Artificial iNtelligence

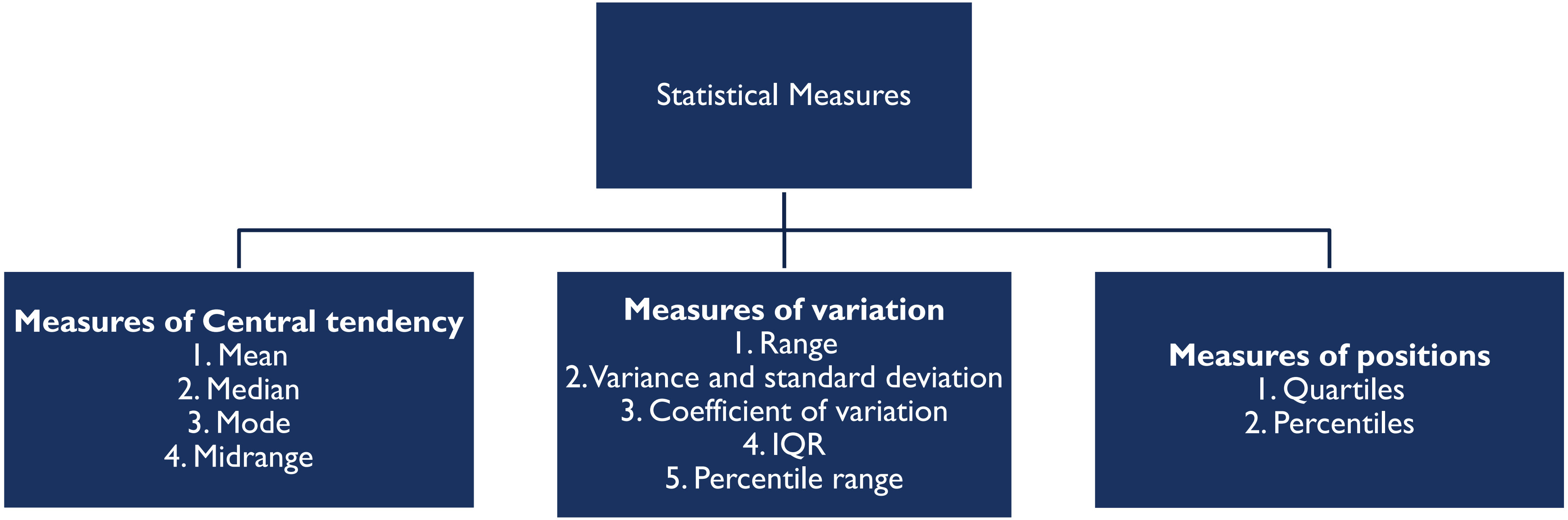
AI in Best Way

AN INTRODUCTION TO STATISTICS

LECTURE 4

BY/ ALY MAHER ABDELFATTAH

STATISTICAL MEASURES



Statistical Measures

Measures of Central tendency
1. Mean
2. Median
3. Mode
4. Midrange

Measures of variation
1. Range
2. Variance and standard deviation
3. Coefficient of variation
4. IQR
5. Percentile range

Measures of positions
1. Quartiles
2. Percentiles

MEASURES OF VARIATION

- Suppose we have the following two sets of data {4,5,6} and {1,5,9}. Although the two sets have a mean of 5 but still the two sets are different (the second set has more variation than the first), so we need another measure that gives a more description about variation of the data

MEASURES OF VARIATION

- **There are different measures of variation or dispersion as**

1. Range
2. Inter quartile range (IQR)
3. Percentile range (PR)
4. Standard deviation ($\boldsymbol{S}$) and variance ($\boldsymbol{S2}$)
5. Coefficient of variation ($\boldsymbol{C.V}$)

MEASURES OF VARIATION

- **1. Range**

- **Definition**: The range of a set of data is the difference between the highest and lowest values in the set.
- **Calculation**: Range = Maximum value - Minimum value.
- **Use**: It provides a basic measure of the spread or dispersion of the data points in the set. The range is heavily influenced by outliers or extreme values in the data set.

MEASURES OF VARIATION

2. Interquartile Range (IQR)

- **Definition:** The Interquartile Range (IQR) is a measure of statistical dispersion and is calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1) in a data set.
- **Calculation:** IQR = Q3 - Q1
- **Use:** It is particularly useful for understanding the spread of the middle 50% of values in a dataset and is less affected by extreme values (outliers) than the range.

MEASURES OF VARIATION

3. Percentile Range (PR)

- **Definition:** Percentile Range refers to the range between two specified percentiles. For example, the range between the 10th and 90th percentiles.
- **Calculation:** PR = P_upper - P_lower, where P_upper and P_lower are the upper and lower percentiles, respectively.
- **Use:** This can be useful for understanding the spread of data in specific portions of the dataset.

MEASURES OF VARIATION

- **4. Standard Deviation (S) and Variance (S^2)**

❑ **Standard Deviation (S)**

- **Definition:** A measure of the amount of variation or dispersion in a set of values.
- **Calculation:** The square root of the variance. It's calculated by finding the square root of the average squared deviation of each number from the mean of the data set.

❑ **Variance (S^2)**

- **Definition:** The average of the squared differences from the Mean.
- **Calculation:** Sum of squared deviations from the mean divided by the number of observations.

