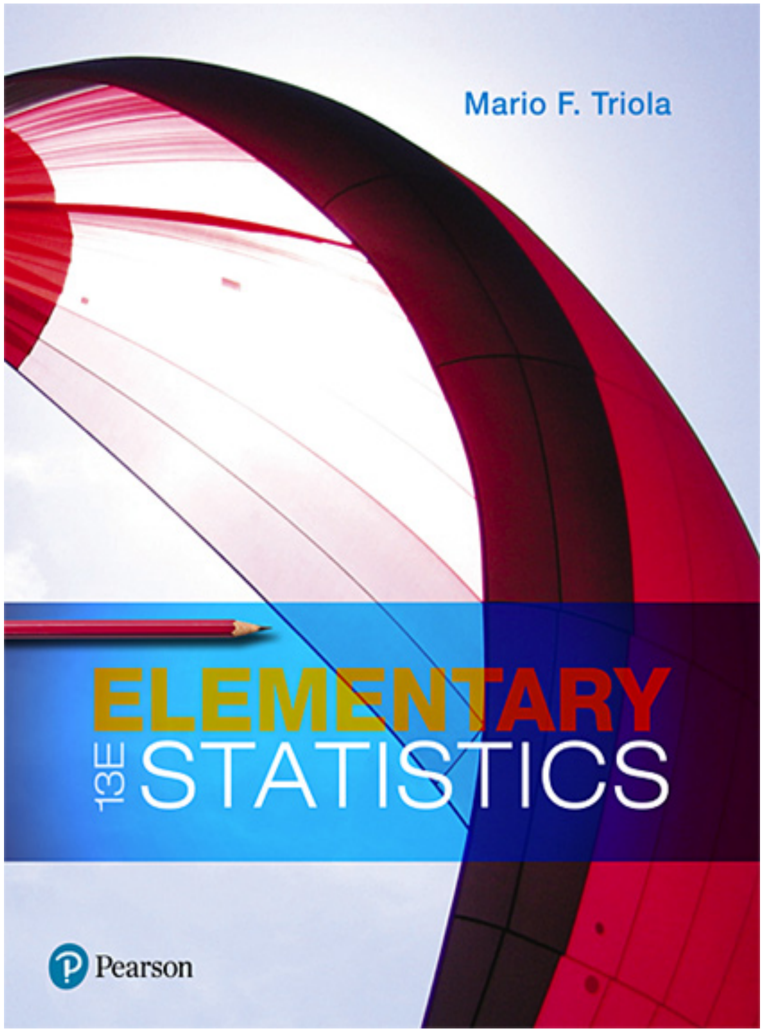
Elementary Statistics

Thirteenth Edition



Chapter 3

Describing, Exploring, and Comparing Data

Describing, Exploring, and Comparing Data

3-1 Measures of Center

3-2 Measures of Variation

3-3 Measures of Relative Standing and Boxplots

Key Concept

Variation is the single most important topic in statistics, so this is the single most important section in this book. This section presents three important measures of variation: **range, standard deviation,** and **variance.**

These statistics are numbers, but our focus is not just computing those numbers but developing the ability to **interpret** and **understand** them.

Round-off Rule for Measures of Variation

- Round-off Rule for Measures of Variation
  - When rounding the value of a measure of variation, carry one more decimal place than is present in the original set of data.

Range

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

Range = (maximum data value) − (minimum data value)

Important Property of Range

- The range uses only the maximum and the minimum data values, so it is very sensitive to extreme values. The range is not **resistant.**
- Because the range uses only the maximum and minimum values, it does not take every value into account and therefore does not truly reflect the variation among all of the data values.

Example: Range

Find the range of these Verizon data speeds (Mbps): 38.5, 55.6, 22.4, 14.1, 23.1.

Solution

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

$$= 55.6 - 14.1 = 41.50 \text{ Mbps}$$

Standard Deviation of a Sample (1 of 2)

- Standard Deviation
  - The **standard deviation** of a set of sample values, denoted by *s*, is a measure of how much data values deviate away from the mean.

Notation

s = **sample** standard deviation

σ = **population** standard deviation

Standard Deviation of a Sample (2 of 2)

- Standard Deviation

sample standard deviation

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

Shortcut formula for sample standard deviation (used by calculators and software)

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

Important Properties of Standard Deviation (1 of 2)

- The standard deviation is a measure of how much data values deviate away from the **mean**.
- The value of the standard deviation *s* is never negative. It is zero only when all of the data values are exactly the same.
- Larger values of *s* indicate greater amounts of variation.

