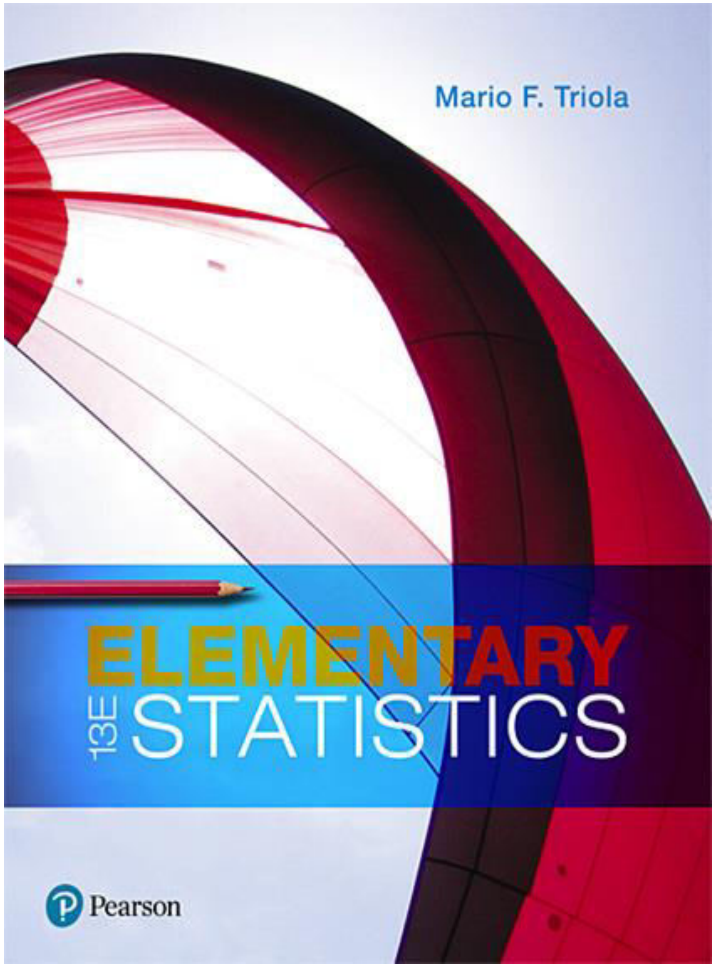
Elementary Statistics

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

This section introduces measures of relative standing, which are numbers showing the location of data values relative to the other values within the same data set.

The most important concept in this section is the *z* **score***.*

We also discuss percentiles and quartiles, which are common statistics, as well as another statistical graph called a boxplot.

z Scores

- *z* Score
  - A ***z* score** (or **standard score** or **standardized value**) is the number of standard deviations that a given value *x* is above or below the mean. The *z* score is calculated by using one of the following:

Sample **Population**

$$z = \frac{x - \bar{x}}{s} \quad \text{or} \quad z = \frac{x - \mu}{\sigma}$$

Round-off Rule for *z* Scores

Round *z* scores to two decimal places (such as 2.31).

Important Properties of *z* Scores

1. A *z* score is the number of standard deviations that a given value *x* is above or below the mean.
2. *z* scores are expressed as numbers with no units of measurement.
3. A data value is **significantly low** if its *z* score is less than or equal to −2 or the value is **significantly high** if its *z* score is greater than or equal to +2.
4. If an individual data value is less than the mean, its corresponding *z* score is a negative number.

Example: Comparing a Baby’s Weight and Adult Body Temperature (1 of 3)

Which of the following two data values is more extreme relative to the data set from which it came?

- The 4000 g weight of a newborn baby (among 400 weights with sample mean $\bar{x}$ = 3152.0 g and sample standard deviation s = 693.4 g)
- The 99°F temperature of an adult (among 106 adults with sample mean $\bar{x}$ = 98.20°F and sample standard deviation s = 0.62°F)

Example: Comparing a Baby's Weight and Adult Body Temperature (2 of 3)

Solution

The 4000 g weight and the 99°F body temperature can be standardized by converting each of them to *z* scores.

- 4000 g birth weight:

$$z = \frac{x - \bar{x}}{s} = \frac{4000\text{ g} - 3152.0\text{ g}}{693.4\text{ g}} = 1.22$$

- 99°F body temperature:

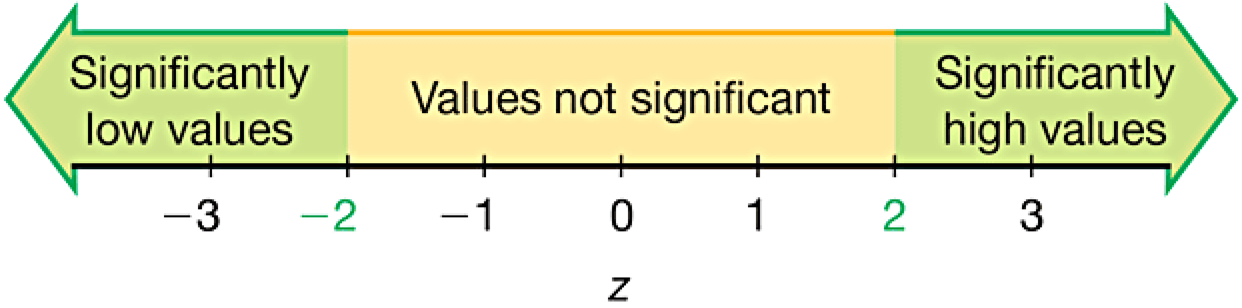
$$z = \frac{x - \bar{x}}{s} = \frac{99°\text{F} - 398.20°\text{F}}{0.62°\text{F}} = 1.29$$

Example: Comparing a Baby’s Weight and Adult Body Temperature (3 of 3)

Interpretation

- The *z* scores show that the 4000 g birth weight is 1.22 standard deviations above the mean, and the 99°F body temperature is 1.29 standard deviations above the mean.
- Because the body temperature is farther above the mean, it is the more extreme value. A 99°F body temperature is slightly more extreme than a birth weight of 4000 g.

Using *z* Scores to Identify Significant Values



Significant values are those with

z scores ≤ −2.00 or ≥ 2.00.

Example: Is a Platelet Count of 75 Significantly Low? (1 of 3)

The lowest platelet count in a dataset is 75. (Platelet counts are measured in 1000 cells/μL). Is that value significantly low? Assume that platelet counts have a mean of $\bar{x} = 239.4$ and a standard deviation of $s = 64.2$.

Example: Is a Platelet Count of 75 Significantly Low? (2 of 3)

Solution

The platelet count of 75 is converted to a *z* score as shown below:

$$z = \frac{x - \bar{x}}{s} = \frac{75 - 239.4}{64.2} = -2.56$$

Example: Is a Platelet Count of 75 Significantly Low? (3 of 3)

Interpretation

The platelet count of 75 converts to the z score of −2.56. $z = -2.56$ is less than −2, so the platelet count of 75 is significantly low. (Low platelet counts are called thrombocytopenia, not for the lack of a better term.)

