Chapter 3 Describing, Exploring, and Comparing Data
Measures of Relative Standing and Boxplots (3.3)

Objectives

In this lesson, we will learn how to

- compute and interpret z scores;
- use z scores to compare values from different data sets;
- identify significantly low and significantly high values;
- find and interpret percentiles and quartiles;
- find a five-number summary;
- construct and interpret boxplots and modified boxplots.

1. Measures of Relative Standing

A measure of relative standing describes the location of a data value relative to the other values in the same data set.

The most important measure of relative standing in this section is the z score.

z Score

A z score, also called a standard score or standardized value, is the number of standard deviations that a value x is above or below the mean.
For a sample:

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

For a population:

$$z = \frac{x - \mu}{\sigma}$$

Round-Off Rule for z Scores

Round z scores to two decimal places, such as 2.31.

2. Important Properties of z Scores

Interpreting z Scores

- A positive z score means the value is above the mean.
- A negative z score means the value is below the mean.
- A z score of 0 means the value is equal to the mean.
- z scores have no units.

Significant Values Using z Scores

- Significantly low values: $z \leq -2$
- Significantly high values: $z \geq 2$
- Values not significant: $-2 < z < 2$

Example 1: Which of the following two data values is more extreme relative to the data set from which it came?

- A $99°$F body temperature among 106 adults with sample mean $\bar{x} = 98.20°$F and sample standard deviation $s = 0.62°$F.
- A 5.7790 g weight of a quarter among 40 quarters with sample mean $\bar{x} = 5.63930$ g and sample standard deviation $s = 0.06194$ g.

Solution

For the body temperature,

$$z = \frac{x - \bar{x}}{s} = \frac{99 - 98.20}{0.62} = \frac{0.80}{0.62} \approx 1.29.$$

For the quarter weight,

$$z = \frac{x - \bar{x}}{s} = \frac{5.7790 - 5.63930}{0.06194} = \frac{0.13970}{0.06194} \approx 2.26.$$

The quarter weight is more extreme because 2.26 is farther from 0 than 1.29.

Example 2: Among 600 earthquake magnitudes, the mean is 2.572 and the standard deviation is 0.651. One earthquake has magnitude 4.01.
Is the magnitude 4.01 significantly high?

Solution

Compute the z score:

$$z = \frac{x - \bar{x}}{s} = \frac{4.01 - 2.572}{0.651} = \frac{1.438}{0.651} \approx 2.21.$$

Since $z = 2.21 \geq 2$, the magnitude 4.01 is significantly high.

3. Percentiles

Percentiles

Percentiles are measures of location, denoted

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

which divide a data set into 100 groups with about 1% of the values in each group.

Finding the Percentile of a Data Value

For a data value x,

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

Round the result to the nearest whole number.

Example 3: The following are 50 sorted Space Mountain wait times, in minutes.

10	15	15	15	15	15	20	20	20	20
25	25	25	25	25	25	30	30	30	30
30	30	30	30	35	35	35	35	35	35
35	35	40	40	40	40	45	50	50	50
50	50	55	55	60	75	75	75	105	110

Find the percentile for a wait time of 45 minutes.

Solution

There are 36 wait times less than 45 minutes, and there are 50 total wait times.

$$\text{Percentile of } 45 = \frac{36}{50} \cdot 100 = 72.$$

So a wait time of 45 minutes is in the 72nd percentile.
This means that 45 minutes separates the lowest 72% of values from the highest 28% of values. $P_{72} = 45$ minutes.