Important Properties of Standard Deviation (2 of 2)

- The standard deviation *s* can increase dramatically with one or more outliers.
- The units of the standard deviation *s* (such as minutes, feet, pounds) are the same as the units of the original data values.
- The sample standard deviation *s* is a **biased estimator** of the population standard deviation *σ*, which means that values of the sample standard deviation *s* do not center around the value of *σ*.

Example: Calculating Standard Deviation

Use sample standard deviation formula to find the standard deviation of these Verizon data speed times (in Mbps): 38.5, 55.6, 22.4, 14.1, 23.1.

Solution

$$\bar{x} = 30.7 \quad \sum\left(x-\bar{x}\right)^2 = 1083.0520 \quad n-1 = 4$$

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

$$= \sqrt{\frac{1083.0520}{4}}$$

$$= \sqrt{270.7630}$$

$$= 16.45 \text{ Mbps}$$

Example: Calculating Standard Deviation Using Shortcut Formula

Find the standard deviation of the Verizon data speeds (Mbps) of 38.5, 55.6, 22.4, 14.1, 23.1

Solution

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

$$= \sqrt{\frac{5(5807.79) - (153.7)^2}{5(5-1)}}$$

$$= \sqrt{\frac{5415.26}{20}}$$

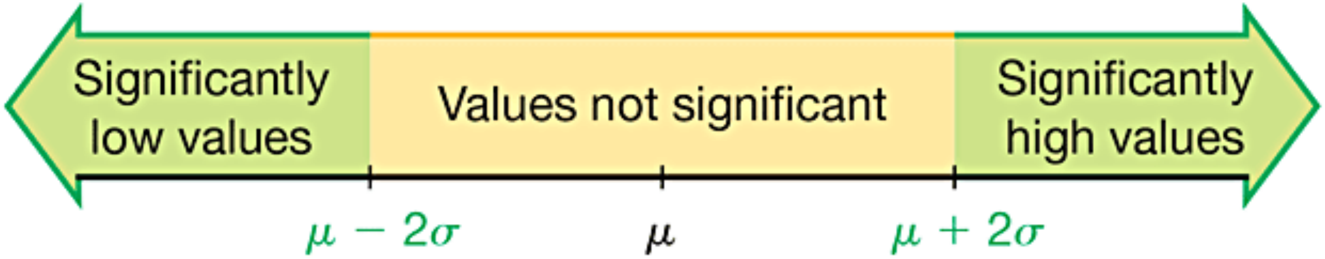
$$= 16.45 \text{ Mbps}$$

Range Rule of Thumb for Understanding Standard Deviation

- Range Rule of Thumb
  - The **range rule of thumb** is a crude but simple tool for understanding and interpreting standard deviation. The vast majority (such as 95%) of sample values lie within 2 standard deviations of the mean.

Range Rule of Thumb for Identifying Significant Values

- **Significantly low** values are $\mu - 2\sigma$ or lower.
- **Significantly high** values are $\mu + 2\sigma$ or higher.
- **Values not significant** are between $(\mu - 2\sigma)$ and $(\mu + 2\sigma)$.



Range Rule of Thumb for Estimating a Value of the Standard Deviation *s*

- Range Rule of Thumb for Estimating a Value of the Standard Deviation
  - To roughly estimate the standard deviation from a collection of known sample data, use

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

Standard Deviation of a Population

- Standard Deviation of a Population
  - A different formula is used to calculate the standard deviation *σ* of a **population**: Instead of dividing by *n* − 1 for a **sample**, we divide by the population size *N*.

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

Variance of a Sample and a Population

- Variance
  - The **variance** of a set of values is a measure of variation equal to the square of the standard deviation.
    - Sample variance: s^2 = square of the standard deviation *s.*
    - Population variance: σ^2 = square of the population standard deviation σ.

Notation Summary

s = **sample** standard deviation

s^2 = **sample** variance

σ = **population** standard deviation

σ^2 = **population** variance

Important Properties of Variance

- The units of the variance are the **squares** of the units of the original data values.
- The value of the variance can increase dramatically with the inclusion of outliers. (The variance is not **resistant.**)
- The value of the variance is never negative. It is zero only when all of the data values are the same number.
- The sample variance s^2 is an **unbiased estimator** of the population variance σ^2.

