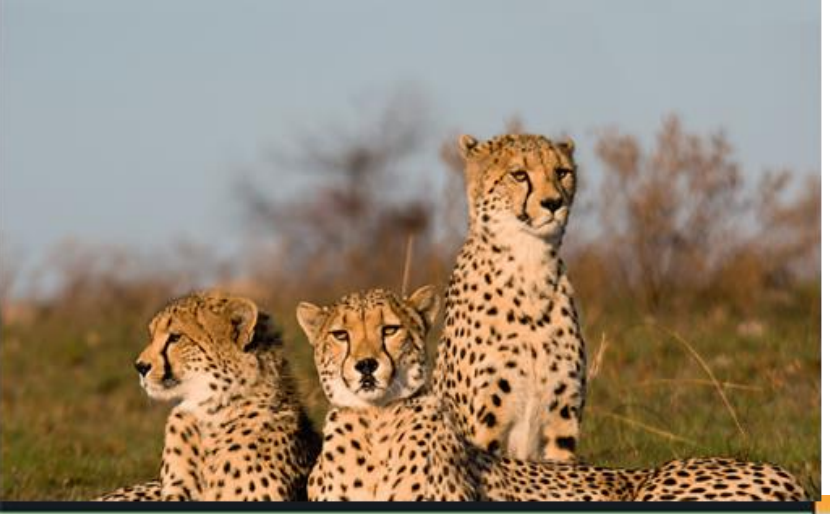
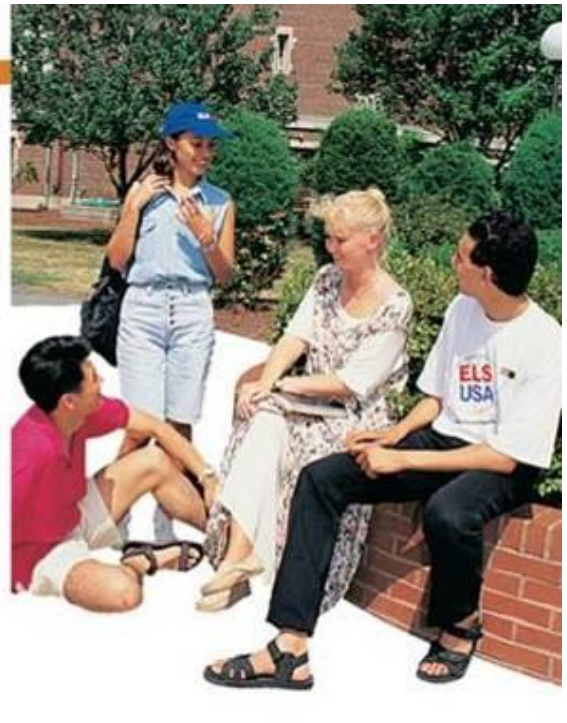


Averages and Variation

3

Section 3.2

Measures of Variation



Focus Points

- Find the range, variance, and standard deviation.
- Compute the coefficient of variation from raw data. Why is the coefficient of variation important?
- Apply Chebyshev's theorem to raw data. What does a Chebyshev interval tell us?

Measures of Variation

An average is an attempt to summarize a set of data using just one number. As some of our examples have shown, an average taken by itself may not always be very meaningful.

We need a statistical cross-reference that measures the spread of the data.

The *range* is one such measure of variation.

The **range** is the difference between the largest and smallest values of a data distribution.

Example 5 – *Range*

A large bakery regularly orders cartons of Maine blueberries.

The average weight of the cartons is supposed to be 22 ounces. Random samples of cartons from two suppliers were weighed.

The weights in ounces of the cartons were

Supplier I:	17	22	22	22	27
Supplier II:	17	19	20	27	27

Example 5 – *Range*

cont'd

(a) Compute the range of carton weights from each supplier.

Range = Largest value – Smallest value

Supplier I = range 27 – 17 = 10 ounces

Supplier II = range 27 – 17 = 10 ounces

(b) Compute the mean weight of cartons from each supplier. In both cases the mean is 22 ounces.

Example 5 – *Range*

cont'd

(c) Look at the two samples again. The samples have the same range and mean. How do they differ?

The bakery uses one carton of blueberries in each blueberry muffin recipe. It is important that the cartons be of consistent weight so that the muffins turn out right.

