(\frac{31.25 - 10}{25} \right) \times 5 \right]$$

$$P_{25} = \left[\left(4.5 + \frac{21.25}{25} \right) \times 5 \right]$$

$$P_{25} = 4.5 + (.85 \times 5)$$

$$P_{25} = 4.5 + 4.25 = 8.75$$

To find the 75th percentile we simply substitute appropriate values into the formula as follows:

$$P_{75} = \left[\left(19.5 + \frac{93.75 - 83}{22} \right) \times 5 \right]$$

$$P_{75} = 19.5 + \left[\left(\frac{10.75}{22} \right) \times 5 \right]$$

$$P_{75} = 19.5 + (.49 \times 5)$$

$$P_{75} = 19.5 + 2.45 = 21.95$$

Having computed both the 25th and the 75th percentiles, we may now compute the IQR:

$$IQR = 21.95 - 8.75 = 13.2$$

The resulting value of 13.2 for the IQR indicates that there is a range of 13.2 points for the middle 50% of the distribution. The advantage of the IQR over the simple range is that any bias that might result from a few extremely high scores or a few extremely low scores (or both) has been eliminated.

Semi-interquartile range

The final range based measure of variation presented in this chapter is the **semi-interquartile range**. The concept “interquartile range” suggests a measure of variation based on a quartile, or 25% of the distribution; however, the IQR actually represents the range of the middle 50% of the distribution. The SIQR is

an alternative to the IQR that comes closer to representing a “quartile or 25%” size range in the distribution. The SIQR is simply the IQR divided by 2.

$$\text{SIQR} = \frac{\text{IQR}}{2}$$

In the case of our previous example, the SIQR is

$$\text{SIQR} = \frac{13.2}{2} = 6.6$$

The IQR and the SIQR are widely used in education research where there always seems to be one or two students at each extreme of the distribution.

The Variance and the Standard Deviation

Variance

The **variance** and the **standard deviation** are two measures of variation that are based on the concept of deviation from the mean. For any distribution of scores measured on a continuous scale we can compute a mean, and then measure the distance of each score from the mean. For example, the set of 6 scores presented below have a mean equal to 8.

X
13
11
9
7
5
3

We may then define deviation from the mean (d_i) as the distance of each score from the mean, or

$$d_i = X - \bar{X}$$

Using our set of 6 scores, we may then calculate the deviation from the mean for each score.

X	$X - \bar{X} = d_i$
13	$(13 - 8) = 5$
11	$(11 - 8) = 3$
9	$(9 - 8) = 1$
7	$(7 - 8) = -1$
5	$(5 - 8) = -3$
3	$(3 - 8) = -5$

One way we might construct a summary measure of variation in a distribution of scores is to compute the average deviation of each score from the score’s mean. To do this, we would simple sum the individual deviations from the mean which we have just calculated, and then divide by the number of observations we have.

$$\frac{\Sigma di}{n} = \frac{5 + 3 + 1 + (-1) + (-3) + (-5)}{6}$$

$$\frac{\Sigma di}{n} = \frac{0}{6} = 0$$

It may seem quite logical to construct a measure of variance by calculating the average deviation from the mean for a set of scores, but there is one small problem. The sum of the deviations from the mean for all distributions is always the same thing—0.

$$\Sigma di = 0$$

One solution to this problem is to base our measure of variance on the *squared deviation* from the mean. By squaring the result of $(X - \bar{X})$, we will eliminate the negative numbers, and prevent the negative deviations and positive deviations from canceling each other out.

Applying this strategy to our original distribution will give us the following result:

X	$X - \bar{X}$	$(X - \bar{X})^2$
13	$(13 - 8) = 5$	$(5)^2 = 25$
11	$(11 - 8) = 3$	$(3)^2 = 9$
9	$(9 - 8) = 1$	$(1)^2 = 1$
7	$(7 - 8) = -1$	$(-1)^2 = 1$
5	$(5 - 8) = -3$	$(-3)^2 = 9$
3	$(3 - 8) = -5$	$(-5)^2 = 25$

Now if we want to construct a measure of variation that gives us a single number representing the average variation of each score in a distribution, we can use the mean (or average) of the squared deviations of each score from the distribution mean. We need only sum the squared deviations from the mean, and then divide by the number of observations.

$$\frac{\Sigma(X - \bar{X})^2}{n} = \frac{25 + 9 + 1 + 1 + 9 + 25}{6}$$

$$\frac{\Sigma(X - \bar{X})^2}{n} = \frac{70}{6} = 11.67$$

The resulting value indicates that the mean squared deviation of each score from the distribution mean is 11.67; or stated differently, on average, the distance squared of each score from the mean is 11.67 units.

The statistic we call the **variance** represents the mean squared deviation from the mean for a set of data. The logical formula for the variance is simply:

$$\text{Variance} = \frac{\Sigma(X - \bar{X})^2}{n}$$

where

- X = each value of X in the distribution
- \bar{X} = the mean of the distribution
- n = the sample size.