Why Divide by (n – 1)?

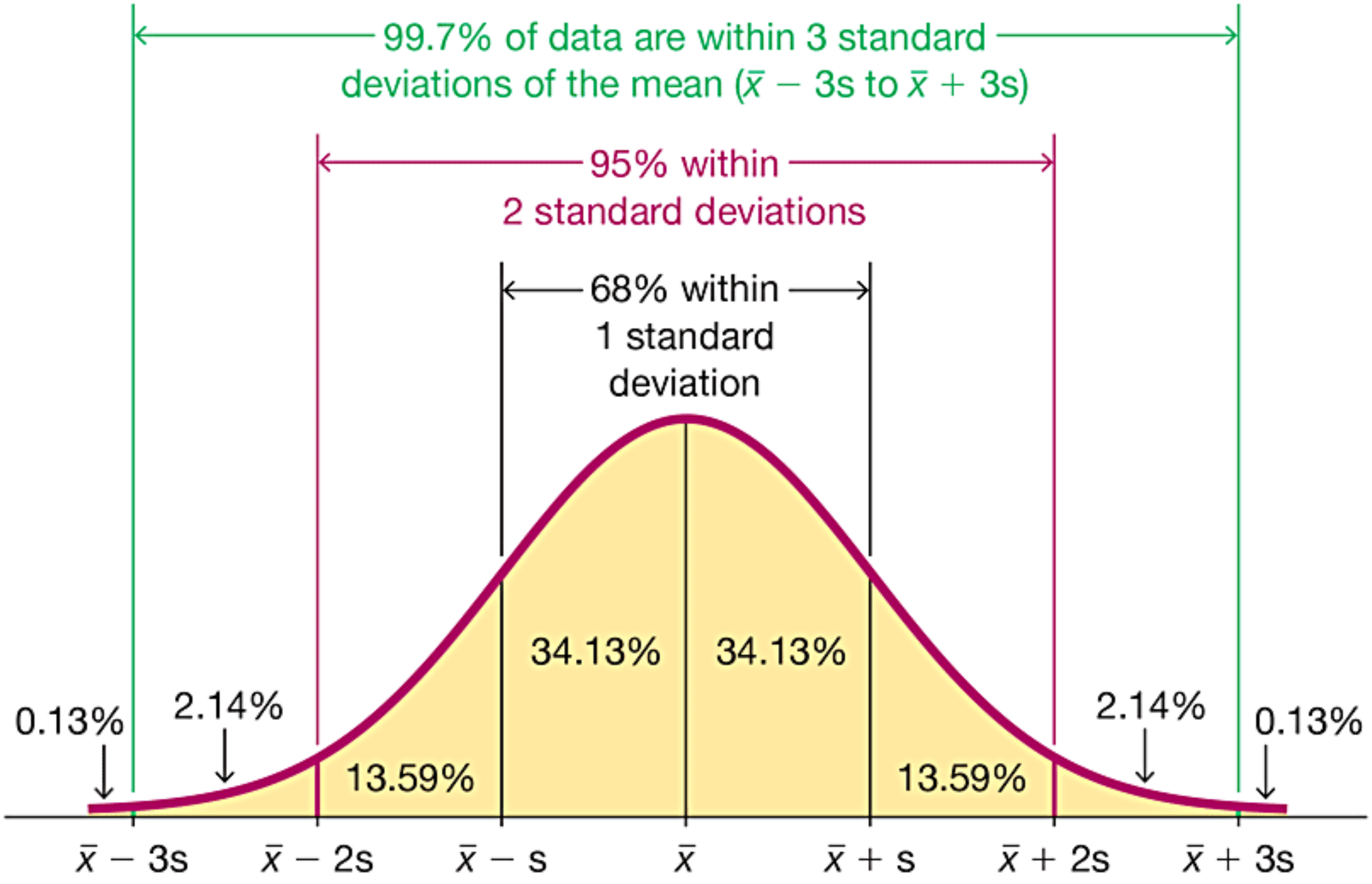
- There are only n − 1 values that can assigned without constraint. With a given mean, we can use any numbers for the first n − 1 values, but the last value will then be automatically determined.
- With division by n − 1, sample variances s^2 tend to center around the value of the population variance σ^2; with division by n, sample variances s^2 tend to **underestimate** the value of the population variance σ^2.

Empirical Rule for Data with a Bell-Shaped Distribution

The **empirical rule** states that **for data sets having a distribution that is approximately bell-shaped,** the following properties apply.