Percentiles

- Percentiles
  - **Percentiles** are measures of location, denoted P_1, P_2, . . . , P_{99}, which divide a set of data into 100 groups with about 1% of the values in each group.

Finding the Percentile of a Data Value

The process of finding the percentile that corresponds to a particular data value *x* is given by the following (round the result to the nearest whole number):

$$\text{Percentile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

Example: Finding a Percentile (1 of 3)

The airport Verizon cell phone data speeds listed below are arranged in increasing order. Find the percentile for the data speed of 11.8 Mbps.

0.8	1.4	1.8	1.9	3.2	3.6	4.5	4.5	4.6	6.2
6.5	7.7	7.9	9.9	10.2	10.3	10.9	11.1	11.1	11.6
11.8	12.0	13.1	13.5	13.7	14.1	14.2	14.7	15.0	15.1
15.5	15.8	16.0	17.5	18.2	20.2	21.1	21.5	22.2	22.4
23.1	24.5	25.7	28.5	34.6	38.5	43.0	55.6	71.3	77.8

Example: Finding a Percentile (2 of 3)

Solution

From the sorted list of airport data speeds in the table, we see that there are 20 data speeds less than 11.8 Mbps, so

$$\text{Percentile of value } 11.8 = \frac{20}{50} \cdot 100 = 40$$

Example: Finding a Percentile (3 of 3)

Interpretation

A data speed of 11.8 Mbps is in the 40th percentile. This can be interpreted loosely as this:

A data speed of 11.8 Mbps separates the lowest 40% of values from the highest 60% of values. We have P_{40} = 11.8 Mbps.

Notation

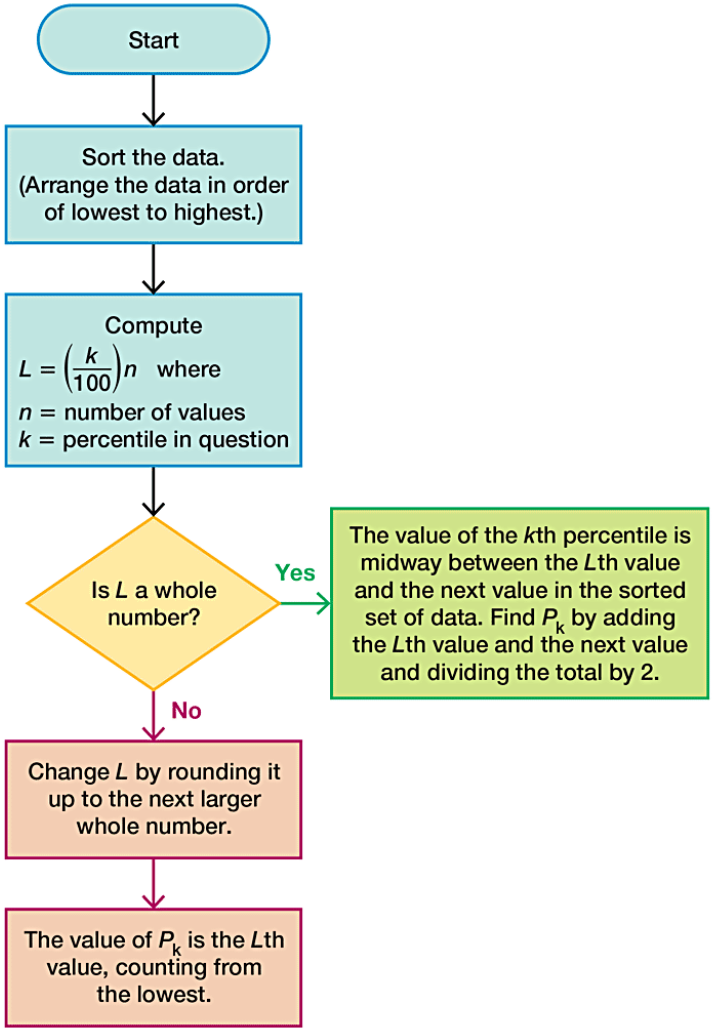
n total number of values in the data set

k percentile being used (Example: For the 25th percentile, *k* = 25.)

L locator that gives the **position** of a value (Example: For the 12th value in the sorted list, *L* = 12.)

P_k *k*th percentile (Example: P_{25} is the 25th percentile.)

Converting a Percentile to a Data Value



Start

Sort the data. (Arrange the data in order of lowest to highest.)

Compute $L = \left(\frac{k}{100}\right)n$ where

n = number of values

k = percentile in question

Is L a whole number?

Yes

The value of the kth percentile is midway between the Lth value and the next value in the sorted set of data. Find P_k by adding the Lth value and the next value and dividing the total by 2.

No

Change L by rounding it up to the next larger whole number.

The value of P_k is the Lth value, counting from the lowest.

Example: Converting a Percentile to a Data Value (1 of 4)

Refer to the sorted data speeds below. Find the 40th percentile, denoted by P_{40}.

0.8	1.4	1.8	1.9	3.2	3.6	4.5	4.5	4.6	6.2
6.5	7.7	7.9	9.9	10.2	10.3	10.9	11.1	11.1	11.6
11.8	12.0	13.1	13.5	13.7	14.1	14.2	14.7	15.0	15.1
15.5	15.8	16.0	17.5	18.2	20.2	21.1	21.5	22.2	22.4
23.1	24.5	25.7	28.5	34.6	38.5	43.0	55.6	71.3	77.8

Example: Converting a Percentile to a Data Value (2 of 4)

Solution

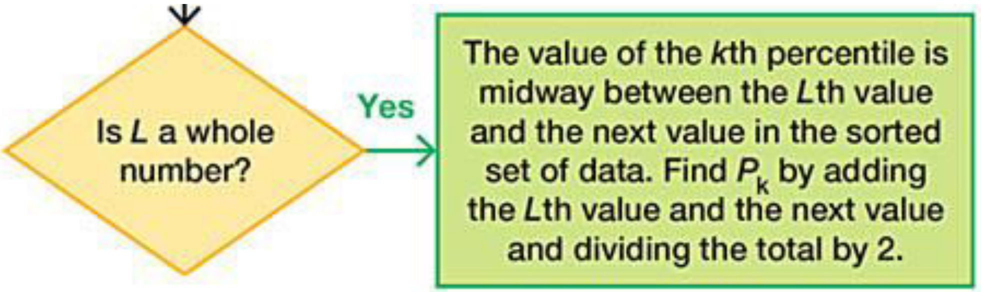
We can proceed to compute the value of the locator *L*. In this computation, we use *k* = 40 because we are attempting to find the value of the 40th percentile, and we use *n* = 50 because there are 50 data values.

$$L = \frac{k}{100} \cdot n = \frac{40}{100} \cdot 50 = 20$$

Example: Converting a Percentile to a Data Value (3 of 4)

Solution

Since *L* = 20 is a whole number, we proceed to the box located at the right.

