Chapter 3 Describing, Exploring, and Comparing Data
Measures of Variation (3.2)

Objectives

In this lesson, we will learn how to

- describe why variation is important in statistics;
- compute and interpret the range, standard deviation, and variance;
- distinguish between sample and population standard deviation;
- use the range rule of thumb to interpret standard deviation and identify unusual values;
- apply the Empirical Rule and Chebyshev's Theorem;
- compare variation using the coefficient of variation.

1. Why Variation Matters

Variation describes how spread out the data values are.

Two data sets may have the same center, but they can still look very different if one set has much more spread than the other.

Key Concept

Variation is one of the most important ideas in statistics. We do not only want to compute numbers such as range, standard deviation, and variance; we also want to understand what those numbers tell us about the data.

Round-Off Rule for Measures of Variation

When rounding a measure of variation, carry one more decimal place than is present in the original set of data.

2. Range

The range of a set of data values is the difference between the maximum data value and the minimum data value.

Range

$$\text{Range} = (\text{maximum data value}) - (\text{minimum data value}).$$

Important Property of the Range

The range uses only the maximum and minimum values. Therefore, it is very sensitive to extreme values and is not resistant.

Example 1: Find the range of these wait times, in minutes, for Space Mountain:

$$50,\ 25,\ 75,\ 35,\ 50,\ 25,\ 30,\ 50,\ 45,\ 25,\ 20.$$

Solution

The maximum value is

$$75.$$

The minimum value is

$$20.$$

Therefore,

$$\text{Range} = 75 - 20 = 55.$$

Using the round-off rule, the range is

$$\boxed{55.0 \text{ minutes}}.$$

3. Standard Deviation

The standard deviation of a set of sample values, denoted by s, measures how much the data values deviate away from the mean.

Notation

$$s = \text{sample standard deviation}$$
$$\sigma = \text{population standard deviation}$$

Sample Standard Deviation

For a sample, the standard deviation is

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

A shortcut formula often used by calculators and software is

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

Important Properties of Standard Deviation

- The standard deviation measures how much values typically vary from the mean.
- The value of s is never negative.
- The value of s is zero only when all data values are exactly the same.

- Larger values of s indicate greater variation.
- The units of s are the same as the units of the original data values.
- The standard deviation can increase dramatically when outliers are included.

4. Calculating Standard Deviation

Example 2: Use the sample standard deviation formula to find the standard deviation of these wait times, in minutes:

$$50,\ 25,\ 75,\ 35,\ 50,\ 25,\ 30,\ 50,\ 45,\ 25,\ 20.$$

Solution

First compute the mean:

$$\bar{x} = \frac{50+25+75+35+50+25+30+50+45+25+20}{11} = \frac{430}{11} \approx 39.1.$$

Now subtract the mean from each data value:

$$10.9,\ -14.1,\ 35.9,\ -4.1,\ 10.9,\ -14.1,$$

$$-9.1,\ 10.9,\ 5.9,\ -14.1,\ -19.1.$$

Square these deviations and add them:

$$\sum (x-\bar{x})^2 \approx 2739.1.$$

Then divide by $n-1=10$:

$$\frac{2739.1}{10} = 273.91.$$

Finally, take the square root:

$$s = \sqrt{273.91} \approx 16.6.$$

So the sample standard deviation is

$$\boxed{16.6 \text{ minutes}}.$$

Example 3: Use the shortcut formula to compute the sample standard deviation for the same data set.

Solution

For the data set,

$$n = 11, \qquad \sum x = 430, \qquad \sum x^2 = 19550.$$

Using the shortcut formula,

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

Substitute:

$$s = \sqrt{\frac{11(19550) - (430)^2}{11(10)}}.$$

$$s = \sqrt{\frac{215050 - 184900}{110}} = \sqrt{274.09} \approx 16.6.$$

The result is approximately

$$\boxed{16.6 \text{ minutes}}.$$

5. Range Rule of Thumb

Range Rule of Thumb for Understanding Standard Deviation

For many data sets, the vast majority of sample values, such as about 95%, lie within 2 standard deviations of the mean.

Identifying Significant Values

- Significantly low values are $\mu - 2\sigma$ or lower.
- Significantly high values are $\mu + 2\sigma$ or higher.
- Values between $\mu - 2\sigma$ and $\mu + 2\sigma$ are not significant by this rule.

Estimating Standard Deviation

A rough estimate of the standard deviation is

$$s \approx \frac{\text{range}}{4}.$$

6. Population Standard Deviation

A different formula is used for the standard deviation of a population. Instead of dividing by $n - 1$, we divide by the population size N.

Population Standard Deviation

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

7. Variance

The variance of a set of values is the square of the standard deviation.

