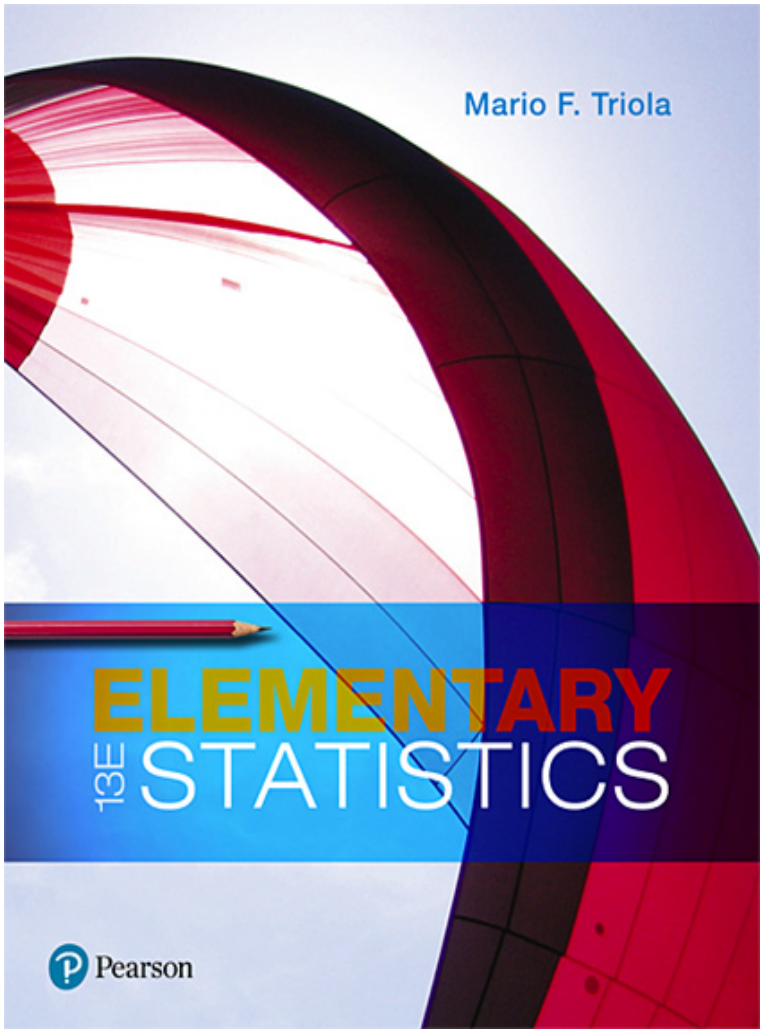
Elementary Statistics

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

$$z = \frac{x - \bar{x}}{s}$$

or

Population

$$z = \frac{x}{\quad}$$

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:

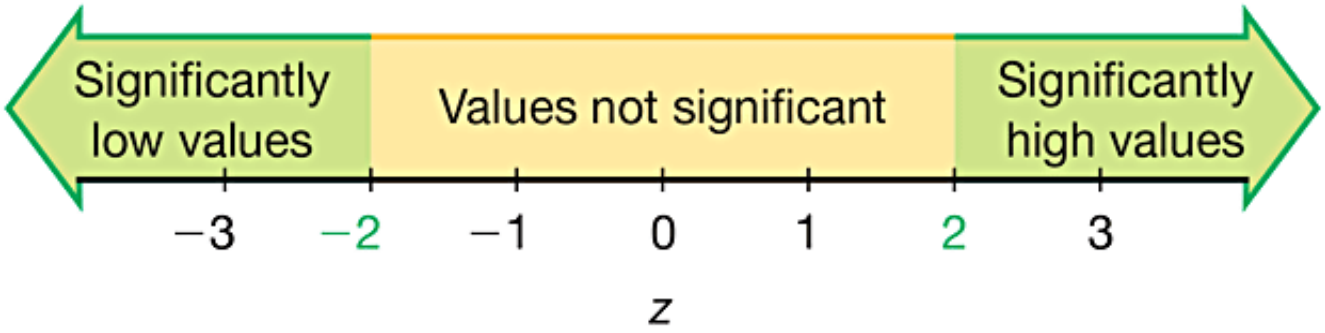
$$z = \frac{x - \bar{x}}{s} = \frac{99°\text{F} - 398.20°\text{F}}{0.62°\text{F}} = 1.29$$

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 $P_1, P_2, \ldots, P_{99}$, 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