What does the variance really tell us?

Recall the situation from earlier in the chapter where we had two distributions representing the performance of two sections of a class on an exam with the same number of observations, the same mean, and the same median. Yet, we could see by simple inspection that the two distributions were very different. The variance of the distribution will tell us how representative the mean is of each of the scores in the distribution. The closer each individual score is to the mean the smaller the variance will be. If each score is at the mean in a distribution the variance will equal zero, indicating that there is no variation from the mean across the entire distribution. The farther each of the individual scores is from the mean the greater the variance will be, indicating that the mean is not as typical of the individual scores in the distribution.

Examine the three simple distributions below.

A	B	C
10	10	8
8	10	6
6	6	6
4	2	6
2	2	4

Each distribution contains 5 observations, and each distribution has a mean equal to 6. Yet the observations differ with respect to how much the individual scores vary from the mean. Since the variance represents the average squared deviation from the mean, which distribution would you expect to have the greatest variance? Which distribution should have the smallest variance?

The mean value of 6 seems to be most typical of the scores in distribution C, so it should have the smallest variance. The scores in distribution B are much farther from the mean value of 6, so it should have the largest variance. The scores in distribution A appear somewhat in between, and should have a variance between that of distribution B and distribution C.

Let's calculate the variance for each distribution below:

A	$(X - \bar{X})$	$(X - \bar{X})^2$	B	$(X - \bar{X})$	$(X - \bar{X})^2$	C	$(X - \bar{X})$	$(X - \bar{X})^2$
10	4	16	10	4	16	8	2	4
8	2	4	10	4	16	6	0	0
6	0	0	6	0	0	6	0	0
4	-2	4	2	-4	16	6	0	0
2	-4	16	2	-4	16	4	-2	4
	$\Sigma(X - \bar{X})^2 = 40$			$\Sigma(X - \bar{X})^2 = 64$			$\Sigma(X - \bar{X})^2 = 8$	
Variance:	$\frac{40}{5} = 8$			$\frac{64}{5} = 12.8$			$\frac{8}{5} = 1.6$	

As suspected, the variance in distribution C is the smallest. For distribution C the average squared deviation from the mean is 1.6 units. Distribution B has the largest variance with an average squared deviation of 12.8 units. Distribution A is in between the two with an average squared deviation from the mean of 8 units.

It is important to realize that the size of the variance does not have any special underlying standard interpretation. The value of the variance does not have a special meaning like your blood pressure, where

you know that you are in reasonably good condition with a systolic blood pressure of 120 and a diastolic pressure of 80. There is no normal or abnormal range for the variance. What the variance is telling you is simply what the squared distance is from the typical or average score to the mean. Variance, along with the mean, can then allow you to have an idea of what a particular distribution might look like, and will allow you to judge how well the mean serves as a measure of central tendency.

Consider the case of the three simple distributions we just examined. Typically, information for such distributions would not provide individual scores, but would be presented in summary form as follows:

Summary Statistics	Distribution		
	A	B	C
Sample Size:	5	5	5
Mean:	6	6	6
Variance:	8.0	12.8	1.6

Even though we do not have the individual scores available, we can reach some fairly accurate conclusions about what each of these distributions would look like. For example, we can see that all three distributions are of the same size ($n = 5$), and that all three have the same mean ($\bar{X} = 6$). The fact that the variance for distribution C is only 1.6 indicates that most of the individual scores in the distribution should be very close to the value of 6. After all, a distribution where every score is equal to the mean will have a variance of zero (remember that variance can never be negative). Similarly, we would assume that the individual scores in distribution B must be much more diverse or spread out around the mean since the variance is so much larger.

A computational formula for variance

Up to this point we have utilized a logical formula for the variance that is useful for demonstrating how the variance is computed, but requires us to go through some unnecessary steps. A computational formula may be used, which simplifies the calculations, especially when a larger data set is involved. The computational formula presented below may look a little more difficult at first, but with a little experience using it you will likely find it to be much easier.

$$\text{Variance} = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n}$$

where

- $\sum X^2$ = the sum of the X s squared
- $(\sum X)^2$ = the quantity, sum of X s squared
- n = the sample size.

Notice that the computational formula for variance contains both the sum of X s squared term, and the quantity sum of X s squared term.

The term in the numerator: $\sum X^2 - \frac{(\sum X)^2}{n}$

represents the sum of the squared deviations from the mean that was formally written as

$$\sum (X - \bar{X})^2$$

This term is also sometimes referred to as simply the sum of squares, and will play a role in several statistical procedures that will be examined in later chapters.

Let's demonstrate that the computational formula for the variance will provide the same results that we previously obtained for our three distributions.