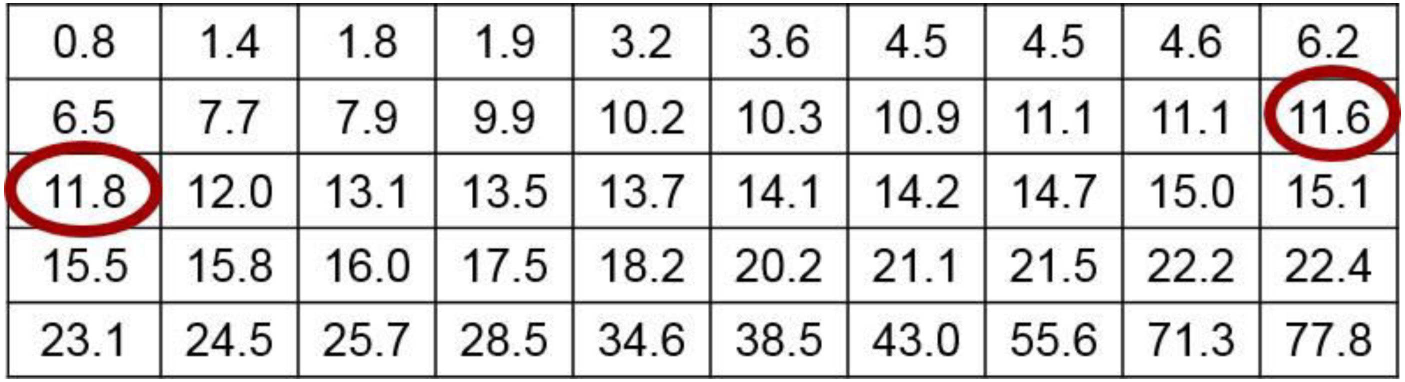

Is *L* a whole number?
Yes
The value of the *k*th percentile is midway between the *L*th value and the next value in the sorted set of data. Find P_k by adding the *L*th value and the next value and dividing the total by 2.

We now see that the value of the 40th percentile is midway between the *L*th (20th) value and the next value in the original set of data. That is, the value of the 40th percentile is midway between the 20th value and the 21st value.

Example: Converting a Percentile to a Data Value (4 of 4)

Solution

The 20th value in the table is 11.6 and the 21st value is 11.8, so the value midway between them is 11.7 Mbps. We conclude that the 40th percentile is P_{40} = 11.7 Mbps.



0.8	1.4	1.8	1.9	3.2	3.6	4.5	4.5	4.6	6.2
6.5	7.7	7.9	9.9	10.2	10.3	10.9	11.1	11.1	11.6
11.8	12.0	13.1	13.5	13.7	14.1	14.2	14.7	15.0	15.1
15.5	15.8	16.0	17.5	18.2	20.2	21.1	21.5	22.2	22.4
23.1	24.5	25.7	28.5	34.6	38.5	43.0	55.6	71.3	77.8

Quartiles

- Quartiles
  - **Quartiles** are measures of location, denoted Q_1, Q_2, and Q_3, which divide a set of data into four groups with about 25% of the values in each group.

Descriptions of Quartiles (1 of 2)

- Q_1 (First quartile):
  - Same value as P_{25}. It separates the bottom 25% of the sorted values from the top 75%.
- Q_2 (Second quartile):
  - Same as P_{50} and same as the median. It separates the bottom 50% of the sorted values from the top 50%.

Descriptions of Quartiles (2 of 2)

- Q_3 (Third quartile):
  - Same as P_{75}. It separates the bottom 75% of the sorted values from the top 25%.

Caution Just as there is not universal agreement on a procedure for finding percentiles, there is not universal agreement on a single procedure for calculating quartiles, and different technologies often yield different results.

Statistics defined using quartiles and percentiles

Interquartile range (or IQR) $= Q_3 - Q_1$

Semi-interquartile range $= \dfrac{Q_3 - Q_1}{2}$

Midquartile range $= \dfrac{Q_3 + Q_1}{2}$

10 – 90 quartile range $= P_{90} - P_{10}$

5-Number Summary

- 5-Number Summary
  - For a set of data, the **5-number summary** consists of these five values:

1. Minimum
2. First quartile, Q_1
3. Second quartile, Q_2 (same as the median)
4. Third quartile, Q_3
5. Maximum

Example: Finding a 5-Number Summary (1 of 3)

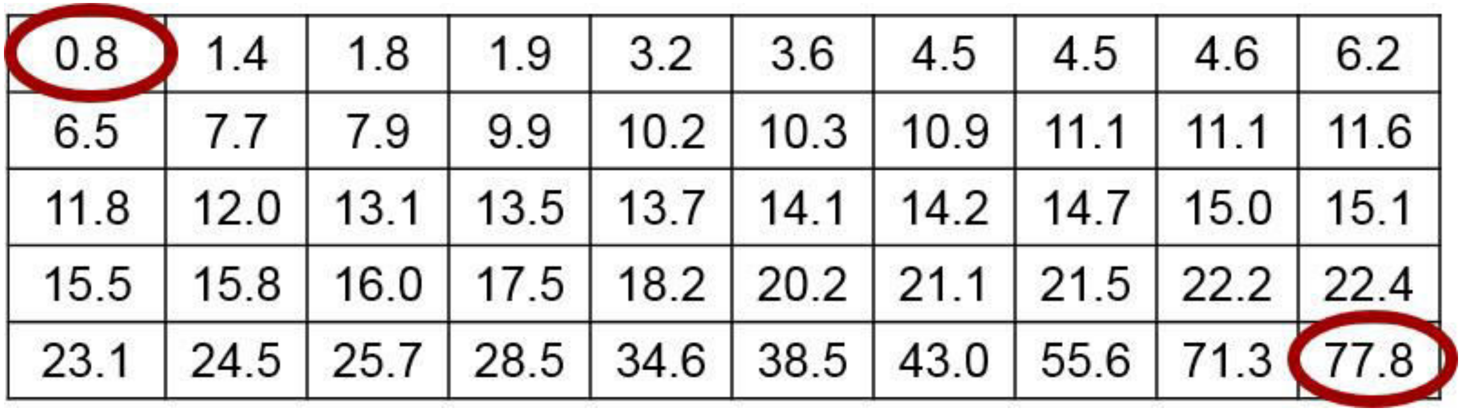
Use the Verizon airport data speeds to find the 5-number summary.

0.8	1.4	1.8	1.9	3.2	3.6	4.5	4.5	4.6	6.2
6.5	7.7	7.9	9.9	10.2	10.3	10.9	11.1	11.1	11.6
11.8	12.0	13.1	13.5	13.7	14.1	14.2	14.7	15.0	15.1
15.5	15.8	16.0	17.5	18.2	20.2	21.1	21.5	22.2	22.4
23.1	24.5	25.7	28.5	34.6	38.5	43.0	55.6	71.3	77.8

Example: Finding a 5-Number Summary (2 of 3)

Solution

Because the Verizon airport data speeds are sorted, it is easy to see that the minimum is 0.8 Mbps and the maximum is 77.8 Mbps.