- 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.

The Empirical Rule



Example: The Empirical Rule (1 of 2)

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?

Example: The Empirical Rule (2 of 2)

Solution

The key is to recognize that 70 and 130 are each exactly 2 standard deviations away from the mean of 100.

2 standard deviations = $2s = 2(15) = 30$

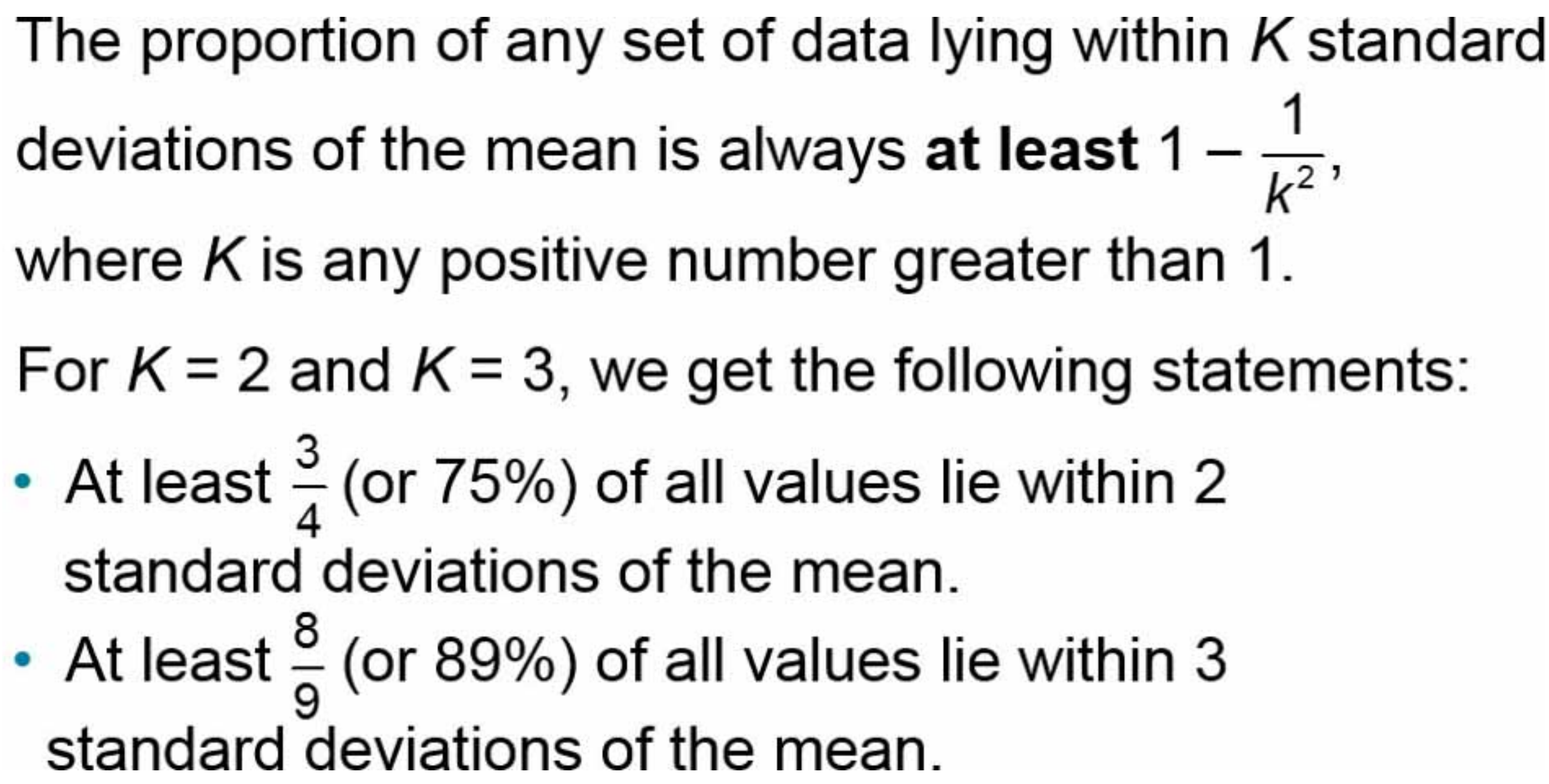
2 standard deviations from the mean is

$$100 - 30 = 70$$

$$\text{or } 100 + 30 = 130$$

About 95% of all IQ scores are between 70 and 130.