A		B		C	
X	X^2	X	X^2	X	X^2
10	100	10	100	8	64
8	64	10	100	6	36
6	36	6	36	6	36
4	16	2	4	6	36
2	4	2	4	4	16
$\Sigma X = 30$	$\Sigma X^2 = 220$	$\Sigma X = 30$	$\Sigma X^2 = 244$	$\Sigma X = 30$	$\Sigma X^2 = 188$
$(\Sigma X)^2 = 900$		$(\Sigma X)^2 = 900$		$(\Sigma X)^2 = 900$	

Variance Computation		
Distribution A	Distribution B	Distribution C
$\frac{220 - \frac{900}{5}}{5}$	$\frac{244 - \frac{900}{5}}{5}$	$\frac{188 - \frac{900}{5}}{5}$
$\frac{220 - 180}{5}$	$\frac{244 - 180}{5}$	$\frac{188 - 180}{5}$
$\frac{40}{5}$	$\frac{64}{5}$	$\frac{8}{5}$
8	12.8	1.6

In each case we obtain the same result for the variance with the computational formula that we previously obtained with the logical formula. Remember that the variance will never be negative. If your calculation of the variance results in a negative number, you can be sure that you have made an error somewhere. Confusing the ΣX^2 with the $(\Sigma X)^2$ in the computational formula is a common mistake that will result in a negative number, as is neglecting to divide $(\Sigma X)^2$ by n before subtracting the result from ΣX^2 .

Some important terminology and symbols for variance

At this point we need to introduce some terminology, and appropriate symbols for the variance. There are three situations where we might want to calculate the variance. The logical formula and its computational alternative that we have been using are appropriate for two of the three situations.

Recall from Chapter 1 the distinction between a population representing the entire collection of units of interest in a research project, and a sample that is the smaller subgroup that we select for actual observations. The variance for a population is represented by the lower case Greek letter sigma squared, or σ^2 . The population variance σ^2 may be calculated using the formula that we have worked with up to this point. The variance for a sample of data, when we are only interested in describing the sample, is represented by the upper case letter S^2 , and it too may be calculated with the formula that we have worked with up to this point.

However, if you also recall from Chapter 1, we made a distinction between descriptive statistics that are used to describe a set of data and inferential statistics that are used to infer something about a population parameter by the observation of sample statistics. We often have an interest in doing just that in behavioral science research, or at the very least we are interested in being able to generalize our results to a larger population. It turns out that if you were to actually know the value of a population's variance, and then take a series of samples from the population and compare the sample variances computed for each sample to the actual population variance, you would find that the sample variances tend to underestimate the true size of the population variance. The sample variance is sometimes referred to as a biased estimator of the population variance, and the direction of the bias is to underestimate the true size of the population variance.

We can reduce the bias of the estimate of the population variance when using data from a sample by making a slight adjustment in the formula for the sample variance when we intended it to serve as an estimate of the population variance. Since the direction of the bias is to underestimate the true population variance, we can increase the size of the estimated variance by using the value $n - 1$ in the denominator of the variance formula in place of the usual denominator n . We will then use the lower case letter s^2 to represent a sample variance that is being used as an estimate of the population variance.

It might be useful at this point to review the computational formulas for the three situations for computing variance.

Symbol	Situation	Computational Formula
σ^2	Population Variance	$\sigma^2 = \frac{\sum X^2 - (\sum X)^2}{n}$
S^2	Sample Variance	$S^2 = \frac{\sum X^2 - (\sum X)^2}{n}$
s^2	Sample Variance used to estimate the population variance σ^2	$s^2 = \frac{\sum X^2 - (\sum X)^2}{n - 1}$

$n - 1$ as degrees of freedom

The use of " $n - 1$ " in the denominator of the formula for variance is the result of the concept of **degrees of freedom**. One of the things that we observed when looking at deviation from the mean was that the sum of the deviations from the mean always equaled zero. However, when dealing with sample data from a population there is no guarantee that the sum of the deviations of the individual sample scores from the *population mean* will actually equal zero. (Don't be confused by the fact that the sum of the deviations of the sample scores from the sample mean will equal zero. What we are concerned with here is whether or not the sum of the deviations of the sample scores from the population mean will equal zero.)

Consider this example of a population of $N = 10$ with a mean $\mu = 30$. (You may want to verify that the sum of the deviations of the population scores from the population mean is actually zero.) Suppose I select a sample of $n = 5$ scores from the population as indicated below.