Supplier I provides more cartons that have weights closer to the mean. Or, put another way, the weights of cartons from Supplier I are more clustered around the mean.

The bakery might find Supplier I more satisfactory.

Variance and Standard Deviation

Variance and Standard Deviation

We need a measure of the distribution or spread of data around an expected value (either $\bar{x}$ or μ). *Variance* and *standard deviation* provide such measures.

Formulas and rationale for these measures are described in the next Procedure display. Then, examples and guided exercises show how to compute and interpret these measures.

As we will see later, the formulas for variance and standard deviation differ slightly, depending on whether we are using a sample or the entire population.

Variance and Standard Deviation

Procedure:

HOW TO COMPUTE THE SAMPLE VARIANCE AND SAMPLE STANDARD DEVIATION

Quantity	Description
x	The variable x represents a **data value** or outcome.
Mean $\bar{x} = \frac{\Sigma x}{n}$	This is the **average of the data values,** or what you "expect" to happen the next time you conduct the statistical experiment. Note that n is the sample size.
$x - \bar{x}$	This is the **difference** between what happened and what you expected to happen. This represents a "deviation" away from what you "expect" and is a measure of risk.
$\Sigma(x - \bar{x})^2$	The expression $\Sigma(x - \bar{x})^2$ is called the **sum of squares.** The $(x - \bar{x})$ quantity is squared to make it nonnegative. The sum is over all the data. If you don't square $(x - \bar{x})$, then the sum $\Sigma(x - \bar{x})$ is equal to 0 because the negative values cancel the positive values. This occurs even if some $(x - \bar{x})$ values are large, indicating a large deviation or risk.

Variance and Standard Deviation cont'd

How to compute the sample variance and sample standard deviation

Quantity	Description
Sum of squares $\Sigma(x - \bar{x})^2$ or $\Sigma x^2 - \frac{(\Sigma x)^2}{n}$	This is an **algebraic simplification of the sum of squares** that is easier to compute. The **defining formula** for the sum of squares is the upper one. The **computation formula** for the sum of squares is the lower one. Both formulas give the same result.
Sample variance $s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$ or $s^2 = \frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1}$	The **sample variance is s^2.** The variance can be thought of as a kind of average of the $(x - \bar{x})^2$ values. However, for technical reasons, we divide the sum by the quantity $n - 1$ rather than n. This gives us the best mathematical estimate for the sample variance. The **defining formula** for the variance is the upper one. The **computation formula** for the variance is the lower one. Both formulas give the same result.

Variance and Standard Deviation cont'd

How to compute the sample variance and sample standard deviation

Quantity	Description
Sample standard deviation $s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1}}$	This is the **sample standard deviation, *s*.** Why do we take the square root? Well, if the original x units were, say, days or dollars, then the s^2 units would be days squared or dollars squared (wow, what's that?). We take the square root to return to the original units of the data measurements. The standard deviation can be thought of as a measure of variability or risk. Larger values of s imply greater variability in the data. The **defining formula** for the standard deviation is the upper one. The **computation formula** for the standard deviation is the lower one. Both formulas give the same result.

Variance and Standard Deviation

In statistics, the sample standard deviation and sample variance are used to describe the spread of data about the mean $\bar{x}$.

The next example shows how to find these quantities by using the defining formulas.

As you will discover, for “hand” calculations, the computation formulas for s^2 and s are much easier to use.

Variance and Standard Deviation

However, the defining formulas for s^2 and s emphasize the fact that the variance and standard deviation are based on the differences between each data value and the mean.

Defining Formulas (Sample Statistic)

$$\text{Sample variance} = s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} \quad (1)$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} \quad (2)$$

where x is a member of the data set, $\bar{x}$ is the mean, and n is the number of data values. The sum is taken over all data values.

Variance and Standard Deviation

Computation Formulas (Sample Statistic)

$$\text{Sample variance} = s^2 = \frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1} \quad (3)$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1}} \quad (4)$$

where x is a member of the data set, $\overline{x}$ is the mean, and n is the number of data values. The sum is taken over all data values.

Example 6 – *Sample Standard Deviation (Defining Formula)*

Big Blossom Greenhouse was commissioned to develop an extra large rose for the Rose Bowl Parade.

A random sample of blossoms from Hybrid A bushes yielded the following diameters (in inches) for mature peak blooms.

2 3 3 8 10 10

Use the defining formula to find the sample variance and standard deviation.