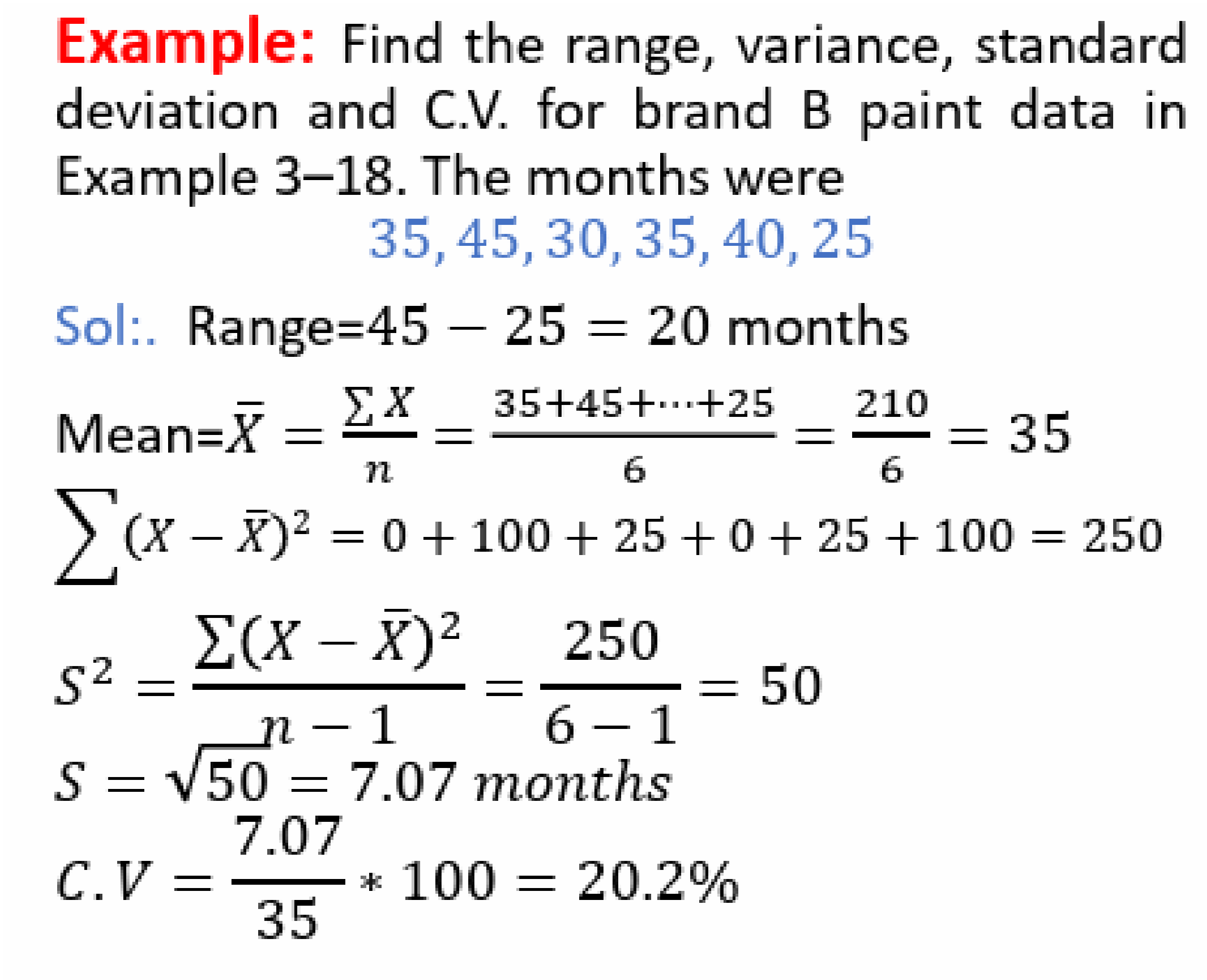
❑ **Use:** Both are widely used in statistics to measure how spread out the numbers in a data set are. The standard deviation is particularly useful as it is in the same units as the data.

MEASURES OF VARIATION

5. Coefficient of Variation ($\boldsymbol{C.V}$)

- **Definition:** The Coefficient of Variation is a standardized measure of dispersion of a probability distribution or frequency distribution.
- **Calculation:** It is calculated as the ratio of the standard deviation to the mean.
- **Formula:** $\boldsymbol{C.V}$. = (Standard Deviation / Mean) × 100%
- **Use:** It is often expressed as a percentage and is used to compare the relative variability between different datasets on a ratio scale.

MEASURES OF VARIATION (EXAMPLE)



Example: Find the range, variance, standard deviation and C.V. for brand B paint data in Example 3–18. The months were

35, 45, 30, 35, 40, 25

Sol:. Range=45 − 25 = 20 months

Mean=$\bar{X} = \frac{\sum X}{n} = \frac{35+45+\cdots+25}{6} = \frac{210}{6} = 35$

$$\sum (X-\bar{X})^2 = 0+100+25+0+25+100 = 250$$

$$S^2 = \frac{\sum(X-\bar{X})^2}{n-1} = \frac{250}{6-1} = 50$$

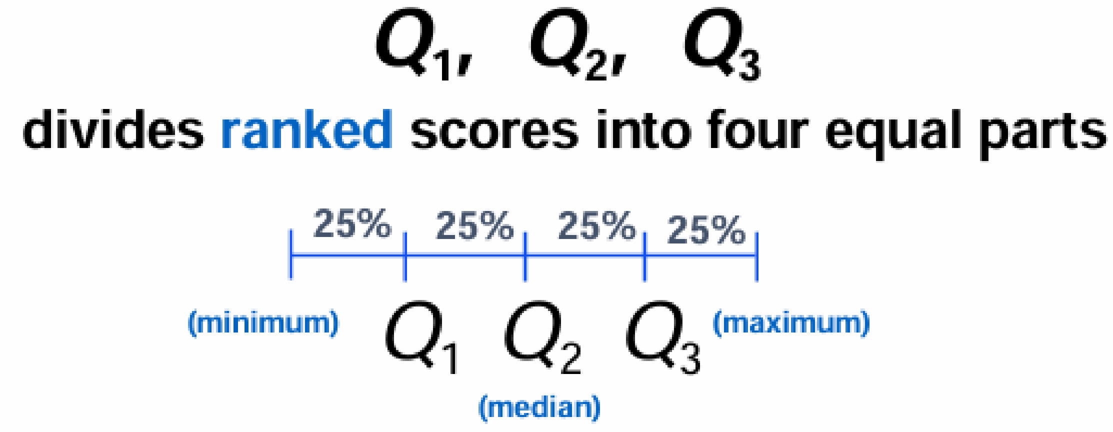
$$S = \sqrt{50} = 7.07\ months$$

$$C.V = \frac{7.07}{35} * 100 = 20.2\%$$

X	$(X - \bar{X})$	$(X - \bar{X})^2$
35	35-35 = 0	0
45	45-35 = 10	100
30	30-35 = -5	25
35	35-35 = 0	0
40	40-35 = 5	25
25	25-35 = -10	100
		∑. = 250

MEASURES OF VARIATION (QUARTILES)

- Q1(First Quartile) separates the bottom 25% of sorted values from the top 75%.
- Q2(Second Quartile) same as the median; separates the bottom 50% of sorted values from the top 50%.
- Q3(Third Quartile) separates the bottom 75% of sorted values from the top 25%.