We can calculate a sample mean for the five scores that were selected from the population, and if we calculate the sum of the deviations from the sample mean for the five scores in the sample we will in fact find the sum equal to. But is the sum of the deviations of the sample scores from the population mean equal to zero? No! The sum of the deviations from the sample scores to the population mean is actually -30 . How

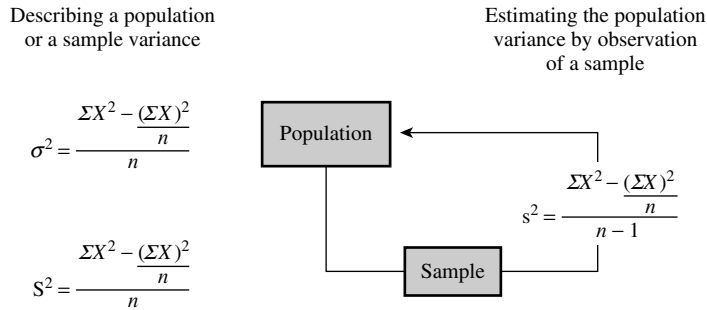


Figure 5.2 Diagram of population and sample with the three formulas

Population			
10	15	20	25
30	35	35	40
45	45		

$N = 10$
 $\mu = 30$

Sample		
10	20	45
15	30	

$n = 5$
 $\bar{X} = 24$

Figure 5.3 Illustration of population data and sample drawn from it

many of the sample scores would I have to change (or control) if I wanted to force the sum of the deviations of the sample scores from the population mean to equal zero? The answer is that *only one* of the sample scores must be controlled. I can always ensure that the sum of the deviations of the sample scores from the population mean will equal zero if I can control one of the sample scores. It does not matter which score I choose to control. I simply must be sure that the value of the score I choose will result in a deviation from the population mean that when added to the other deviations from the population mean will give me a sum of zero. Since I must control only one score, it means that the other scores are free to take on any value. In other words, I have $n - 1$ degrees of freedom. The concept of degrees of freedom will be examined again when we begin our investigation of hypothesis testing in later chapters.

It is easy to be confused by the concept of degrees of freedom, or the difference between describing a sample variance through the use of one formula, and estimating the population variance from sample data using a different formula. Like many of the concepts in statistics, it takes time for reflection and experience before everything falls into place. At this point, the most important thing to keep in mind is what variance tells us about a distribution of scores. Remember, the larger the variance the less likely that the individual scores of a distribution are close to the mean, and the less likely that the mean is a good indicator of what the typical score in a distribution was. *The smaller the variance the more likely that the individual scores of the distribution are close to the mean, and the more likely that the mean is a good indicator of the typical score in a distribution.*

We have examined the logical and computational formulas for calculating the variance when using raw data or a set of individual scores. We will follow a pattern similar to what we did in Chapter 4 on measures of central tendency and also briefly examine the method for computing variance when data are presented in a simple frequency distribution of size $i = 1$, and a frequency distribution when $i > 1$. It is useful to examine these two situations since we do not always control the way data are presented to us. But before moving to these other approaches I want to introduce the concept of **standard deviation**, which is closely related to variance.

Standard deviation

Many more people have heard the term “standard deviation” than the number who actually knows what it means. But what is a “standard deviation?” The **standard deviation** of a set of data is simply the square root of the variance. Just as the variance can be defined as the mean of the squared deviations from the mean for a set of data, the standard deviation can be defined as the square root of the mean of the squared deviations from the mean for a set of data.

You have every reason to be wondering why do we care about the standard deviation when we already know the variance? There are really two reasons. First, recall that we were not able to base a measure of variation in a set of data on the simple deviation from the mean due to the fact that the sum of the simple deviations from the mean was always the same thing—zero. We eliminated that problem by squaring the deviations from the mean, and using the sum of the squared deviations as the basis for our measurement of variance. But in solving one problem we artificially inflated our measure of variance when we squared all of the deviations from the mean. In one sense you can think of the standard deviation as being a measure of variation that is more in line with what we were intending to measure in the first place since by taking the square root of the variance we are “unsquaring” the deviations from the mean.

The information conveyed by the variance and the standard deviation is essentially the same, but the standard deviation is usually a much smaller value. There is an exception of course when the variance is less than 1.00, since the square root of a number greater than zero and less than 1.00 is larger than the original number. For example, the square root of 0.36 is the larger value 0.60, but in most cases the variance is a much larger number, and we will find it much easier to work with the smaller value of a standard deviation.

However, it is the second reason for working with the standard deviation that is much more important. It turns out that by knowing the mean and the standard deviation for certain types of distributions we can also know a great deal about how the individual observations in that distribution are organized. The mean and the standard deviation will tell us a great deal about any distribution that is **normal in form**. We discussed the idea of a distribution being normal in form or having a bell-shaped curve in Chapter 3, and we will be examining the concept of a normal distribution in great detail in the next chapter. Furthermore, the standard deviation, and a related concept of the **standard error** that is based on the standard deviation, will be key terms when we begin our investigation of hypothesis testing in later chapters.

Just as we represented the population variance by the symbol of lower case sigma squared (σ^2), we represent the symbol for a population standard deviation by the lower case sigma (σ). Similarly, a sample standard deviation is represented by the upper case S, and a standard deviation that is being used to infer a population standard deviation is symbolized by a lower case s. We will still have the same problem with standard deviation that we had with variance when we attempt to use sample data to estimate the population standard deviation. We will use the same technique of altering the denominator of the formula by using $n - 1$ in place of the usual denominator n when we wish to estimate the population standard deviation. With this in mind, we can present both the logical formula and the computational formula for the standard deviation.

