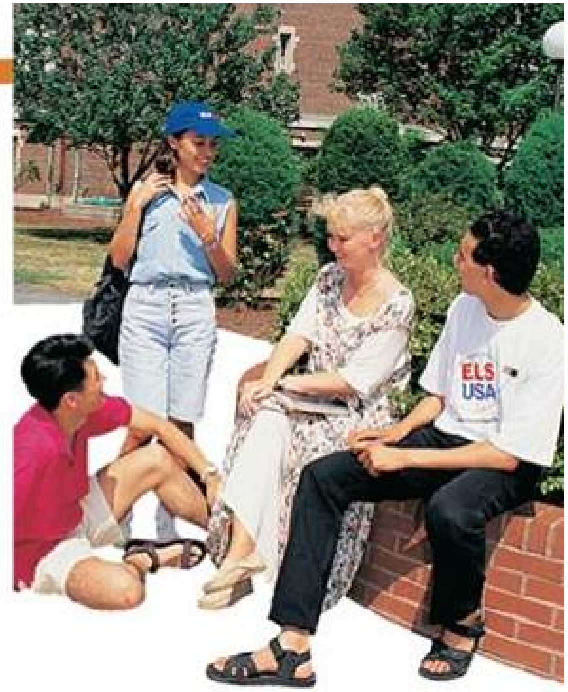
Section 3.2

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

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

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.

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

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

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

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.

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) Find the value of mean and standard deviation and verify that for the Trading Post data, and $\mu \approx$ \$2.14 and $\sigma \approx$ \$0.22.

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

Solution:

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

$$= \frac{0.22}{2.14} \times 100 \qquad = 10.28\%$$

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

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.

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.

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

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.

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

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

Example 8 – *Solution*

cont'd

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

$$25.7 \text{ to } 32.5$$

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

Each of the following data sets has a mean of .

(i) 8 9 10 11 12 (ii) 7 9 10 11 13 (iii) 7 8 10 12 13

A) Without doing any computations, order the data sets according to increasing value of standard deviations.

((i), (ii), (iii))

B) Why do you expect the difference in standard deviations between data sets (i) and (ii) to be greater than the difference in standard deviations between data sets (ii) and (iii)? Hint: consider how much the data in the respective sets differ from the mean.

(The data change between data sets (i) and (ii) increased the square difference $\sum(x-\overline{x})^2$ by 10, whereas the data change between the data sets (ii) and (iii) increased the squared difference $\sum(x-\overline{x})^2$ by only 6).