Q_1, Q_2, Q_3

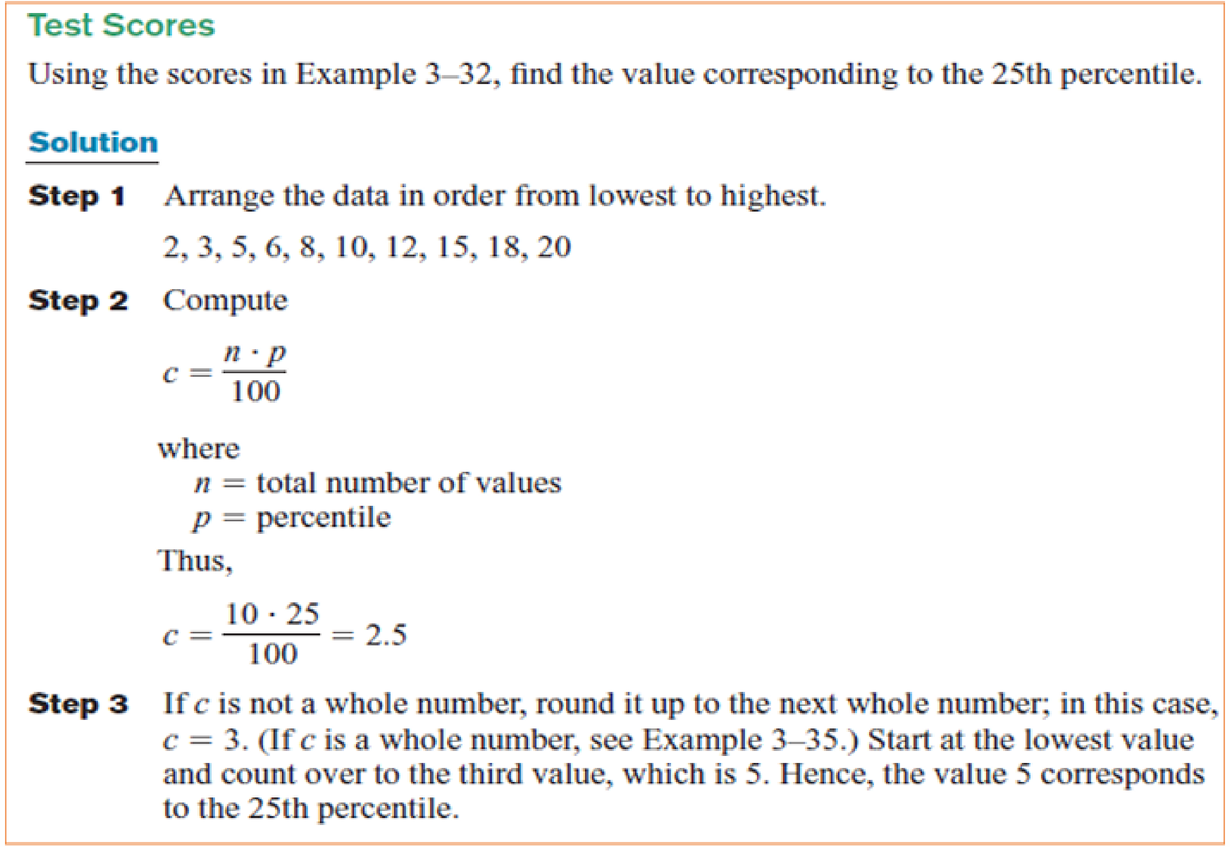
divides ranked scores into four equal parts

MEASURES OF VARIATION (PERCENTILES)

- Just as there are quartiles separating data into four parts, there are 99 percentiles denoted P1, P2, . . . P99, which partition the data into 100 groups.

- Note that $\boldsymbol{Q1 = P25, Q2 = median = P50}$ and $\boldsymbol{Q3 = P75}$

MEASURES OF VARIATION (EXAMPLE)



Test Scores

Using the scores in Example 3–32, find the value corresponding to the 25th percentile.

Solution

Step 1 Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Step 2 Compute

$$c = \frac{n \cdot p}{100}$$

where

n = total number of values

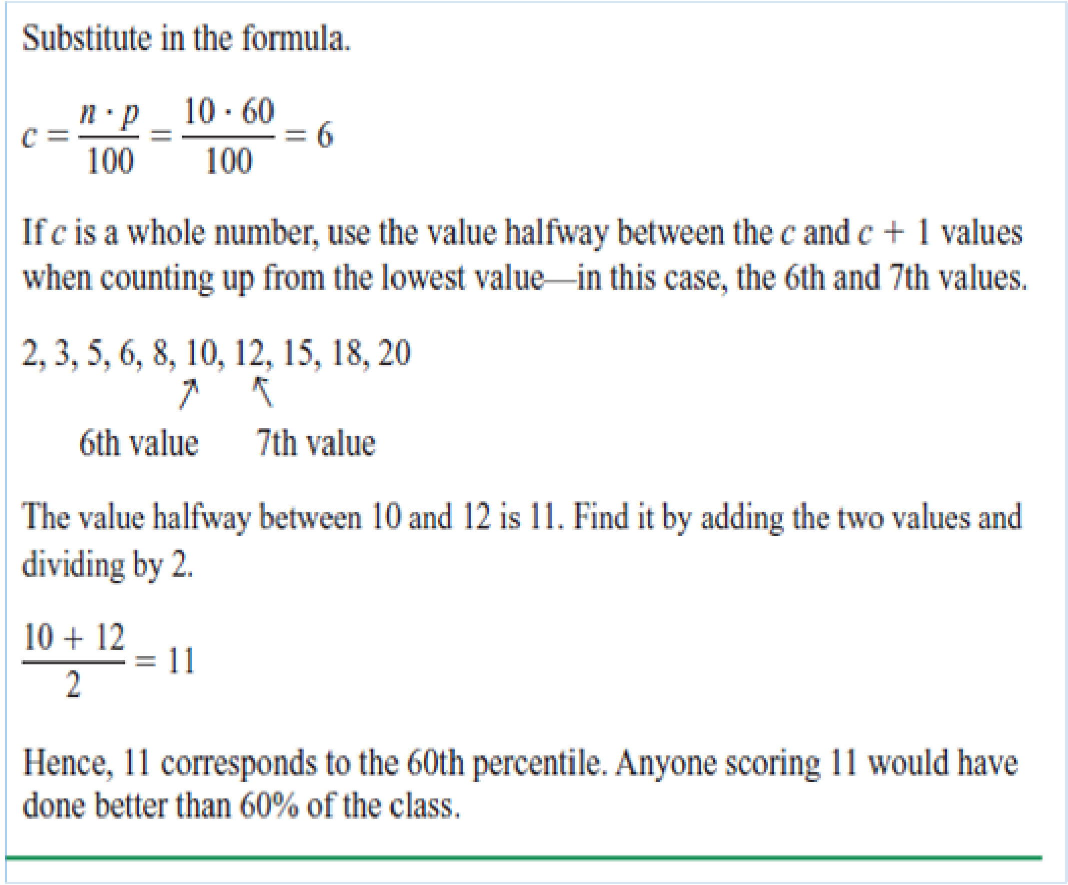
p = percentile

Thus,

$$c = \frac{10 \cdot 25}{100} = 2.5$$

Step 3 If c is not a whole number, round it up to the next whole number; in this case, $c = 3$. (If c is a whole number, see Example 3–35.) Start at the lowest value and count over to the third value, which is 5. Hence, the value 5 corresponds to the 25th percentile.

Find also P_{60}



Substitute in the formula.

$$c = \frac{n \cdot p}{100} = \frac{10 \cdot 60}{100} = 6$$

If c is a whole number, use the value halfway between the c and $c + 1$ values when counting up from the lowest value—in this case, the 6th and 7th values.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