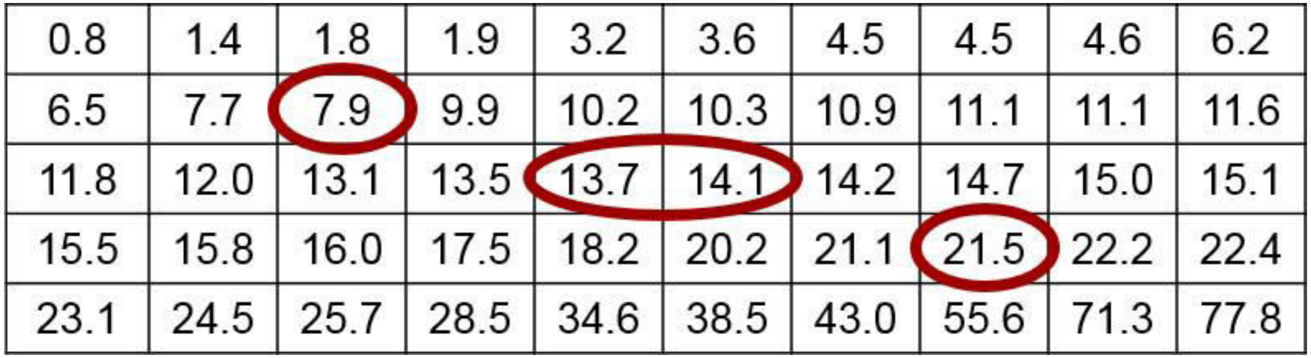


0.8	1.4	1.8	1.9	3.2	3.6	4.5	4.5	4.6	6.2
6.5	7.7	7.9	9.9	10.2	10.3	10.9	11.1	11.1	11.6
11.8	12.0	13.1	13.5	13.7	14.1	14.2	14.7	15.0	15.1
15.5	15.8	16.0	17.5	18.2	20.2	21.1	21.5	22.2	22.4
23.1	24.5	25.7	28.5	34.6	38.5	43.0	55.6	71.3	77.8

Example: Finding a 5-Number Summary (3 of 3)

Solution

The value of the first quartile is Q_1 = 7.9 Mbps. The median is equal to Q_2, and it is 13.9 Mbps. Also, we can find that Q_3 = 21.5 Mbps by using the same procedure for finding P_{75}.



0.8	1.4	1.8	1.9	3.2	3.6	4.5	4.5	4.6	6.2
6.5	7.7	7.9	9.9	10.2	10.3	10.9	11.1	11.1	11.6
11.8	12.0	13.1	13.5	13.7	14.1	14.2	14.7	15.0	15.1
15.5	15.8	16.0	17.5	18.2	20.2	21.1	21.5	22.2	22.4
23.1	24.5	25.7	28.5	34.6	38.5	43.0	55.6	71.3	77.8

The 5-number summary is therefore 0.8, 7.9, 13.9, 21.5, and 77.8 (all in units of Mbps).

Boxplot (or Box-and-Whisker Diagram)

- Boxplot (or Box-and-Whisker Diagram)
  - A **boxplot** (or **box-and-whisker diagram**) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile Q_1, the median, and the third quartile Q_3.

Procedure for Constructing a Boxplot

1. Find the 5-number summary (minimum value, Q_1, Q_2, Q_3, maximum value).
2. Construct a line segment extending from the minimum data value to the maximum data value.
3. Construct a box (rectangle) extending from Q_1 to Q_3, and draw a line in the box at the value of Q_2 (median).

Example: Constructing a Boxplot (1 of 2)

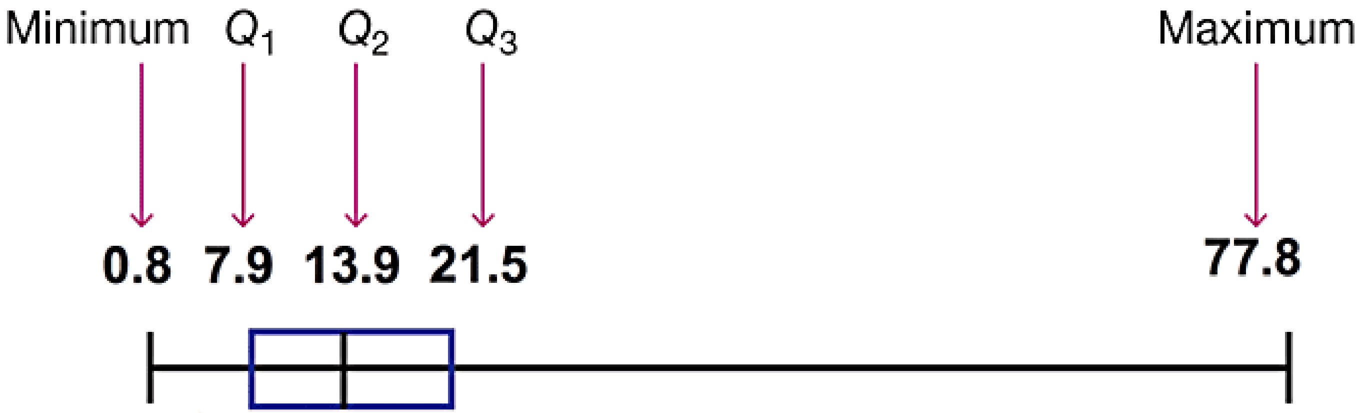
Use the Verizon airport data speeds to construct a boxplot.

0.8	1.4	1.8	1.9	3.2	3.6	4.5	4.5	4.6	6.2
6.5	7.7	7.9	9.9	10.2	10.3	10.9	11.1	11.1	11.6
11.8	12.0	13.1	13.5	13.7	14.1	14.2	14.7	15.0	15.1
15.5	15.8	16.0	17.5	18.2	20.2	21.1	21.5	22.2	22.4
23.1	24.5	25.7	28.5	34.6	38.5	43.0	55.6	71.3	77.8

Example: Constructing a Boxplot (2 of 2)

Solution

The boxplot uses the 5-number summary found in the previous example: 0.8, 7.9, 13.9, 21.5, and 77.8 (all in units of Mbps). Below is the boxplot representing the Verizon airport data speeds.



Skewness

- Skewness
  - A boxplot can often be used to identify skewness. A distribution of data is **skewed** if it is not symmetric and extends more to one side than to the other.