Logical Formula for Standard Deviation:

$$\sigma \text{ or } S = \sqrt{\frac{\sum(X - \bar{X})^2}{n}}$$

Computational formula for Standard Deviation:

$$\sigma \text{ or } S = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n}}$$

When using data from a sample to estimate a population standard deviation we would simply substitute $n - 1$ in the denominator of the formula. For example, the computational formula would become:

Computational formula for Sample Standard Deviation used to estimate the Population Standard Deviation σ :

$$s = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}}$$

Technically, the correction factor of $n - 1$ is sufficient to adjust for the bias when using the sample variance s^2 as an estimate of the population variance σ^2 . However, the sample based standard deviation s will still be a biased estimator of the population standard deviation σ even when using the correction factor of $n - 1$. This is especially true when working with a very small sample size. Fortunately, once the sample size moves above an $n = 20$ or so, the bias becomes very slight, and because many research applications in the behavioral sciences involve a large sample size we need not worry in most cases. Those of you with statistical function calculators might take a moment to examine your function keys. Many statistical calculators will automatically compute the mean and standard deviation for a set of data, and some will give you a choice on how you want the standard deviation computed. You may see two keys marked as: σ_n , and σ_{n-1} , providing you with a choice of a descriptive or an inferential computation of the standard deviation.

Computing the standard deviation is relatively simple. Just find the variance, and then take the square root. For example, we previously computed the variance for three simple distributions, and obtained the results of

A		B		C	
X	X^2	X	X^2	X	X^2
10	100	10	100	8	64
8	64	10	100	6	36
6	36	6	36	6	36
4	16	2	4	6	36
2	4	2	4	4	16
$\Sigma X = 30$	$\Sigma X^2 = 220$	$\Sigma X = 30$	$\Sigma X^2 = 244$	$\Sigma X = 30$	$\Sigma X^2 = 188$
$(\Sigma X)^2 = 900$		$(\Sigma X)^2 = 900$		$(\Sigma X)^2 = 900$	
$S^2 = 8;$		$S^2 = 12.8;$		and $S^2 = 1.6.$	

Standard Deviation Computation

Distribution A	Distribution B	Distribution C
$\frac{220 - \frac{900}{5}}{5} =$	$\frac{244 - \frac{900}{5}}{5} =$	$\frac{188 - \frac{900}{5}}{5} =$
$\frac{220 - 180}{5} = \frac{40}{5} =$	$\frac{244 - 180}{5} = \frac{64}{5} =$	$\frac{188 - 180}{5} = \frac{8}{5} =$
8	12.8	1.6
$S_A = \sqrt{8} = 2.83$	$S_B = \sqrt{12.8} = 3.58$	$S_C = \sqrt{1.6} = 1.26$

To compute the standard deviation we simply took the square root of each of the observed variances.

Computing Variance and Standard Deviation for Frequency Distributions

As we have seen earlier, we do not always control the way data are presented to us. On occasion we may be presented with data already in the form of a frequency distribution with no access to the original raw scores, and yet we may still want to compute a mean, or a variance and standard deviation. Variance and standard deviation may be computed for frequency distributions by making a simple adjustment in the formula following the same pattern we used when computing the mean for frequency distributions in Chapter 4.

To compute the variance we will need to alter our original computational formula for raw data:

$$S^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n}$$

as follows:

$$S^2 = \frac{\sum fX^2 - \frac{(\sum fX)^2}{n}}{n}$$

where

fX = each value of X times its frequency

fX^2 = each value of X squared times its frequency

We must simply use the number of frequencies f in each interval to weight the value of $\sum X^2$ and $(\sum X)^2$ in each interval.

Consider the simple frequency distribution below, which we used in Chapter 4 to compute a mean. We can add the appropriate columns to the frequency distribution to obtain the necessary sums.

X	f	fX	fX^2
11	1	11	121
9	5	45	405
8	4	32	256
5	3	15	75
3	3	9	27
	$\Sigma f = 16$	$\Sigma fX = 112$	$\Sigma fX^2 = 884$

Note that the fX^2 column represents the product of $f \times X \times X$, and as such you may obtain it two ways. Consider the top two entries in the fX^2 column of the frequency distribution above. You may first square each X value, and then multiply the resulting sum by the appropriate value of f as follows:

$$X = 11 \quad X^2 = 121 \quad f = 1 \quad fX^2 = 1 \times 121 = 121$$

$$X = 9 \quad X^2 = 81 \quad f = 5 \quad fX^2 = 5 \times 81 = 405$$

and so on, or you might see that you already have the value of fX , and may simply multiply that column by the value of X . Since multiplication is commutative it does not matter what order you perform the computation. That is:

$$fX^2 = f \times X \times X = f \times X^2 = fX \times X$$

So we can also obtain the fX^2 column as follows:

$$X = 1 \quad f = 1 \quad fX = 11 \quad fX^2 = fX \times X = 11 \times 11 = 121$$

$$X = 9 \quad f = 5 \quad fX = 45 \quad fX^2 = fX \times X = 45 \times 9 = 405$$

Use which ever method is easier for you.