↗ ↖

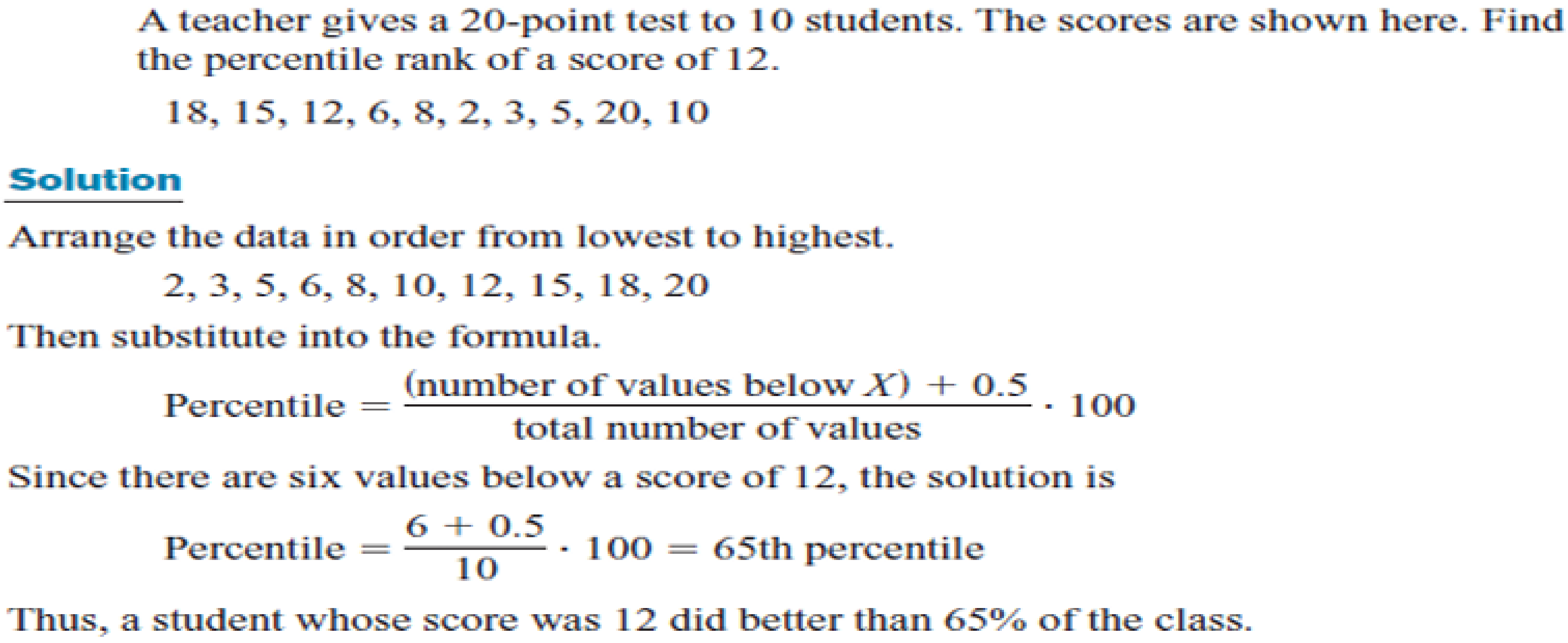
6th value 7th value

The value halfway between 10 and 12 is 11. Find it by adding the two values and dividing by 2.

$$\frac{10 + 12}{2} = 11$$

Hence, 11 corresponds to the 60th percentile. Anyone scoring 11 would have done better than 60% of the class.

MEASURES OF VARIATION (EXAMPLE)



A teacher gives a 20-point test to 10 students. The scores are shown here. Find the percentile rank of a score of 12.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Solution

Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

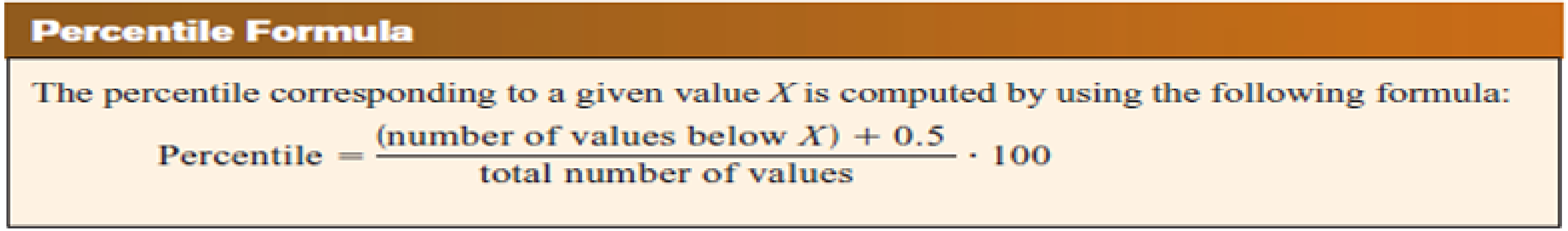
Then substitute into the formula.

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100$$

Since there are six values below a score of 12, the solution is

$$\text{Percentile} = \frac{6 + 0.5}{10} \cdot 100 = 65\text{th percentile}$$

Thus, a student whose score was 12 did better than 65% of the class.



Percentile Formula

The percentile corresponding to a given value X is computed by using the following formula:

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100$$

EXPLORATORY DATA ANALYSIS

- Exploratory Data Analysis is the process of using statistical tools (such as graphs, measures of center, and measures of variation) to investigate data sets in order to understand their important characteristics

OUTLIER

- An outlier is a value that is located very far away from almost all the other values
- An outlier can have a dramatic effect on the mean
- An outlier have a dramatic effect on the standard deviation
- An outlier can have a dramatic effect on the scale of the histogram so that the true nature of the distribution is totally obscured

OUTLIER

- Procedure for identifying outliers

Step1: Arrange the data in order and find Q1 and Q3.

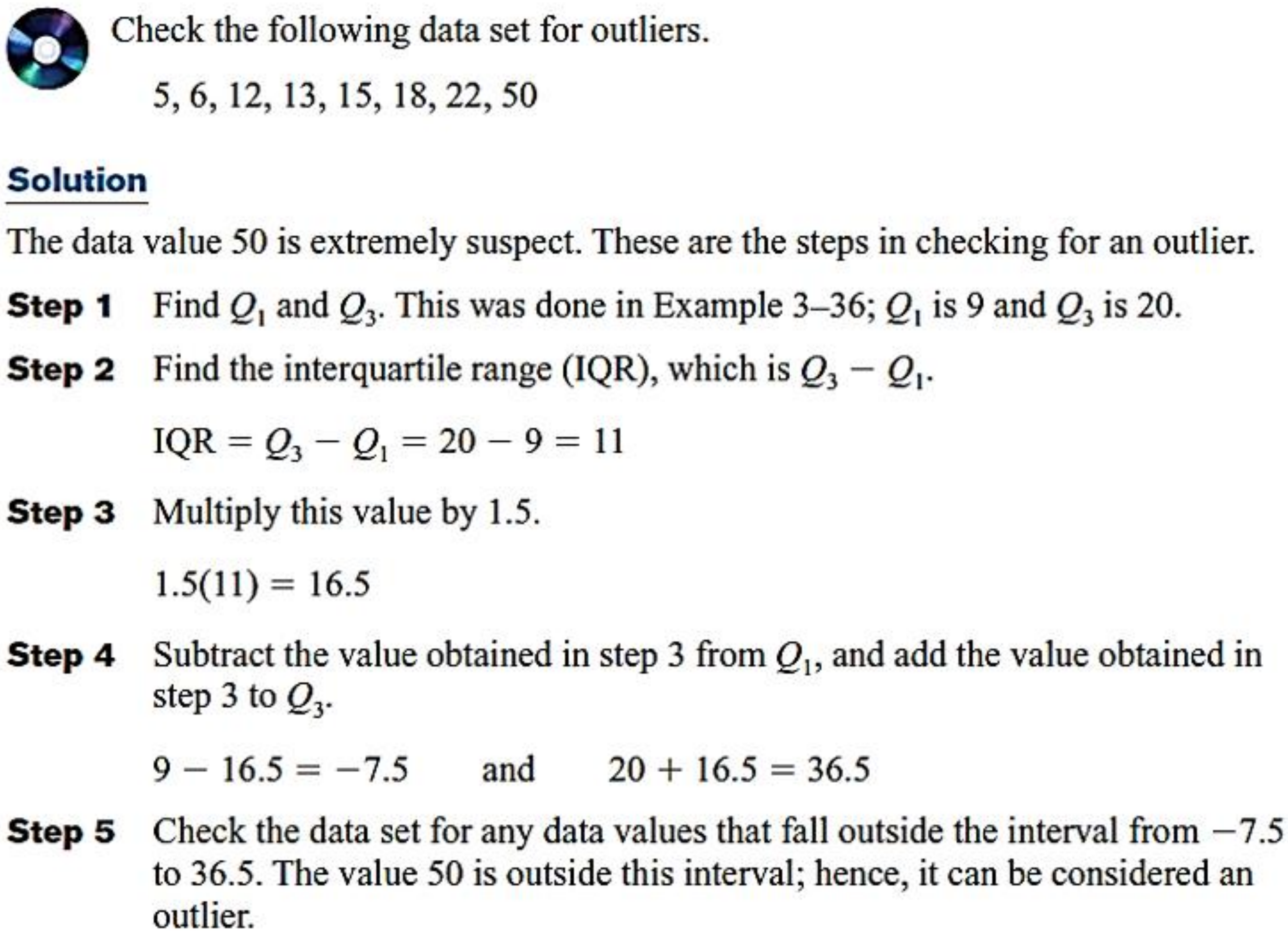
Step2: Find the interquartile range: IQR = Q3-Q1.