Example 6 – *Solution*

Several steps are involved in computing the variance and standard deviation. A table will be helpful (see Table 3-1).

Column I x	Column II $x - \bar{x}$	Column III $(x - \bar{x})^2$
2	$2 - 6 = -4$	$(-4)^2 = 16$
3	$3 - 6 = -3$	$(-3)^2 = 9$
3	$3 - 6 = -3$	$(-3)^2 = 9$
8	$8 - 6 = 2$	$(2)^2 = 4$
10	$10 - 6 = 4$	$(4)^2 = 16$
10	$10 - 6 = 4$	$(4)^2 = 16$
$\Sigma x = 36$		$\Sigma(x - \bar{x})^2 = 70$

Diameters of Rose Blossoms (in inches)

Table 3-1

Since $n = 6$, we take the sum of the entries in column I of Table 3-1 and divide by 6 to find the mean $\bar{x}$.

$$\bar{x} = \frac{\Sigma x}{n}$$

Example 6 – *Solution*

cont'd

$$= \frac{36}{6}$$

$$= 6.0 \text{ inches}$$

Using this value for $\bar{x}$, we obtain Column II. Square each value in column II to obtain Column III, and then add the values in Column III.

To get the sample variance, divide the sum of Column III by $n - 1$. Since $n = 6$, $n - 1 = 5$.

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

Example 6 – *Solution*

cont'd

$$= \frac{70}{5}$$

$$= 14$$

Now obtain the sample standard deviation by taking the square root of the variance.

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

$$= \sqrt{14}$$

$$\approx 3.74$$

Variance and Standard Deviation

In most applications of statistics, we work with a random sample of data rather than the entire population of *all* possible data values.

Variance and Standard Deviation

However, if we have data for the entire population, we can compute the *population mean* μ, *population variance* σ^2, and *population standard deviation* σ (lowercase Greek letter sigma) using the following formulas:

Population Parameters

$$\text{Population mean} = \mu = \frac{\Sigma x}{N}$$

$$\text{Population variance} = \sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

$$\text{Population standard deviation} = \sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{N}}$$

where N is the number of data values in the population and x represents the individual data values of the population.

Variance and Standard Deviation

We note that the formula for μ is the same as the formula for $\bar{x}$ (the sample mean) and that the formulas for σ^2 and σ are the same as those for s^2 and s (sample variance and sample standard deviation), except that the population size N is used instead of $n - 1$.

Also, μ is used instead of $\bar{x}$ in the formulas for σ^2 and σ.

In the formulas for s and σ, we use $n - 1$ to compute s and N to compute σ. Why?

The reason is that N (capital letter) represents the *population size*, whereas n (lowercase letter) represents the sample size.

Variance and Standard Deviation

Since a random sample usually will not contain extreme data values (large or small), we divide by $n - 1$ in the formula for s to make s a little larger than it would have been had we divided by n.

Courses in advanced theoretical statistics show that this procedure will give us the best possible estimate for the standard deviation σ.

In fact, s is called the *unbiased estimate* for σ. If we have the population of all data values, then extreme data values are, of course, present, so we divide by N instead of $N - 1$.

Variance and Standard Deviation

Comment

The computation formula for the population standard deviation is

$$\sigma = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/N}{N}}$$

Now let's look at two immediate applications of the standard deviation. The first is the coefficient of variation, and the second is Chebyshev's theorem.

Coefficient of Variation

Coefficient of Variation

A disadvantage of the standard deviation as a comparative measure of variation is that it depends on the units of measurement.

This means that it is difficult to use the standard deviation to compare measurements from different populations.

For this reason, statisticians have defined the *coefficient of variation*, which expresses the standard deviation as a percentage of the sample or population mean.

Coefficient of Variation

If $\bar{x}$ and s represent the sample mean and sample standard deviation, respectively, then the sample **coefficient of variation** CV is defined to be

$$CV = \frac{s}{\bar{x}} \cdot 100$$

If μ and σ represent the population mean and population standard deviation, respectively, then the population coefficient of variation CV is defined to be

$$CV = \frac{\sigma}{\mu} \cdot 100$$

Notice that the numerator and denominator in the definition of *CV* have the same units, so *CV* itself has no units of measurement.

Coefficient of Variation

This gives us the advantage of being able to directly compare the variability of two different populations using the coefficient of variation.

In the next example, we will compute the *CV* of a population and of a sample and then compare the results.