To complete the computation for the variance we need only plug the numbers into the formula. Keep in mind that n is equal to the total number of observations in our sample, not the number of intervals in which they happen to be categorized. In this case, $n = 16$, not 5.

$$S^2 = \frac{\Sigma fX^2 - \frac{(\Sigma fX)^2}{n}}{n}$$

$$S^2 = \frac{884 - \frac{12544}{16}}{16}$$

$$S^2 = \frac{884 - 784}{16}$$

$$S^2 = \frac{100}{16} = 6.25$$

So our variance is equal to 6.25.

To compute the standard deviation we would take the square root of the variance.

$$S = \sqrt{6.25} = 2.5$$

Notice that we are assuming that we are interested in describing the observed sample, and are not attempting to estimate a population variance or standard deviation. This is evident from our use of $n = 16$ in the formula instead of $n - 1 = 15$.

We follow the same general procedure when working with a frequency distribution where the interval size is greater than 1. The formula for variance must be adjusted just as it was when working with the simpler frequency distribution. We may write the formula as before as:

$$S^2 = \frac{\Sigma fX'^2 - \frac{(\Sigma fX')^2}{n}}{n}$$

where

X' = the interval midpoint.

We will again use one of the frequency distributions from Chapter 4. Our first step is to find the midpoint of each interval. From that point on we are simply repeating the procedure that we followed

for the frequency distribution where $i = 1$. We will need an fX' column, and an fX'^2 column to compute the variance.

X	f	X'	fX'	fX'^2
45-49	6	47	282	13,254
40-44	8	42	336	14,112
35-39	12	37	444	16,428
30-34	10	32	320	10,240
25-29	9	27	243	6,561
20-24	5	22	110	2,420
	$\Sigma f = 50$		$\Sigma fX' = 1,735$	$\Sigma fX'^2 = 63,015$

$$S^2 = \frac{63015 - \frac{(1735)^2}{50}}{50}$$

$$S^2 = \frac{63015 - \frac{3,010,225}{50}}{50}$$

$$S^2 = \frac{63015 - 60204.5}{50}$$

$$S^2 = \frac{2810.5}{50} = 56.21$$

We have computed the variance, and may now compute the standard deviation by simply taking the square root of 56.21.

$$S = \sqrt{56.21} = 7.50$$

Variance as Prediction Error (or Cabo San Lucas Here I Come!)

We have examined the idea of variation in data in several different ways such as the range, the IQR, the SIQR, variance, and standard deviation. Toward that end, we have spent a great deal of time doing a variety of computations. While it is important to be able to take a statistical formula, apply it to a set of data, and generate the correct result, it is much more important to know why we do it. In other words, what does the resulting statistic mean? What does it tell us about a set of data that we did not know before?

Up until now I have stressed the idea that variance is important because it tells us something about the way that individual scores are distributed around their mean. By knowing the size of the variance we know how representative the mean is of the individual scores in the distribution. The variance viewed in those terms is an important piece of information, but in the behavioral sciences we can, and often do, look at variance in another way. We look at variance as a type of prediction error, and try to find ways to reduce the size of prediction error. Or you might think of the process as simply trying to do a better job of predicting the value of some variable that we consider important in our research.

Suppose we have a set of data representing annual income in thousands of dollars for a group of $n = 10$ individuals.

Income ($\times 1,000$)
30
20
50
15
12
75
40
15
22
16
$\Sigma f = 295$

We could calculate the mean income and would find that it is 29.5, or \$29,500 per year. By examining the 10 scores in the distribution you can see that there is variation present. That is, not everyone has an income of 29.5; the incomes vary, some are higher and some are lower.

Now suppose I told you that I had the income of each person written on a piece of paper, and that I was going to draw the pieces of paper at random and let you guess what the person's income was. The only restriction is that you have to make the same guess each time. You are free to choose any of the incomes represented, or any other value for that matter as your guess. All 10 incomes will eventually be selected, and I am going to measure how well you guess by comparing your guess to the income that is chosen. Since some of your guesses will be too high and others too low, I am going to square the difference between the actual income and your guess to eliminate any negative numbers. After all 10 incomes have been selected, I will calculate a mean of your squared difference for each guess as an indication of how well you have done.

What income value would you choose as your guess? You want to select a value to guess each time that will give you the smallest amount of squared error possible. (Choose wisely because there might be a prize in this for you if you win. I'm thinking maybe some nice luggage and a trip to Cabo San Lucas, but I haven't made up my mind yet.)

If you examine the distribution you will notice that there are two scores at 15, and you might be tempted to select the value 15 as your guess since your mean squared difference for those two cases would be zero. However, your mean squared difference across all 10 incomes would be 575.9 if you select the value 15 as your guess. (You might want to verify this by subtracting 15 from each of the observed incomes, squaring the difference, and then averaging the 10 results). Is that a winning performance? I doubt it, it seems high.