Step3: Multiply IQR by 1.5.

Step4: Subtract the value obtained in step3 from Q1 and add the value to Q3.

Step5: Check the data set for any data value that is smaller than Q1 – 1.5(IQR) or larger than Q3 + 1.5(IQR).

OUTLIERS (EXAMPLE)



Check the following data set for outliers.

5, 6, 12, 13, 15, 18, 22, 50

Solution

The data value 50 is extremely suspect. These are the steps in checking for an outlier.

Step 1 Find Q_1 and Q_3. This was done in Example 3–36; Q_1 is 9 and Q_3 is 20.

Step 2 Find the interquartile range (IQR), which is $Q_3 - Q_1$.

$$\text{IQR} = Q_3 - Q_1 = 20 - 9 = 11$$

Step 3 Multiply this value by 1.5.

$$1.5(11) = 16.5$$

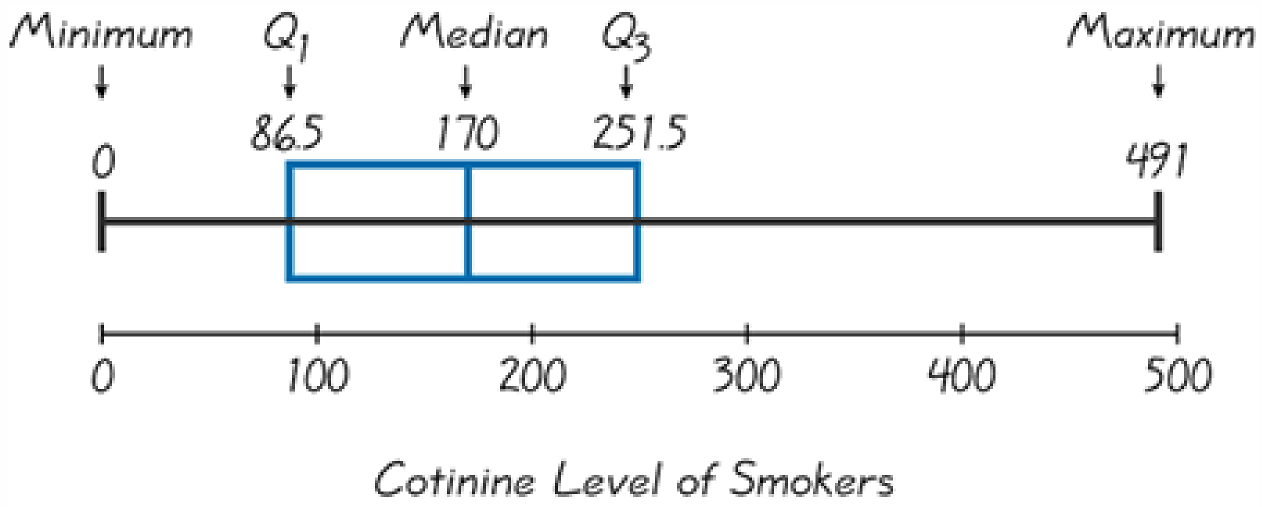
Step 4 Subtract the value obtained in step 3 from Q_1, and add the value obtained in step 3 to Q_3.

$$9 - 16.5 = -7.5 \qquad \text{and} \qquad 20 + 16.5 = 36.5$$

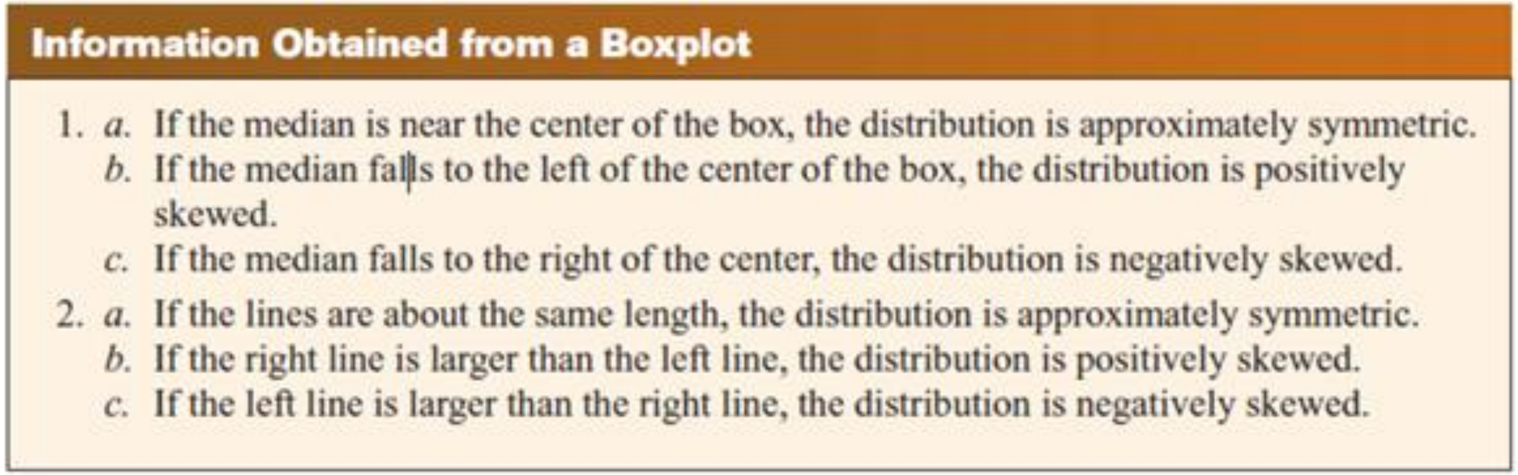
Step 5 Check the data set for any data values that fall outside the interval from −7.5 to 36.5. The value 50 is outside this interval; hence, it can be considered an outlier.