Example 7 – *Coefficient of Variation*

The Trading Post on Grand Mesa is a small, family-run store in a remote part of Colorado. The Grand Mesa region contains many good fishing lakes, so the Trading Post sells spinners (a type of fishing lure).

The store has a very limited selection of spinners. In fact, the Trading Post has only eight different types of spinners for sale. The prices (in dollars) are

2.10 1.95 2.60 2.00 1.85 2.25 2.15 2.25

Since the Trading Post has only eight different kinds of spinners for sale, we consider the eight data values to be the *population*.

Example 7 – *Coefficient of Variation* cont'd

(a) Use a calculator with appropriate statistics keys to verify that for the Trading Post data, and $\mu \approx \$2.14$ and $\sigma \approx \$0.22$.

Solution:

Since the computation formulas for $\bar{x}$ and μ are identical, most calculators provide the value of $\bar{x}$ only.

Use the output of this key for μ. The computation formulas for the sample standard deviation s and the population standard deviation σ are slightly different.

Be sure that you use the key for σ (sometimes designated as σ_n or σ_x).

Example 7 – *Coefficient of Variation* cont'd

(b) Compute the *CV* of prices for the Trading Post and comment on the meaning of the result.

Solution:

$$CV = \frac{\sigma}{\mu} \times 100\%$$

$$= \frac{0.22}{2.14} \times 100\%$$

$$= 10.28\%$$

Example 7 – *Solution*

cont'd

Interpretation

The coefficient of variation can be thought of as a measure of the spread of the data relative to the average of the data.

Since the Trading Post is very small, it carries a small selection of spinners that are all priced similarly.

The *CV* tells us that the standard deviation of the spinner prices is only 10.28% of the mean.

Chebyshev's Theorem

Chebyshev's Theorem

However, the concept of data spread about the mean can be expressed quite generally for *all data distributions* (skewed, symmetric, or other shapes) by using the remarkable theorem of Chebyshev.

Chebyshev's Theorem

For *any* set of data (either population or sample) and for any constant k greater than 1, the proportion of the data that must lie within k standard deviations on either side of the mean is *at least*

$$1 - \frac{1}{k^2}$$

Chebyshev's Theorem

Results of Chebyshev's Theorem

For *any* set of data:

- *at least* 75% of the data fall in the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$.
- *at least* 88.9% of the data fall in the interval from $\mu - 3\sigma$ to $\mu + 3\sigma$.
- *at least* 93.8% of the data fall in the interval from $\mu - 4\sigma$ to $\mu + 4\sigma$.

The results of Chebyshev's theorem can be derived by using the theorem and a little arithmetic.

Chebyshev's Theorem

For instance, if we create an interval $k = 2$ standard deviations on either side of the mean, Chebyshev's theorem tells us that

$$1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} \text{ or } 75\%$$

is the minimum percentage of data in the $\mu - 2\sigma$ to $\mu + 2\mu$ interval.

Chebyshev's Theorem

Notice that Chebyshev's theorem refers to the *minimum* percentage of data that must fall within the specified number of standard deviations of the mean.

If the distribution is mound-shaped, an even *greater* percentage of data will fall into the specified intervals.

Example 8 – *Chebyshev's theorem*

Students Who Care is a student volunteer program in which college students donate work time to various community projects such as planting trees.

Professor Gill is the faculty sponsor for this student volunteer program.

For several years, Dr. Gill has kept a careful record of x = total number of work hours volunteered by a student in the program each semester.

Example 8 – *Chebyshev's theorem* cont'd

For a random sample of students in the program, the mean number of hours was $\bar{x} = 29.1$ hours each semester, with a standard deviation $s = 1.7$ of hours each semester.

Find an interval A to B for the number of hours volunteered into which at least 75% of the students in this program would fit.

Solution:

According to results of Chebyshev's theorem, at least 75% of the data must fall within 2 standard deviations of the mean.

Example 8 – *Solution*

cont'd

Because the mean is $\bar{x} = 29.1$ and the standard deviation is $s = 1.7$, the interval is

$$\bar{x} - 2s \text{ to } \bar{x} + 2s$$

$$29.1 - 2(1.7) \text{ to } 29.1 + 2(1.7)$$

$$25.7 \text{ to } 32.5$$

At least 75% of the students would fit into the group that volunteered from 25.7 to 32.5 hours each semester.