Selecting the mode of 15 did not seem to be a wise strategy. What if you selected the median income as your guess? If you rearrange the numbers you will find the median to be equal to 21. If 21 becomes your guess, you will not be exactly right on any of the 10 incomes, but your mean squared difference across all 10 incomes will be 437.9, which is much improved over the strategy of selecting 15, but is it the best you could do?

One final strategy might be to select the mean value of 29.5 as your guess. Again, you will not be exactly right on any of the 10 incomes, but how would you do across all 10? By selecting the mean of 29.5 as your guess, your mean squared difference would be 365.65. That is the best we have seen yet, and in fact, there is no other guess that would be any better. Now if you think about how we have measured

the accuracy of each guess, you might recognize that the value of 365.65 represents something else we have examined in this chapter. We took each score, subtracted your guess of the mean, squared the difference, and then calculated the mean of the 10 squared deviations. In other words, we computed the *variance*.

Income ($\times 1,000$)	$(X - \bar{X})$	$(X - \bar{X})^2$
30	0.5	0.25
20	-9.5	90.25
50	20.5	420.25
15	-14.5	210.25
12	-17.5	306.25
75	45.5	2070.25
40	10.5	110.25
15	-14.5	210.25
22	-7.5	56.25
16	-13.5	182.25
		$\Sigma(X - \bar{X})^2 = 3656.50$

$$s^2 = \frac{3656.50}{10} = 365.65$$

Since there is no other single value that would serve as a better guess of the individual scores than the mean, we can think of the variance as the maximum amount of prediction error that we would have to accept when trying to predict an individual score. Or to think of it a different way, your best guess of a score in a distribution is the mean, assuming that no additional information is available to you. (For those of you who originally guessed the mean, the prize committee informs me that neither the trip to Cabo nor the luggage is available. You do win our home game allowing you hours of fun guessing anyone's income you please.)

Now let's change the rules a little bit. Suppose before I have you guess the income value, I am willing to give you one piece of additional information. It would be in your best interest to ask for something that might help you better predict income. What sorts of things (variables) are related to income? You can probably think of many things, but certainly education is a key variable that helps determine one's income. Suppose I am willing to tell you if the income I have selected is that of a person who has a college degree or not. Let's also suppose that I am willing to let you provide two different incomes as your guess; one for the college graduates, and one for the noncollege graduates. (I hope you understand that by changing the rules of the game the trip to Cabo San Lucas is definitely out of the question for this round. No, you're not going to get the luggage either!)

As you might suspect, your best strategy for guessing has changed. In the light of this new information, your best strategy is to guess the mean of the college graduates as the selected income when you know I have selected a college graduate, and to guess the mean of the noncollege graduates when you know I have selected a noncollege graduate. Let's look at the income distribution again with the new information added.

College Graduate	Income (× 1,000)	
YES	30	
NO	20	
YES	50	
NO	15	
NO	12	College Graduate's Mean = 43.4
YES	75	Noncollege Graduate's Mean = 15.6
YES	40	
NO	15	
YES	22	
NO	16	

The mean income of the college graduates is 43.4 or \$43,400, and the mean income of the noncollege graduates is 15.6, or \$15,600. How well can you guess income now if you guess the college graduate mean of 43.4 when you know the individual has a college degree, and guess the noncollege graduate mean of 15.6 when you know the individual does not have a college degree? We will substitute the appropriate mean into the calculation of $(X - \bar{X})$, and $(X - \bar{X})^2$, and then compute what we can think of as a modified variance (symbolized by S^2).

College Graduate	Income (× 1,000)	$(X - \bar{X})$	$(X - \bar{X})^2$
YES	30	$(30 - 43.4) = -13.4$	179.56
NO	20	$(20 - 15.6) = 4.4$	19.36
YES	50	$(50 - 43.4) = 6.6$	43.56
NO	15	$(15 - 15.6) = -0.6$	0.36
NO	12	$(12 - 15.6) = -3.6$	12.96
YES	75	$(75 - 43.4) = 31.6$	998.56
YES	40	$(40 - 43.4) = -3.4$	11.56
NO	15	$(15 - 15.6) = -0.6$	0.36
YES	22	$(22 - 43.4) = -21.4$	457.96
NO	16	$(16 - 15.6) = 0.4$	0.16
			$\Sigma(X - \bar{X})^2 = 1724.40$

$$S^2 = \frac{1724.40}{10} = 172.44$$

When we use the mean income of the college graduates to guess a college graduate's income, and the mean income of the noncollege graduates to guess the mean income of the noncollege graduates, and then find the mean squared deviation, we arrive at a modified measure of variance; one that uses the specific group mean in place of the overall group mean. In this case using educational level to help predict income results in a modified variance of 172.44, which we can compare with our previous variance of 365.65. By using the appropriate mean income for each group we have been able to reduce the amount

of variance by 193.21 points ($365.65 - 172.44$), or we can express that difference as a percentage of the original variance and say that we have reduced the variance by 52.8%.