4. Converting a Percentile to a Data Value

Notation

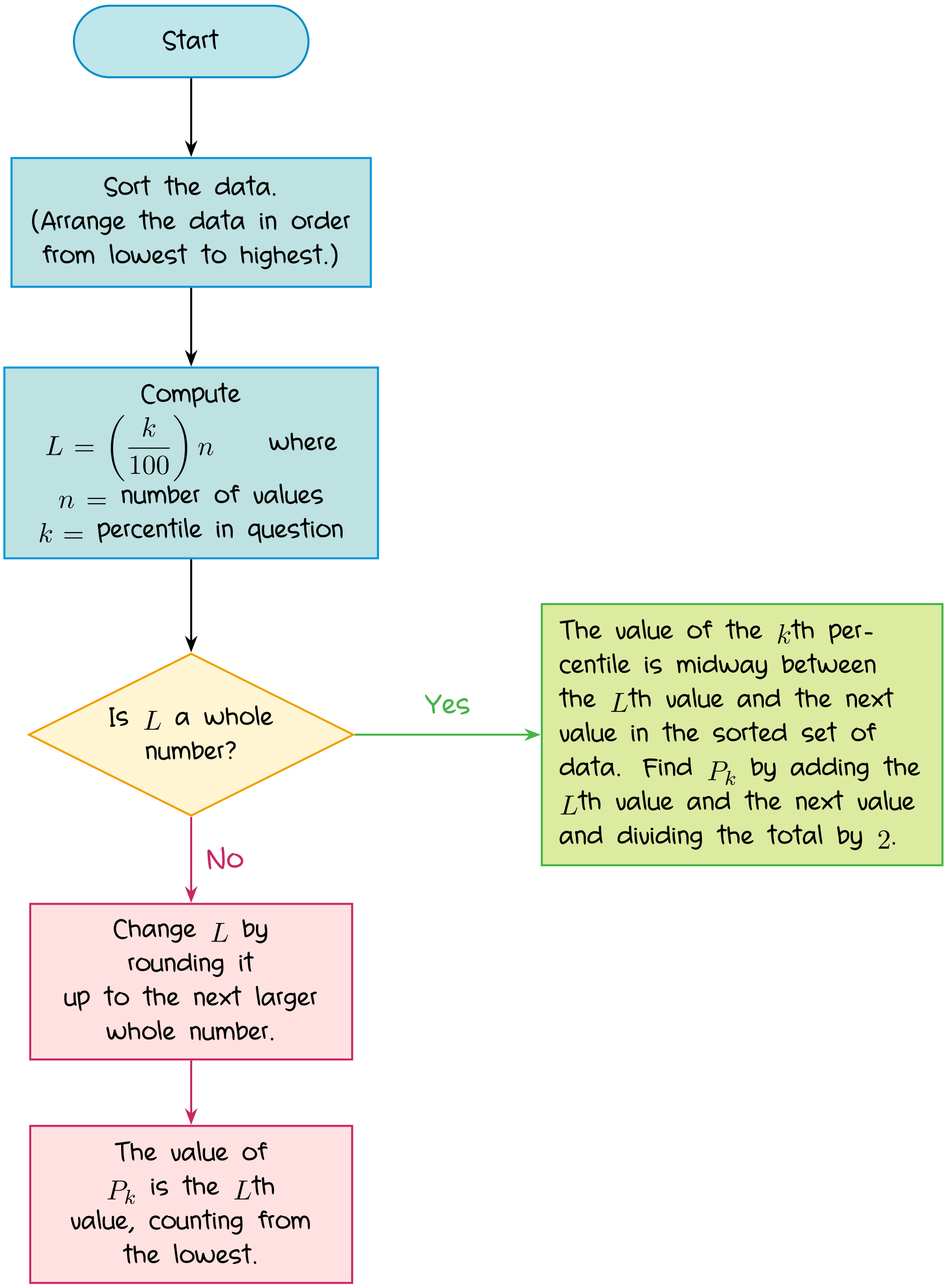
- n: total number of values in the data set
- k: percentile being used
- L: locator giving the position of a value
- P_k: kth percentile

Locator Formula

To find the data value corresponding to the kth percentile, compute

$$L = \frac{k}{100}n.$$

- If L is not a whole number, round L up to the next larger whole number. The value of P_k is the value in that position.
- If L is a whole number, P_k is the mean of the values in positions L and $L + 1$.


Start
Sort the data. (Arrange the data in order from 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 4: Use the 50 sorted Space Mountain wait times from Example 3 to find P_{25}, the 25th percentile.

Solution

Here $k = 25$ and $n = 50$, so

$$L = \frac{25}{100}(50) = 12.5.$$

Since $L = 12.5$ is not a whole number, round up to 13.
The 13th value in the sorted list is 25. Therefore,

$$P_{25} = 25 \text{ minutes}.$$

5. Quartiles

Quartiles

Quartiles divide a data set into four groups with about 25% of the values in each group.

- Q_1: first quartile, same as P_{25}
- Q_2: second quartile, same as P_{50}, and also the median
- Q_3: third quartile, same as P_{75}

Statistics Using Quartiles and Percentiles

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

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

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

$$10\text{-}90 \text{ Percentile Range} = P_{90} - P_{10}$$

6. Five-Number Summary

Five-Number Summary

For a data set, the five-number summary consists of:

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

Example 5: Use the 50 sorted Space Mountain wait times to find the five-number summary.

Solution

From the sorted list:

$$\text{Minimum} = 10, \qquad \text{Maximum} = 110.$$

Using the percentile/quartile procedure,

$$Q_1 = 25, \qquad Q_2 = 35, \qquad Q_3 = 50.$$

Therefore, the five-number summary is

$$10, \quad 25, \quad 35, \quad 50, \quad 110.$$

All values are in minutes.

7. Boxplots

Boxplot

A boxplot, or box-and-whisker diagram, is a graph based on the five-number summary.
It has:

- a line extending from the minimum to the maximum;
- a box extending from Q_1 to Q_3;
- a line inside the box at the median Q_2.