Variance Notation

$$s^2 = \text{sample variance}$$

$$\sigma^2 = \text{population variance}$$

Important Properties of Variance

- The units of variance are the squares of the original units.
- Variance is not resistant; it can increase dramatically with outliers.
- Variance is never negative.
- Variance is zero only when all data values are the same.
- The sample variance s^2 is an unbiased estimator of the population variance σ^2.

8. Why Divide by $n - 1$?

When calculating the sample standard deviation or sample variance, we divide by $n - 1$ instead of n. The number $n - 1$ is called the degrees of freedom.

Reason

With a given sample mean, only $n - 1$ of the sample values can vary freely. Once the first $n - 1$ values are known, the last value is forced by the mean.

Using $n-1$ helps sample variances s^2 center around the population variance σ^2.

9. Empirical Rule

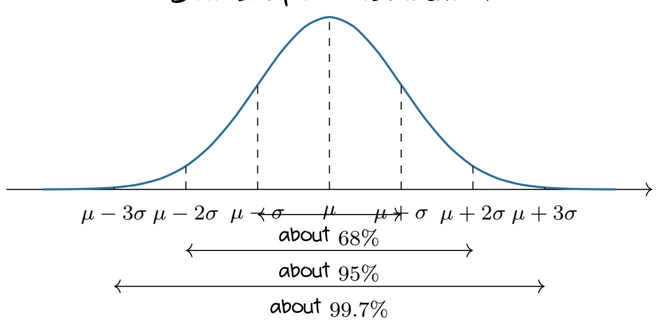
The Empirical Rule applies to data sets with distributions that are approximately bell-shaped.

Empirical Rule

For a bell-shaped distribution:

- About 68% of all values fall within 1 standard deviation of the mean.
- About 95% of all values fall within 2 standard deviations of the mean.
- About 99.7% of all values fall within 3 standard deviations of the mean.

Bell-shaped distribution



Example 4: IQ scores have a bell-shaped distribution with a mean of 100 and a standard deviation of 15. What percentage of IQ scores are between 70 and 130?

Solution

The mean is

$$\mu = 100,$$

and the standard deviation is

$$\sigma = 15.$$

Two standard deviations are

$$2\sigma = 2(15) = 30.$$

The interval from two standard deviations below the mean to two standard deviations above the mean is

$$100 - 30 = 70$$

and

$$100 + 30 = 130.$$

By the Empirical Rule, about 95% of values lie within two standard deviations of the mean.
Therefore, about

$$\boxed{95\%}$$

of IQ scores are between 70 and 130.

10. Chebyshev's Theorem

Chebyshev's Theorem applies to any data set, even if the distribution is not bell-shaped.

Chebyshev's Theorem

The proportion of any set of data lying within K standard deviations of the mean is at least

$$1 - \frac{1}{K^2},$$

where $K > 1$.

Important special cases:

$$K = 2: \qquad 1 - \frac{1}{2^2} = \frac{3}{4} = 75\%,$$

$$K = 3: \qquad 1 - \frac{1}{3^2} = \frac{8}{9} \approx 89\%.$$

Example 5: IQ scores have a mean of 100 and a standard deviation of 15. What can we conclude from Chebyshev's Theorem?

Solution

For $K = 2$, two standard deviations are

$$2(15) = 30.$$

So the interval is

$$100 - 30 = 70$$

to

$$100 + 30 = 130.$$

Chebyshev's Theorem says that at least 75% of IQ scores are between 70 and 130.

For $K = 3$, three standard deviations are

$$3(15) = 45.$$

So the interval is

$$100 - 45 = 55$$

to

$$100 + 45 = 145.$$

Chebyshev's Theorem says that at least 89% of IQ scores are between 55 and 145.

11. Comparing Variation: Coefficient of Variation

The coefficient of variation, or CV, compares the standard deviation relative to the mean. It is useful when comparing variation between data sets with different units or very different means.

Coefficient of Variation

For sample data,

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

For population data,

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

Round-Off Rule for CV

Round the coefficient of variation to one decimal place, such as 25.3%.

12. Biased and Unbiased Estimators

Biased and Unbiased Estimators

- Biased Estimator: The sample standard deviation s is a biased estimator of the population standard deviation σ. Its values do not center around σ.
- Unbiased Estimator: The sample variance s^2 is an unbiased estimator of the population variance σ^2. Its values tend to center around σ^2, rather than systematically overestimating or underestimating the population variance.

Summary

Key Ideas

- The range is maximum $-$ minimum, but it is not resistant.
- The standard deviation measures typical variation from the mean.
- The sample standard deviation uses $n - 1$ in the denominator.
- The variance is the square of the standard deviation.
- The range rule of thumb helps identify significantly low or high values.

- The Empirical Rule applies to approximately bell-shaped distributions.
- Chebyshev's Theorem applies to any distribution.
- The coefficient of variation compares variation relative to the mean.