Chebyshev’s Theorem



The proportion of any set of data lying within K standard deviations of the mean is always **at least** $1 - \frac{1}{k^2}$, where K is any positive number greater than 1.

For $K = 2$ and $K = 3$, we get the following statements:

- At least $\frac{3}{4}$ (or 75%) of all values lie within 2 standard deviations of the mean.
- At least $\frac{8}{9}$ (or 89%) of all values lie within 3 standard deviations of the mean.

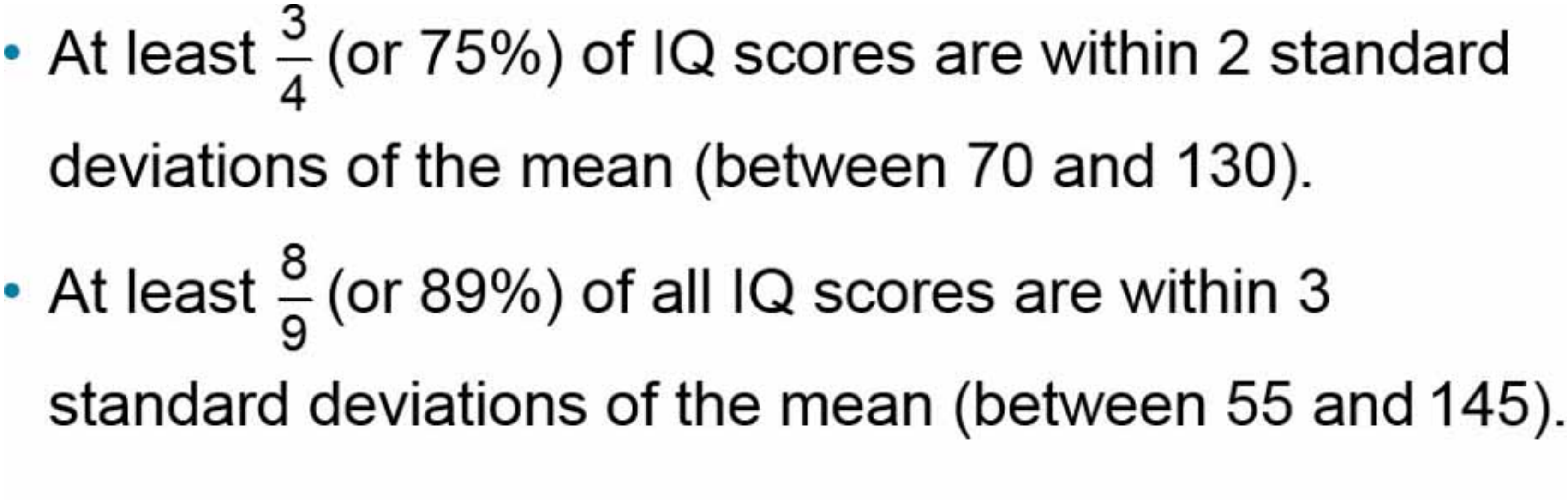
Example: Chebyshev’s Theorem (1 of 2)

IQ scores have a mean of 100 and a standard deviation of 15. What can we conclude from Chebyshev’s theorem?

Example: Chebyshev’s Theorem (2 of 2)

Solution

Applying Chebyshev’s theorem with a mean of 100 and a standard deviation of 15, we can reach the following conclusions:



- At least $\frac{3}{4}$ (or 75%) of IQ scores are within 2 standard deviations of the mean (between 70 and 130).
- At least $\frac{8}{9}$ (or 89%) of all IQ scores are within 3 standard deviations of the mean (between 55 and 145).

Comparing Variation in Different Samples or Populations

- Coefficient of Variation
  - The **coefficient of variation** (or **CV**) for a set of nonnegative sample or population data, expressed as a percent, describes the standard deviation relative to the mean, and is given by the following:

Sample

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

Population

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

Round-off Rule for the Coefficient of Variation

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

Biased and Unbiased Estimators

- The sample standard deviation s is a **biased estimator** of the population standard deviation s, which means that values of the sample standard deviation s do **not** tend to center around the value of the population standard deviation σ.
- The sample variance s^2 is an **unbiased estimator** of the population variance σ^2, which means that values of s^2 tend to center around the value of σ^2 instead of systematically tending to overestimate or underestimate σ^2.