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 the median; and the third quartile, Q3

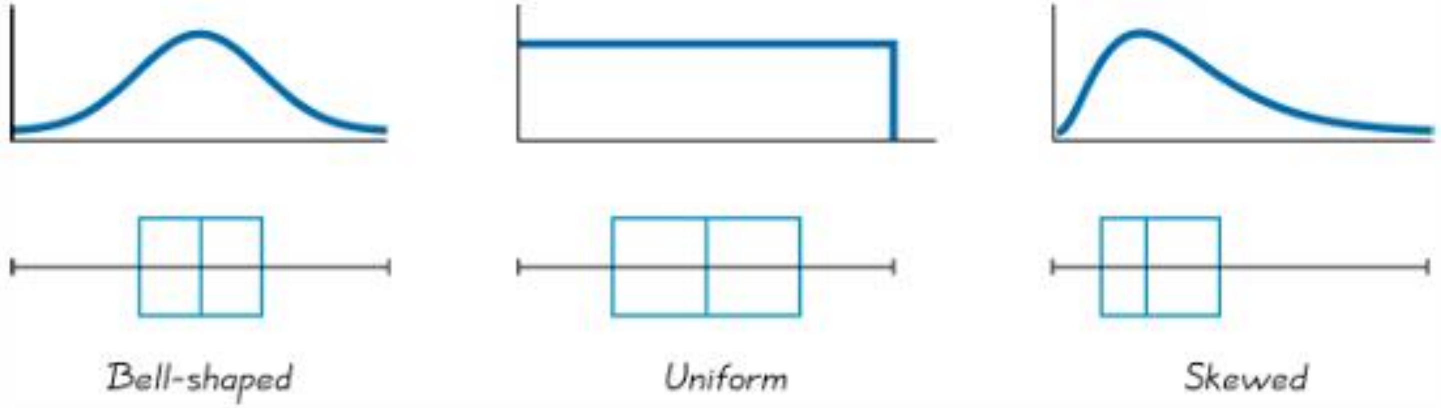


BOXPLOT (OR BOX-AND-WHISKER-DIAGRAM)

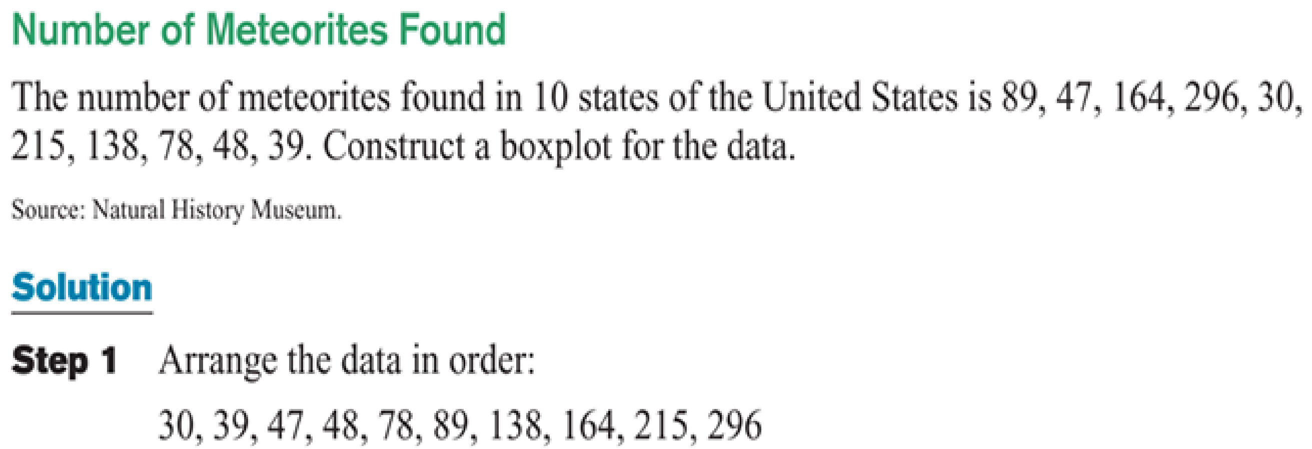


Information Obtained from a Boxplot

1. *a.* If the median is near the center of the box, the distribution is approximately symmetric.
   b. If the median falls to the left of the center of the box, the distribution is positively skewed.
   c. If the median falls to the right of the center, the distribution is negatively skewed.
2. *a.* If the lines are about the same length, the distribution is approximately symmetric.
   b. If the right line is larger than the left line, the distribution is positively skewed.
   c. If the left line is larger than the right line, the distribution is negatively skewed.



BOXPLOT (OR BOX-AND-WHISKER-DIAGRAM)



Number of Meteorites Found

The number of meteorites found in 10 states of the United States is 89, 47, 164, 296, 30, 215, 138, 78, 48, 39. Construct a boxplot for the data.

Source: Natural History Museum.

Solution

Step 1 Arrange the data in order:

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

To find $Q_1 = P_{25}$

$$C = \frac{n.P}{100} = \frac{(10)\,(25)}{100} = 2.5$$

$$C = 3$$

$$Q_1 = P_{25} = 47$$

To find $Q_2 = P_{50}$

$$C = \frac{n.P}{100} = \frac{(10)\,(50)}{100} = 5$$

$$Q_2 = P_{50} = \frac{78 + 89}{2} = 83.5$$

To find $Q_3 = P_{75}$

$$C = \frac{n.P}{100} = \frac{(10)\,(75)}{100} = 7.5$$

$$C = 8$$

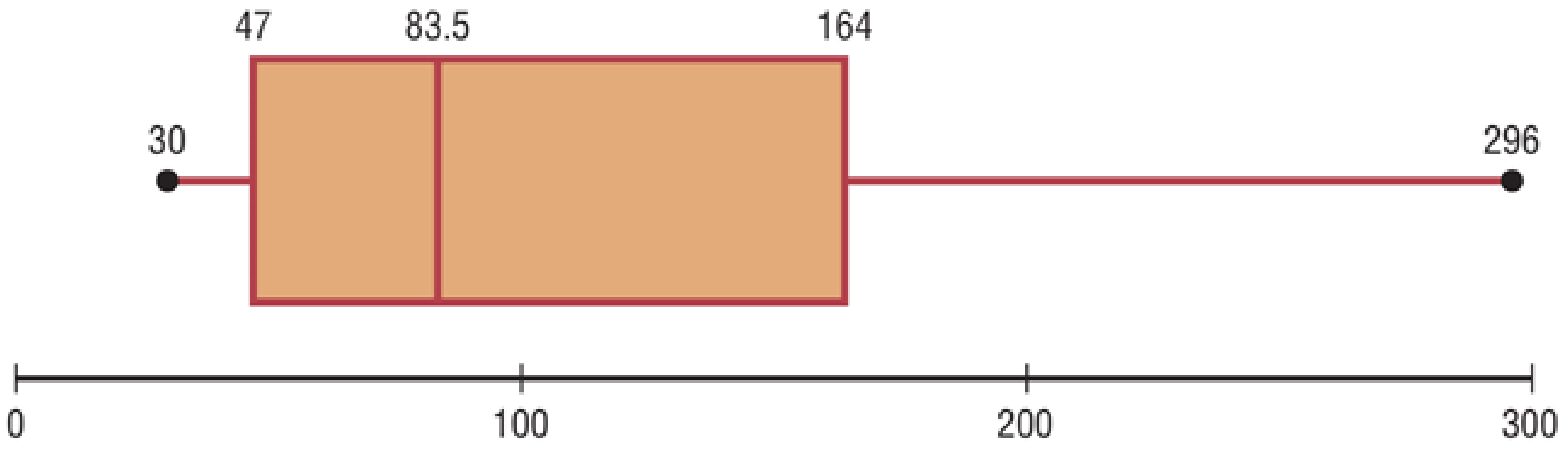
$$Q_3 = P_{75} = 164$$

BOXPLOT (OR BOX-AND-WHISKER-DIAGRAM)

Step 5 Draw a scale for the data on the *x* axis.

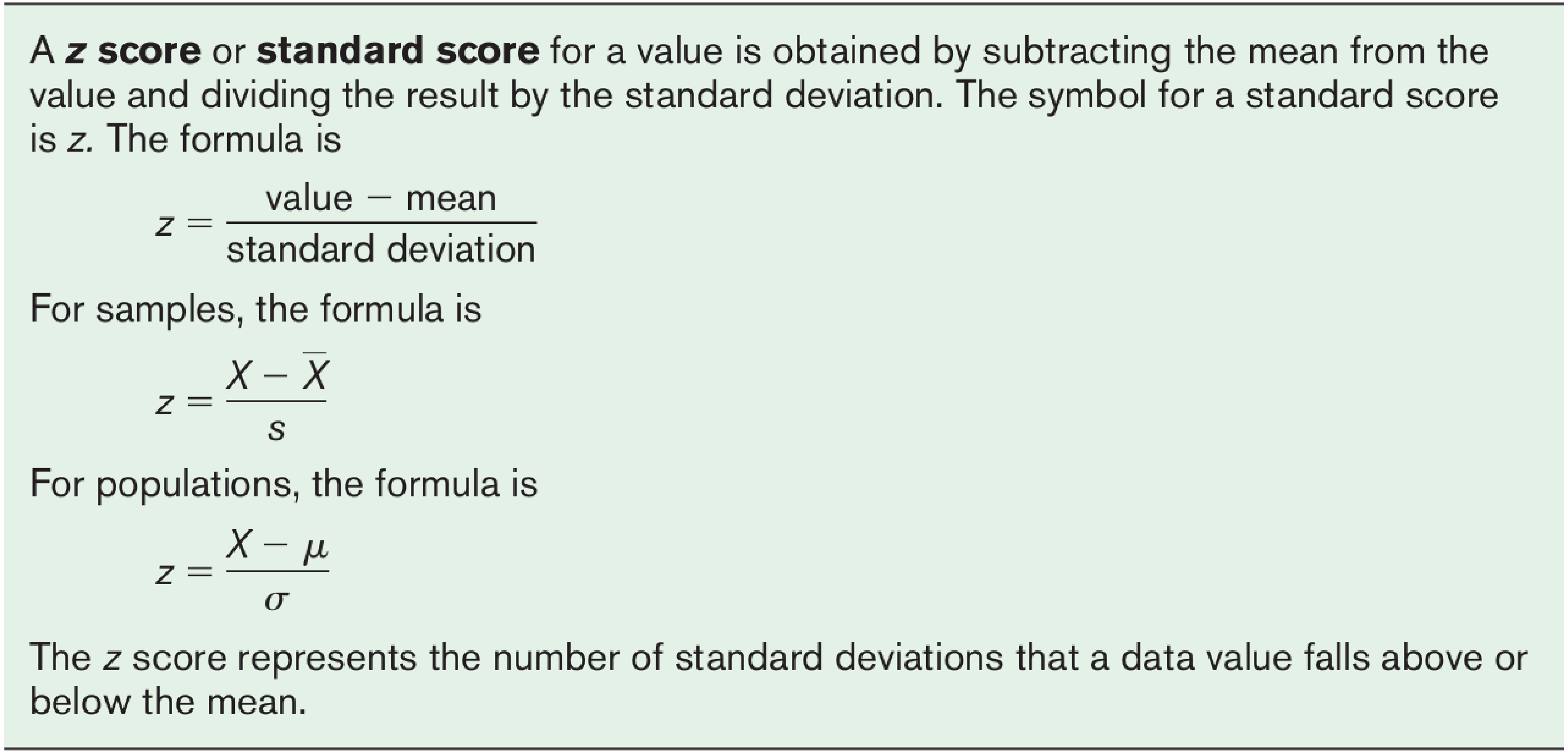
Step 6 Locate the lowest value, Q_1, median, Q_3, and the highest value on the scale.

Step 7 Draw a box around Q_1 and Q_3, draw a vertical line through the median, and connect the upper value and the lower value to the box. See Figure 3–7.



The distribution is somewhat positively skewed.

Z-SCORE



A ***z* score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is *z*. The formula is