Procedure for Constructing a Boxplot

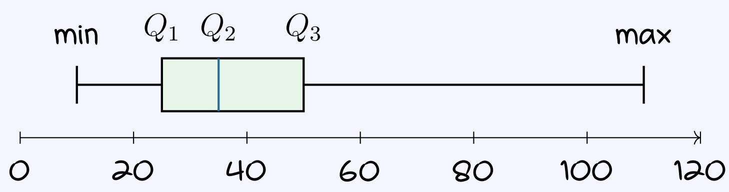
1. Find the five-number summary.
2. Draw a number line.
3. Draw whiskers from the minimum to Q_1 and from Q_3 to the maximum.
4. Draw a box from Q_1 to Q_3.
5. Draw a vertical line inside the box at the median.

Example 6: Construct a boxplot for the Space Mountain wait times using the five-number summary

$$10, \quad 25, \quad 35, \quad 50, \quad 110.$$

Solution

The minimum is 10, $Q_1 = 25$, median $Q_2 = 35$, $Q_3 = 50$, and maximum is 110.

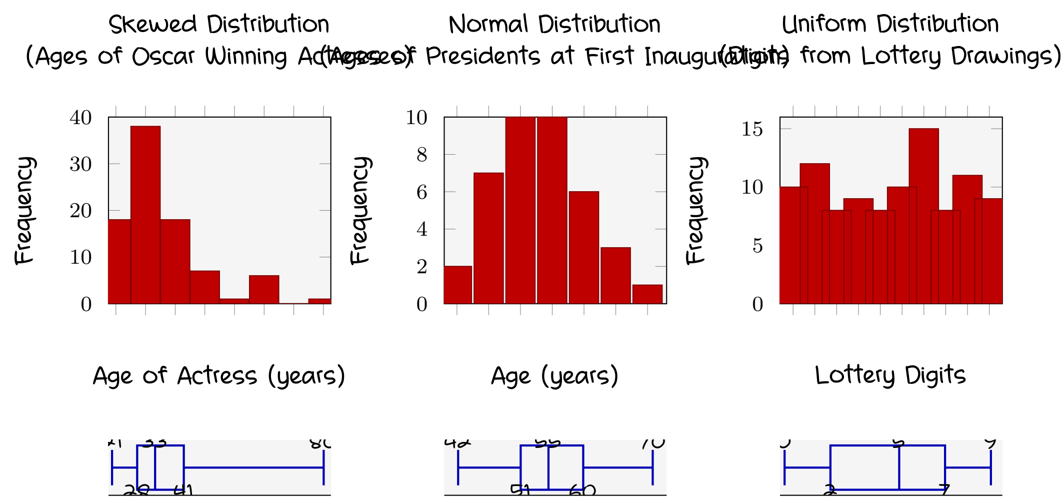


8. Skewness and Modified Boxplots

Skewness

A distribution is skewed if it is not symmetric and extends more to one side than to the other.
Boxplots can help us see skewness:

- A long right whisker suggests right skewness.
- A long left whisker suggests left skewness.



Identifying Outliers for Modified Boxplots

1. Find Q_1, Q_2, and Q_3.
2. Find the interquartile range:

$$\text{IQR} = Q_3 - Q_1.$$

3. Compute

$$1.5 \times \text{IQR}.$$

4. A data value is an outlier if it is more than $1.5 \times \text{IQR}$ above Q_3, or more than $1.5 \times \text{IQR}$ below Q_1.

Modified Boxplot

A modified boxplot is a boxplot that shows outliers with a special symbol, such as an asterisk or point.
The whiskers extend only to the smallest and largest values that are not outliers.

Example 7: For the Space Mountain wait times, suppose the five-number information includes

$$Q_1 = 25, \qquad Q_3 = 50.$$

Find the IQR and the upper outlier boundary.

Solution

First find the interquartile range:

$$\text{IQR} = Q_3 - Q_1 = 50 - 25 = 25.$$

Then compute

$$1.5(\text{IQR}) = 1.5(25) = 37.5.$$

The upper outlier boundary is

$$Q_3 + 1.5(\text{IQR}) = 50 + 37.5 = 87.5.$$

Any value greater than 87.5 minutes is an outlier.

Summary

Key Ideas

- A z score tells how many standard deviations a value is above or below the mean.
- Values with $z \leq -2$ are significantly low, and values with $z \geq 2$ are significantly high.
- Percentiles divide data into 100 groups.
- Quartiles divide data into four groups.
- The five-number summary is minimum, Q_1, median, Q_3, maximum.
- A boxplot displays the five-number summary visually.
- A modified boxplot marks outliers separately.