Identifying Outliers for Modified Boxplots

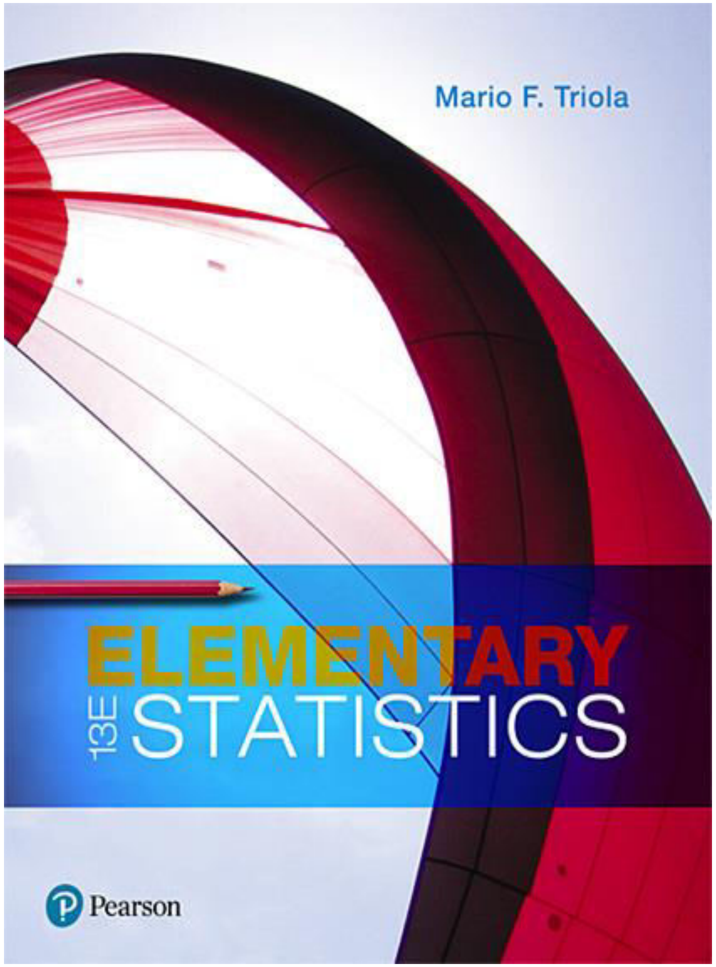
1. Find the quartiles Q_1, Q_2, and Q_3.
2. Find the interquartile range (IQR), where IQR = $Q_3 - Q_1$.
3. Evaluate 1.5 × IQR.
4. In a modified boxplot, a data value is an **outlier** if it is above Q_3, by an amount greater than 1.5 × IQR or below Q_1, by an amount greater than 1.5 × IQR.

Modified Boxplots

- Modified Boxplots
  - A **modified boxplot** is a regular boxplot constructed with these modifications:
    1. A special symbol (such as an asterisk or point) is used to identify outliers as defined above, and
    2. the solid horizontal line extends only as far as the minimum data value that is not an outlier and the maximum data value that is not an outlier.

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

Range = (maximum value) − (minimum value)

= 55.6 − 14.1 = 41.50 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

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

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

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

$$= \sqrt{270.7630}$$

$$= 16.45 \text{ Mbps}$$

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

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

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

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

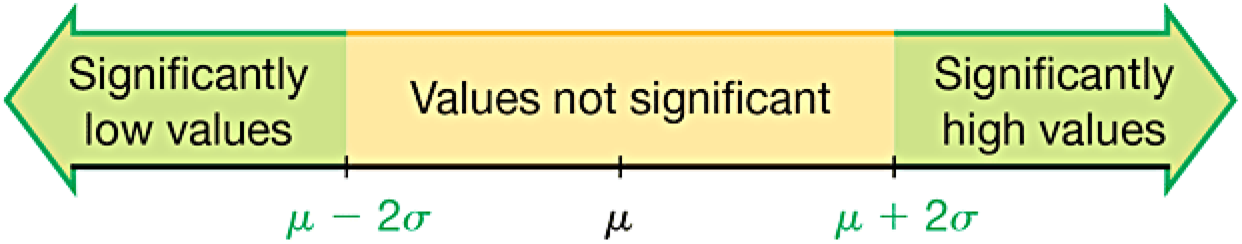
$$= 16.45 \text{ Mbps}$$

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)?

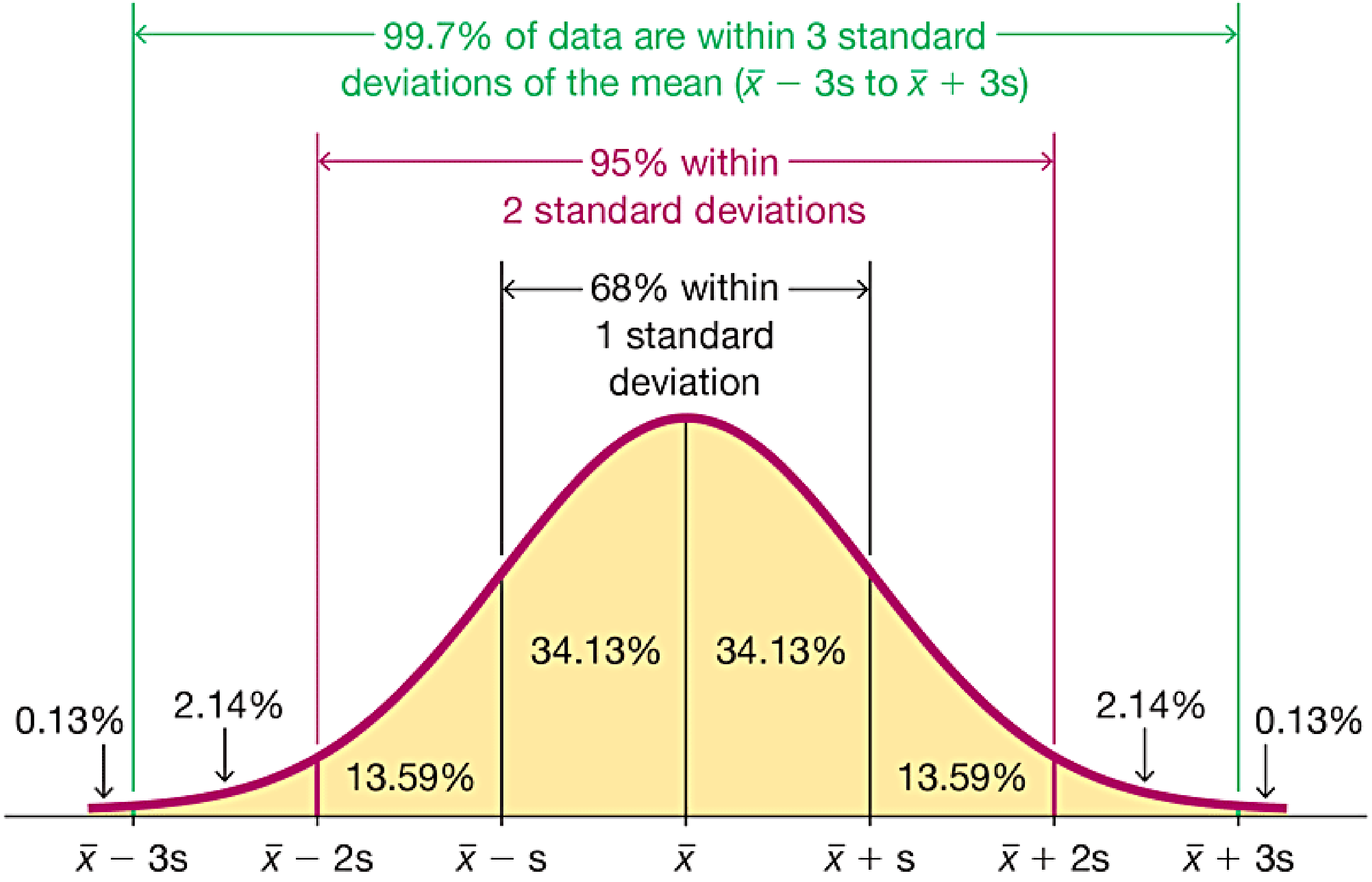
- There are only $n - 1$ values that can assigned without constraint. With a given mean, we can use any numbers for the first $n - 1$ values, but the last value will then be automatically determined.
- With division by $n - 1$, sample variances s^2 tend to center around the value of the population variance 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 = 2*s* = 2(15) = 30