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

For samples, the formula is

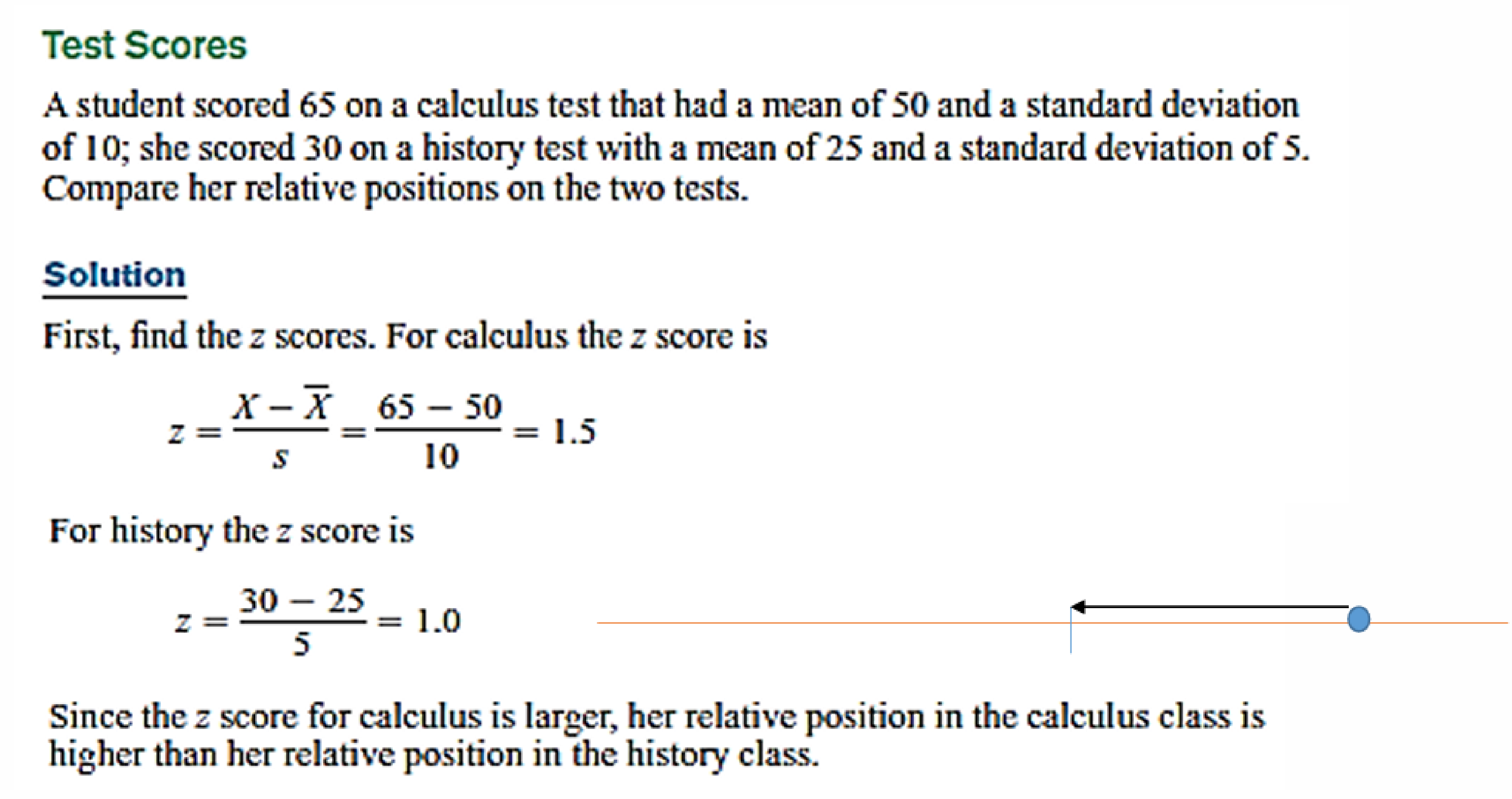
$$z = \frac{X - \overline{X}}{s}$$

For populations, the formula is

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

The *z* score represents the number of standard deviations that a data value falls above or below the mean.

Z-SCORE (EXAMPLE)



Test Scores

A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

Solution

First, find the z scores. For calculus the z score is

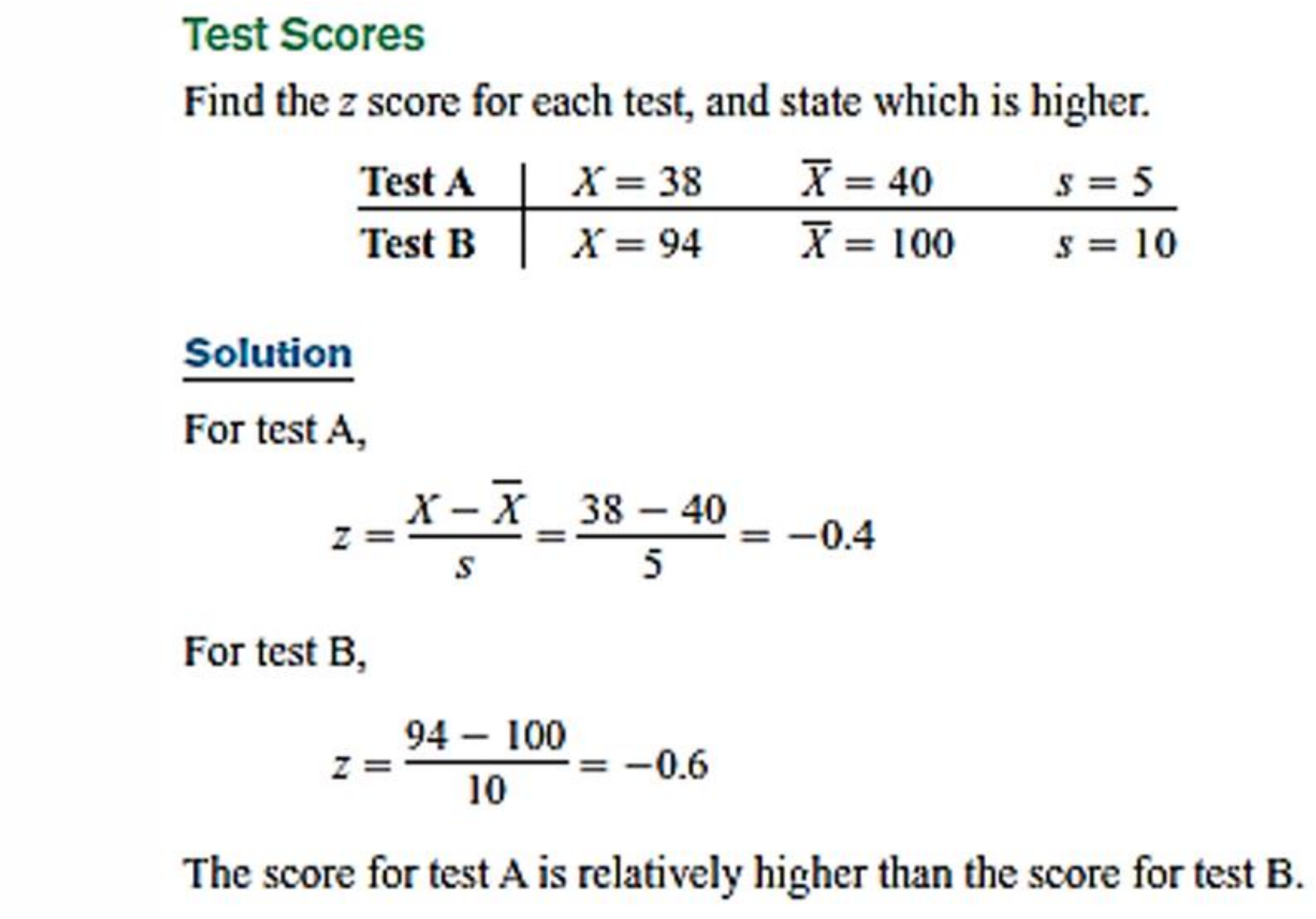
$$z = \frac{X - \overline{X}}{s} = \frac{65 - 50}{10} = 1.5$$

For history the z score is

$$z = \frac{30 - 25}{5} = 1.0$$

Since the z score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

Z-SCORE (EXAMPLE)



Test Scores

Find the z score for each test, and state which is higher.

Test A	$X = 38$	$\overline{X} = 40$	$s = 5$
Test B	$X = 94$	$\overline{X} = 100$	$s = 10$

Solution

For test A,

$$z = \frac{X - \overline{X}}{s} = \frac{38 - 40}{5} = -0.4$$

For test B,

$$z = \frac{94 - 100}{10} = -0.6$$

The score for test A is relatively higher than the score for test B.