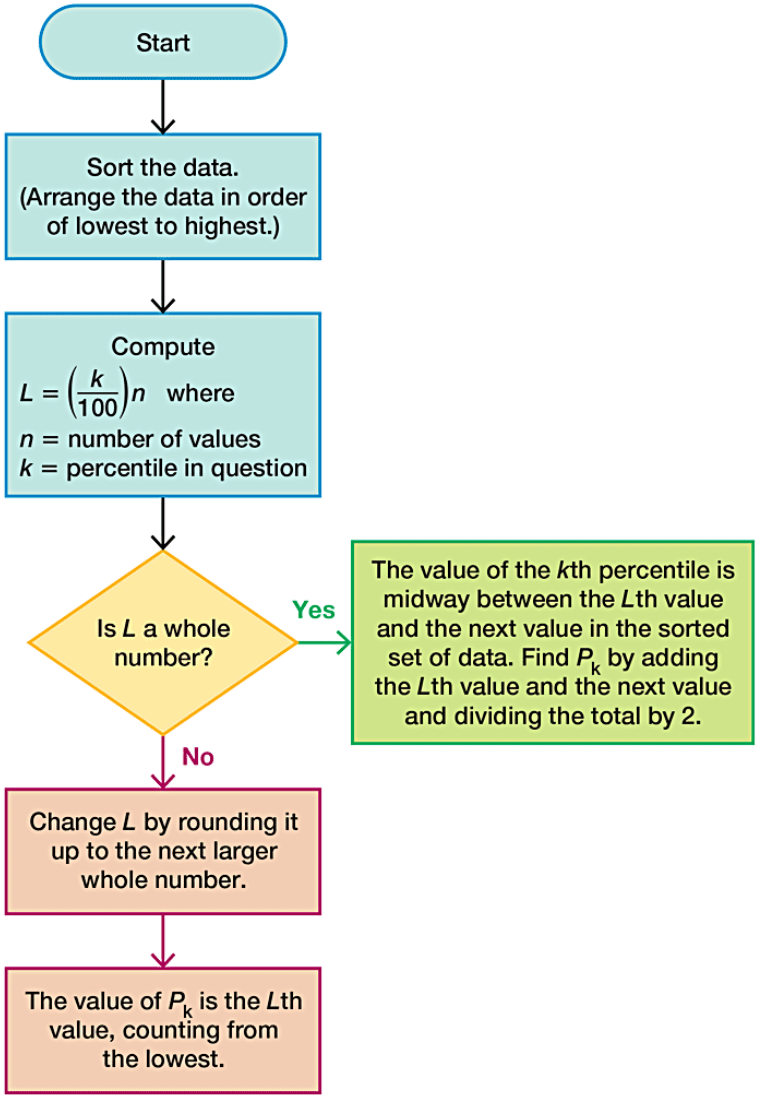
n total number of values in the data set

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 kth percentile (Example: P_{25} is the 25th percentile.)

Converting a Percentile to a Data Value



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

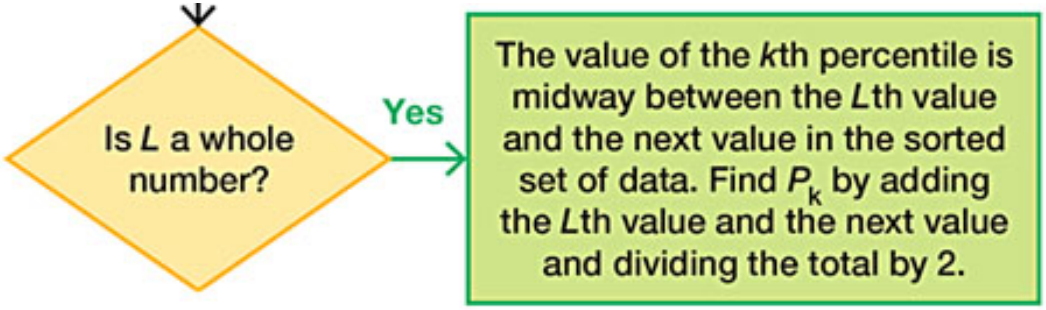
We can proceed to compute the value of the locator L. In this computation, we use $k = 40$ because we are attempting to find the value of the 40th percentile, and we use $n = 50$ because there are 50 data values.

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

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.



We now see that the value of the 40th percentile is midway between the Lth (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

$$\text{Interquartile range (or IQR)} = Q_3 - Q_1$$

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

$$\text{Midquartile range} = \frac{Q_3 + Q_1}{2}$$

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

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)

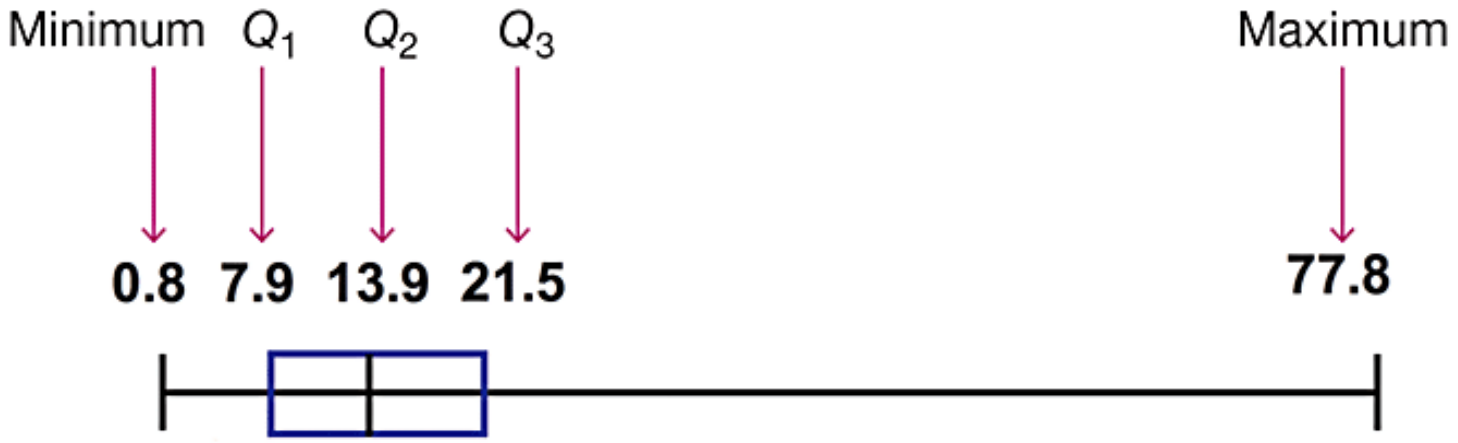
Use the Verizon airport data speeds to construct a boxplot.

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.