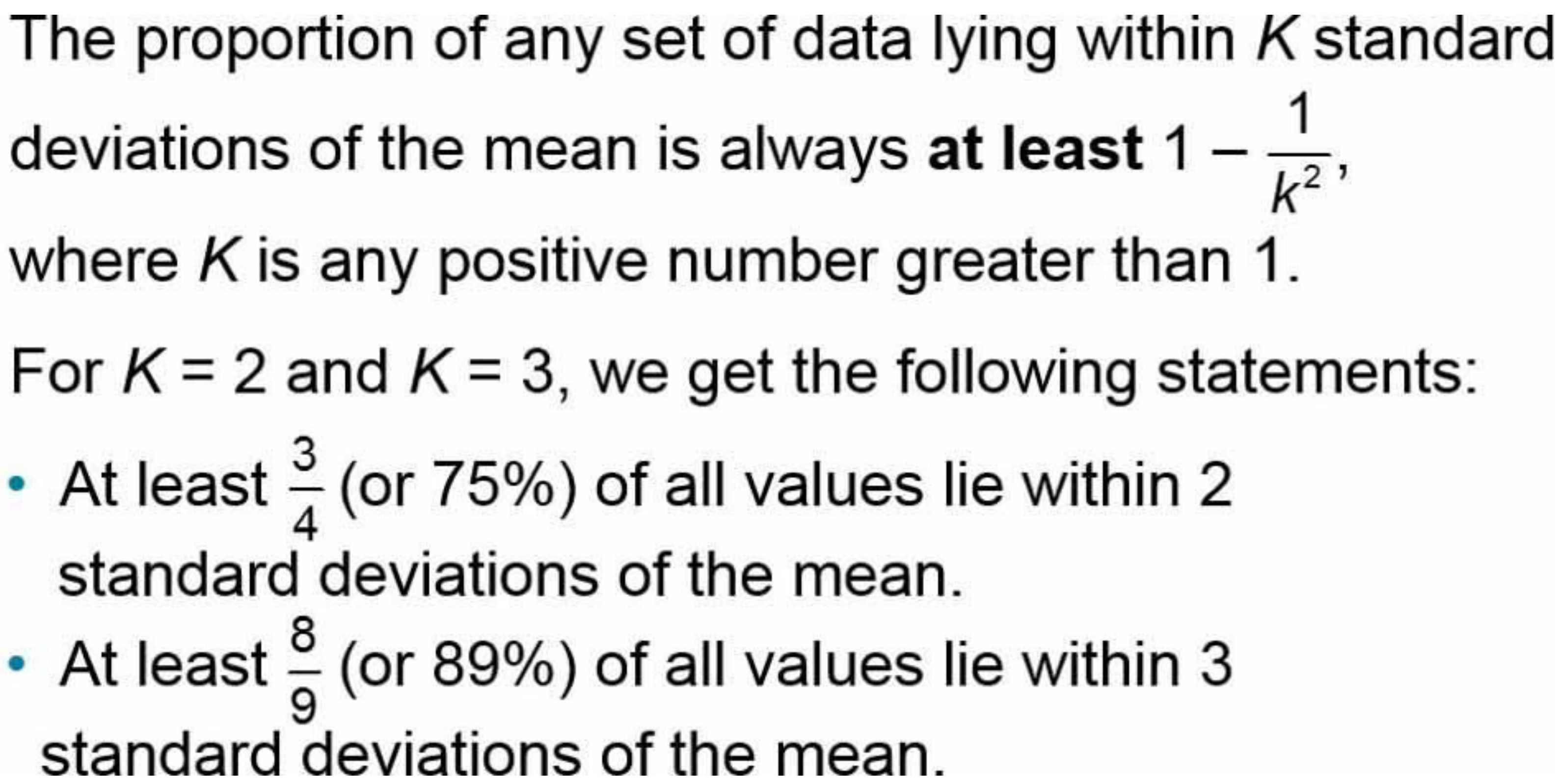
2 standard deviations from the mean is

$$100 - 30 = 70$$

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

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.

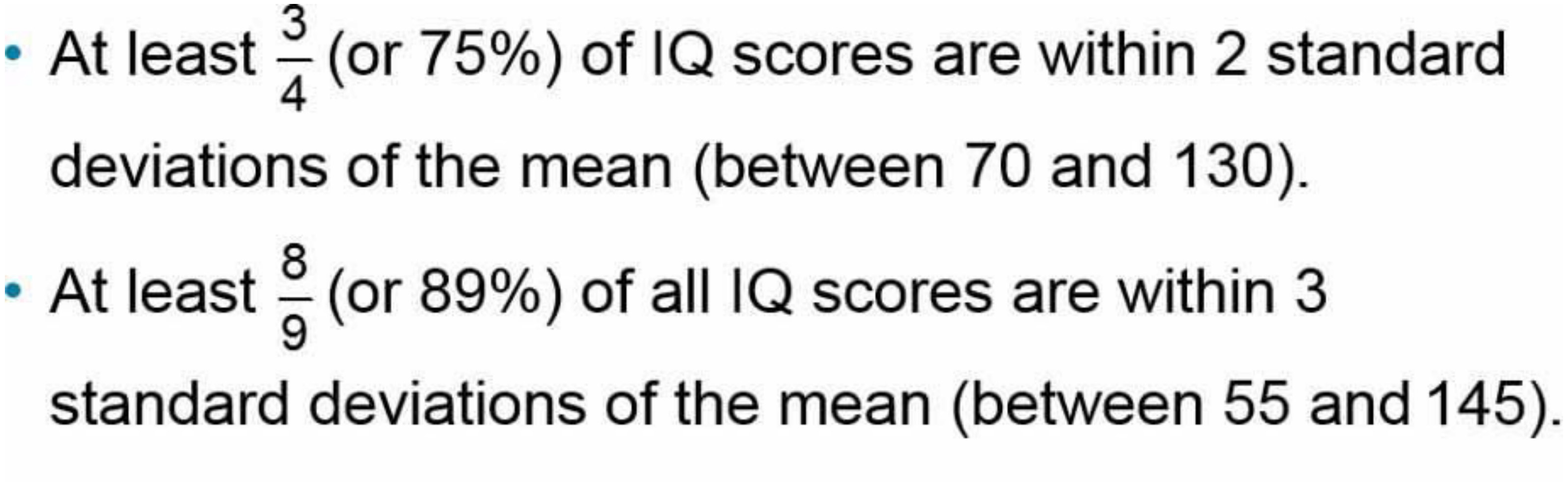
Example: Chebyshev’s Theorem (1 of 2)