Using educational level to help predict income has explained over half (52.8%) of the variance in income. Not everyone in our small sample of 10 individuals has the same income, variation is present. By knowing what the individual's educational level is we are able to explain or account for over half of the variation in income. This idea of being able to explain or **reduce variance** in one variable by knowing the value of a second variable is one of the more important concepts in statistical analysis in the behavioral sciences. We will deal with this concept again when looking at the interpretation of correlation between two variables in Chapter 9, and in assessing the quality of a linear regression analysis in Chapter 10.

Computer Applications

1. Select several variables from the GSS data set, and generate descriptive statistics.
2. Be sure to click on the "Options" button and request the variance and range in addition to the default statistics of mean, standard deviation, minimum, and maximum.
3. Enter data from one of the short examples from the text and compare the results to SPSS. Is SPSS using the variance formula with " n " as the denominator, or " $n - 1$?"

How to do it

Open the GSS data set, and then click on "Analyze," "Descriptive Statistics," and then "Descriptives." Highlight the desired variables and select them by clicking on the direction arrow. Click on "Options" to request additional statistics such as the variance and range. Click on "OK" to run the procedure.

Clear the GSS data set by clicking on "File," "New," and then "Data." Use the new empty Data Editor Screen to input data from one of the simple examples from the text. Run the descriptive statistics procedure with "variance" requested, and determine the formula used.

Summary of Key Points

Measures of variation are a group of statistics that indicate how a set of scores is distributed. Some measures of variation indicate the variation from the bottom or lower end of the distribution to the top or the upper end of the distribution, while other measures of variation indicate the variation of each score from a central point such as the mean. The former techniques typically measure the range of the distribution, while the latter techniques typically measure deviation from the mean.

Range—The range is the distance from the smallest score in a distribution to the largest score. It is one of the simplest measures of variation.

Interquartile Range—The interquartile range is the distance from the 75th percentile to the 25th percentile in a distribution.

Semi-Interquartile Range—The semi-interquartile range is the interquartile range divided by 2.

Deviation from the Mean—Deviation from the mean is a measure of each score's distance from the mean. The sum of the deviations from the mean is always zero.

Variance—The variance is the average of the squared deviations from the mean for a set of scores. We square the deviations to keep the positive and negative deviations from cancelling each other out. The smaller the variance the more closely the scores of a distribution are to their mean.

Standard Deviation—The standard deviation is the square root of the variance.

Degrees of Freedom—The number of values in a sample that are free to take on any value and still represent an unbiased estimate of a population parameter.

Normal in Form—A distribution of scores whose shape approximates that of a normal or bell-shaped curve.

Reduce Variance—The ability to reduce the amount of error when predicting a variable by making use of information obtained from a second variable. Reducing or explaining variance is an important concept that is central to several statistical procedures.

Questions and Problems for Review

1. Compute the range for the following sets of data:
 A. 24 26 28 30 38 49 55 56 67 75
 B. 0.50 1.25 3.50 4.55 8.95 10.50
2. Under what circumstances would it be wise to compute the IQR or the SIQR?
3. Examine the frequency distribution below. Do you think it would be better to report the range for these data, or the IQR? Why?

X	f	Cf
95–99	3	135
90–94	3	132
85–89	20	129
80–84	25	109
75–79	29	84
70–74	24	55
65–69	20	31
60–64	4	11
55–59	5	7
50–54	2	2
$\Sigma f = 135$		

4. Compute the range, IQR, and SIQR for the data in Problem 3 above.
5. Examine the summary information presented below. What can you conclude about the income distribution for each occupational category? For which groups does the mean seem to be a better measure of central tendency? For which groups is the mean less indicative of overall group income?

Occupation	Income	
	Mean (\$)	Standard Deviation (\$)
Accountant	35,500	8,500
Attorney	72,200	23,100
Engineer	57,800	12,100
Evangelist	66,600	57,800
Physician	108,000	8,200
Psychic	23,700	4,250

6. Compute the variance and standard deviation for the two sections of the statistics class illustrated in Figure 5.1.

7. Compute the variance and standard deviation for the simple frequency distribution below.

X	f
20	2
18	4
15	7
12	4
10	3
$\Sigma f = 20$	

8. What does it mean to be able to explain variance?
 9. Scores for the verbal section of the SAT are presented below for a group of $n = 10$ students.
 A. Compute the variance for the entire group.
 B. Compute the mean score of the females, and the mean score of the males.

Gender	Verbal SAT
Female	500
Female	650
Female	485
Female	720
Male	450
Male	395
Male	700
Male	630
Male	450
Male	585

10. How much of the variance in the SAT scores in problem 9 can be explained by gender? (Hint: you will need to use the mean score of the females when predicting a female score, and the mean score of the males when predicting a male score to compute a modified variance).