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

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

Population

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

Round-off Rule for the Coefficient of Variation

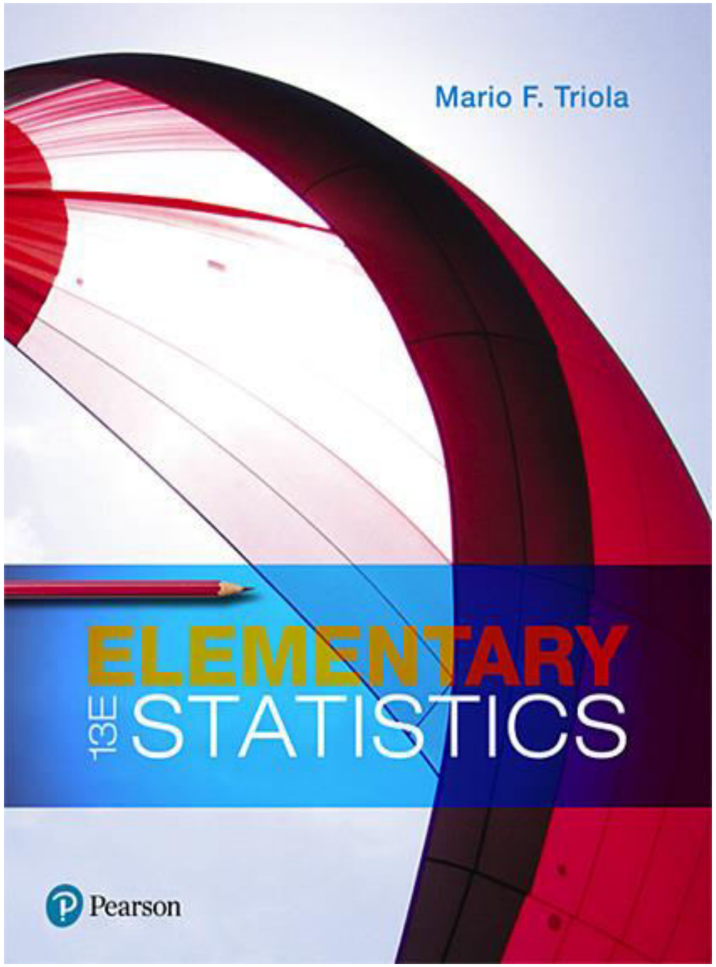
Round the coefficient of variation to one decimal place (such as 25.3%).

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.

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

The focus of this section is to obtain a value that measures the **center** of a data set. In particular, we present measures of center, including **mean** and **median**. Our objective here is not only to find the value of each measure of center, but also to interpret those values.

Measure of Center

- Measure of Center
  - A **measure of center** is a value at the center or middle of a data set.

Mean (or Arithmetic Mean)

- Mean (or Arithmetic Mean)
  - The **mean** (or **arithmetic mean**) of a set of data is the measure of center found by adding all of the data values and dividing the total by the number of data values.

Important Properties of the Mean

- Sample means drawn from the same population tend to vary less than other measures of center.
- The mean of a data set uses every data value.
- A disadvantage of the mean is that just one extreme value (outlier) can change the value of the mean substantially. (Using the following definition, we say that the mean is not **resistant.**)

Resistant

- Resistant
  - A statistic is **resistant** if the presence of extreme values (outliers) does not cause it to change very much.

Notation (1 of 2)

$\sum$ denotes the **sum** of a set of data values.

x is the variable usually used to represent the individual data values.

n represents the number of data values in a **sample.**

N represents the number of data values in a **population.**

Notation (2 of 2)

$\bar{x}$ is pronounced “x-bar” and is the mean of a set of **sample** values.

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

μ is pronounced “mu” and is the mean of all values in a **population**.

$$\mu = \frac{\Sigma x}{N}$$

Example: Mean (1 of 2)

Data Set 32 “Airport Data Speeds” in Appendix B includes measures of data speeds of smartphones from four different carriers. Find the mean of the first five data speeds for Verizon: 38.5, 55.6, 22.4, 14.1, and 23.1 (all in megabits per second, or Mbps).

Example: Mean (2 of 2)

Solution

$$\bar{x} = \frac{\Sigma x}{n} = \frac{38.5+55.6+22.4+14.1+23.1}{5}$$

$$= \frac{153.7}{5}$$

$$= 30.74 \text{ Mbps}$$

Mean

- Caution
  - Never use the term **average** when referring to a measure of center. The word **average** is often used for the mean, but it is sometimes used for other measures of center.
- The term **average** is not used by statisticians.
- The term **average** is not used by the statistics community or professional journals.

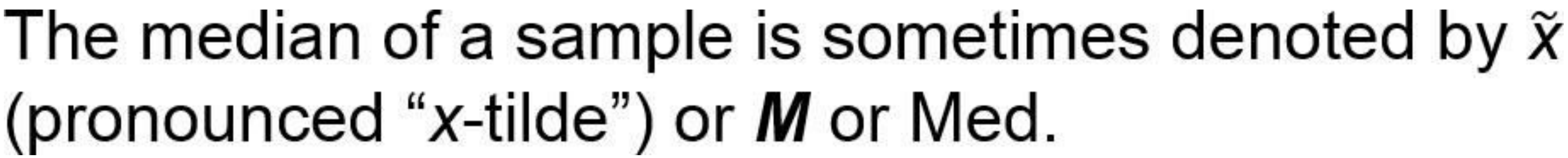
Median

- Median
  - The **median** of a data set is the measure of center that is the **middle value** when the original data values are arranged in order of increasing (or decreasing) magnitude.

Important Properties of the Median

- The median does not change by large amounts when we include just a few extreme values, so the median is a **resistant** measure of center.
- The median does not directly use every data value. (For example, if the largest value is changed to a much larger value, the median does not change.)

Calculation and Notation of the Median



The median of a sample is sometimes denoted by $\tilde{x}$ (pronounced "*x*-tilde") or ***M*** or Med.

To find the median, first **sort** the values (arrange them in order) and then follow one of these two procedures:

1. If the number of data values is **odd**, the median is the number located in the exact middle of the sorted list.
2. If the number of data values is **even,** the median is found by computing the mean of the two middle numbers in the sorted list.

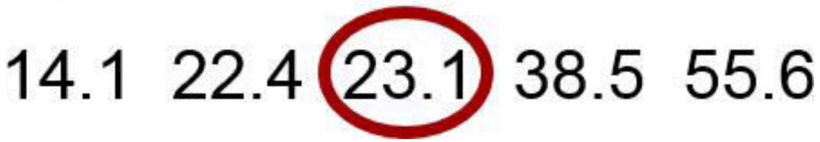
Example: Median with an Odd Number of Data Values (1 of 2)

Find the median of the first five data speeds for Verizon: 38.5, 55.6, 22.4, 14.1, and 23.1 (all in megabits per second, or Mbps).

Example: Median with an Odd Number of Data Values (2 of 2)

Solution

First sort the data values by arranging them in ascending order, as shown below:



14.1 22.4 23.1 38.5 55.6

Because there are 5 data values, the number of data values is an odd number (5), so the median is the number located in the exact middle of the sorted list, which is 23.1 Mbps.

Example: Median with an Even Number of Data Values (1 of 2)

Repeat of the previous example after including the sixth data speed of 24.5 Mbps. That is, find the median of these data speeds: 38.5, 55.6, 22.4, 14.1, 23.1, 24.5 (all in Mbps).

Example: Median with an Even Number of Data Values (2 of 2)

Solution

First arrange the values in ascending order:

14.1 22.4 23.1 24.5 38.5 55.6

Because the number of data values is an even number (6), the median is found by computing the mean of the two middle numbers, which are 23.1 and 24.5.

$$\text{Median} = \frac{23.1 + 24.5}{2} = \frac{47.6}{2} = 23.80 \text{ Mbps}$$

Mode

- Mode
  - The **mode** of a data set is the value(s) that occur(s) with the greatest frequency.

Important Properties of the Mode

- The mode can be found with qualitative data.
- A data set can have no mode or one mode or multiple modes.

Finding the Mode

A data set can have one mode, more than one mode, or no mode.

- When two data values occur with the same greatest frequency, each one is a mode and the data set is said to be **bimodal**.
- When more than two data values occur with the same greatest frequency, each is a mode and the data set is said to be **multimodal**.
- When no data value is repeated, we say that there is **no mode**.

Example: Mode

Find the mode of these Sprint data speeds (in Mbps):

0.2 0.3 0.3 0.3 0.6 0.6 1.2

Solution

The mode is 0.3 Mbps, because it is the data speed occurring most often (three times).

Other Mode Examples

Two modes: The data speeds (Mbps) of 0.3, 0.3, 0.6, 4.0, and 4.0 have two modes: 0.3 Mbps and 4.0 Mbps.

No mode: The data speeds (Mbps) of 0.3, 1.1, 2.4, 4.0, and 5.0 have no mode because no value is repeated.

Midrange

- Midrange
  - The **midrange** of a data set is the measure of center that is the value midway between the maximum and minimum values in the original data set. It is found by adding the maximum data value to the minimum data value and then dividing the sum by 2, as in the following formula:

$$\text{Midrange} = \frac{\text{maximum data value} + \text{minimum data value}}{2}$$

Important Properties of the Midrange (1 of 2)

Because the midrange uses only the maximum and minimum values, it is very sensitive to those extremes so the midrange is not **resistant.**

Important Properties of the Midrange (2 of 2)

- In practice, the midrange is rarely used, but it has three redeeming features:
  1. The midrange is very easy to compute.
  2. The midrange helps reinforce the very important point that there are several different ways to define the center of a data set.
  3. The value of the midrange is sometimes used incorrectly for the median, so confusion can be reduced by clearly defining the midrange along with the median.

Example: Midrange

Find the midrange of these Verizon data speeds: 38.5, 55.6, 22.4, 14.1, and 23.1 (all in Mbps)

Solution

The midrange is found as follows:

$$
\begin{aligned}
\text{Midrange} &= \frac{\text{maximum data value} + \text{minimum data value}}{2} \\
&= \frac{55.6 + 14.1}{2} \\
&= 34.85 \text{ Mbps}
\end{aligned}
$$

Round-Off Rules for Measures of Center

- For the mean, median, and midrange, carry one more decimal place than is present in the original set of values.
- For the mode, leave the value as is without rounding (because values of the mode are the same as some of the original data values).

Critical Thinking

- We can always calculate measures of center from a sample of numbers, but we should always think about whether it makes sense to do that.
- We should also think about the sampling method used to collect the data.

Example: Critical Thinking and Measures of Center (1 of 2)

See each of the following illustrating situations in which the mean and median are **not** meaningful statistics.

a. Zip codes of the Gateway Arch in St. Louis, White House, Air Force division of the Pentagon, Empire State Building, and Statue of Liberty: 63102, 20500, 20330, 10118, 10004. The zip codes don’t measure or count anything. The numbers are just labels for geographic locations.

Example: Critical Thinking and Measures of Center (2 of 2)

See each of the following illustrating situations in which the mean and median are **not** meaningful statistics.

b. Ranks of selected national universities of Harvard, Yale, Duke, Dartmouth, and Brown (from **U.S. News & World Report**): 2, 3, 7, 10, 14. The ranks reflect an ordering, but they don’t measure or count anything.

Calculating the Mean from a Frequency Distribution

- Mean from a Frequency Distribution
  - First multiply each frequency and class midpoint; then add the products.

$$\bar{x} = \frac{\sum (f \cdot x)}{\sum f}$$

(Result is an approximation.)

Sum of frequencies (equal to *n*)

Example: Computing the Mean from a Frequency Distribution (1 of 2)

The first two columns of the table shown here are the same as the frequency distribution of Table 2-2 from Chapter 2. Use the frequency distribution in the first two columns to find the mean.

Time (seconds)	**Frequency f**	**Class Midpoint x**	**$f \cdot x$**
75 – 124	11	99.5	1094.5
125 – 174	24	149.5	3588.0
175 – 224	10	199.5	1995.0
225 – 274	3	249.5	748.5
275 – 324	2	299.5	599.0
Totals:	$\sum f = 50$		$\sum (f \cdot x) = 8025.0$

Example: Computing the Mean from a Frequency Distribution (2 of 2)

Solution

When working with data summarized in a frequency distribution, we make calculations possible by pretending that all sample values in each class are equal to the class midpoint.

$$\bar{x} = \frac{\sum(f \cdot x)}{\sum f} = \frac{8025.0}{50} = 160.5 \text{ seconds}$$

The result of x = 160.5 seconds is an **approximation** because it is based on the use of class midpoint values instead of the original list of service times.

Calculating a Weighted Mean

- Weighted Mean
  - When different *x* data values are assigned different weights *w*, we can compute a weighted mean.

$$\bar{x} = \frac{\Sigma(w \cdot x)}{\Sigma w}$$

Example: Computing Grade-Point Average (1 of 4)

In her first semester of college, a student of the author took five courses. Her final grades, along with the number of credits for each course, were A (3 credits), A (4 credits), B (3 credits), C (3 credits), and F (1 credit).

The grading system assigns quality points to letter grades as follows: A = 4; B = 3; C = 2; D = 1; F = 0. Compute her grade-point average.

Example: Computing Grade-Point Average (2 of 4)

Solution

- Use the numbers of credits as weights: w = 3, 4, 3, 3, 1.
- Replace the letter grades of A, A, B, C, and F with the corresponding quality points: x = 4, 4, 3, 2, 0.

Example: Computing Grade-Point Average (3 of 4)

Solution

$$\bar{x} = \frac{\Sigma(w \cdot x)}{\Sigma w}$$

$$= \frac{(3 \times 4) + (4 \times 4) + (3 \times 3) + (3 \times 2) + (1 \times 0)}{3 + 4 + 3 + 3 + 1}$$

$$= \frac{43}{14} = 3.07$$

Example: Computing Grade-Point Average (4 of 4)

Solution

The result is a first-semester grade-point average of 3.07. (In using the preceding round-off rule, the result should be rounded to 3.1, but it is common to round grade-point averages to two decimal places.)