

SEC “ 3 ”

Descriptive Statistics



(A) MEASURE OF CENTRAL TENDENCY :

- A measure of central tendency is a summary statistic that represents the center point or typical value of a dataset. You can think of it as the tendency of data to cluster around a middle value.
- In statistics, the three most common measures of central tendency are the mean, median, and mode. Each of these measures calculates the location of the central point using a different method.

(1) Mean (\bar{X}):

- The mean is the arithmetic average, and it is probably the measure of central tendency **that you are most familiar**. **Calculating the mean is very simple**, you just add up all of the values and divide by the number of observations in your dataset. For the data set x_1, x_2, \dots, x_n the mean is given as follow

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n},$$

Example :- **Statistics grade : 20, 30, 35, 45**

$$\bar{X} = \frac{20+30+35+45}{4} = 32,$$

Statistics grade : 1, 20, 30, 35, 45

$$\bar{X} = \frac{1+20+30+35+45}{5} = 26.2.$$

Statistics grade : 20, 30, 35, 45, 100

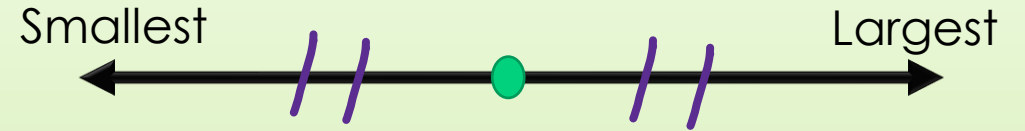
$$\bar{X} = \frac{20+30+35+45+100}{5} = 46,$$

- The calculation of the mean incorporates all values in the data.
- If you change any value, the mean changes. So the mean is sensitive to extreme values.

(2) Median :

- The median is the middle value. **It is the value that splits the dataset in half.** To find the median :

➤ Order your data from smallest to largest.



➤ The method for locating the median varies slightly depending on whether your dataset has an even or odd number of values.

$$\text{Median} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{if the sample size is an odd number} \longrightarrow 10, \boxed{20}, 30 \\ x_2 = 20 \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{if the sample size is an even number} \longrightarrow 10, \boxed{20}, \boxed{30}, 40 \\ \frac{x_2 + x_3}{2} \\ = \frac{20 + 30}{2} \\ = 25 \end{cases}$$

Example :- Statistics grade : 30, 20, 60, 45, 35

➤ 20, 30, **35**, 45, 60

Sample size is an odd number “ $n = 5$ ” \longrightarrow Median = $x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{5+1}{2}\right)} = x_{(3)} = 35$.

Statistics grade : 30, 20, 60, 45, 35, 10

➤ 10, 20, **30**, **35**, 45, 60

Sample size is an even number “ $n = 6$ ” \longrightarrow Median = $\frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} = \frac{x_{\left(\frac{6}{2}\right)} + x_{\left(\frac{6}{2}+1\right)}}{2} = \frac{x_{(3)} + x_{(4)}}{2}$
 $= \frac{30+35}{2} = 32.5$

- The median value doesn't depend on all the values in the dataset. Consequently, when some of the values are more extreme, the effect on the median is smaller.

(3) Mode :

- The mode is the value that occurs the most frequently in your data set.

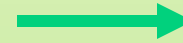
Example:- Statistics grade : 30, 30, 20, 60, 60, 60, 45, 35

Mode = 60.

Statistics grade : 30, 30, 30, 20, 60, 60, 60, 45, 35

Mode = 30 , 60.

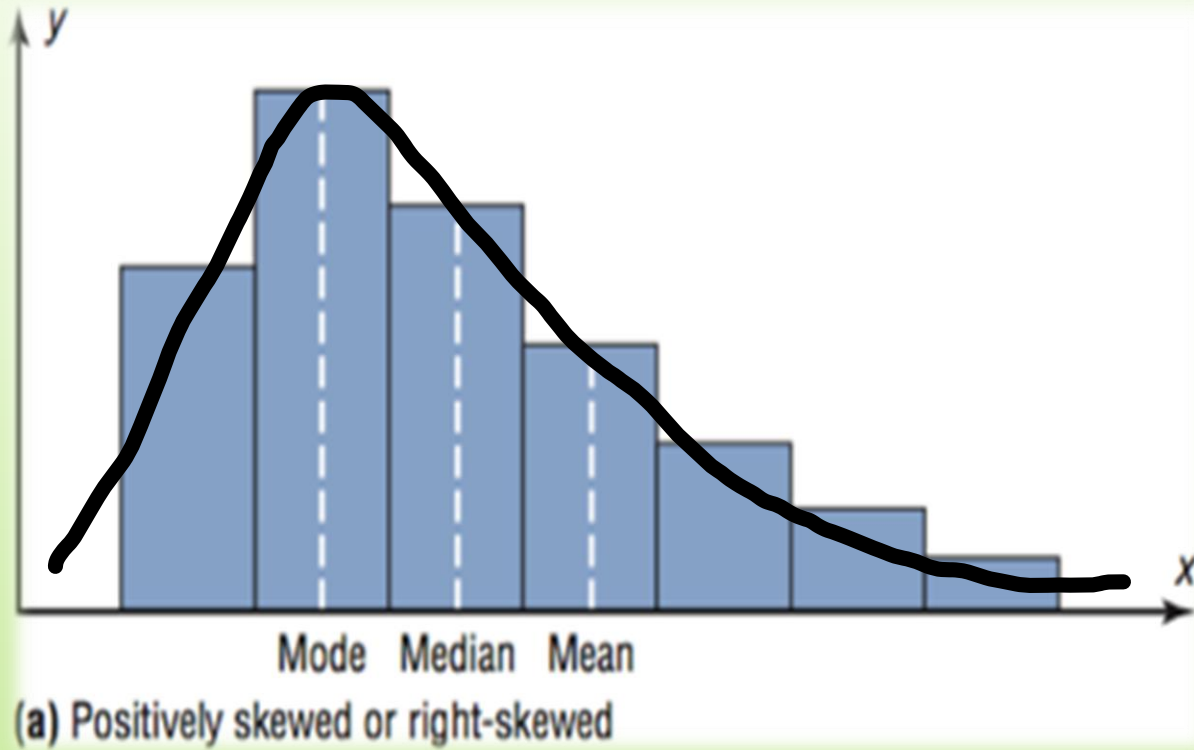
Statistics grade : 30, 20, 60, 45, 35



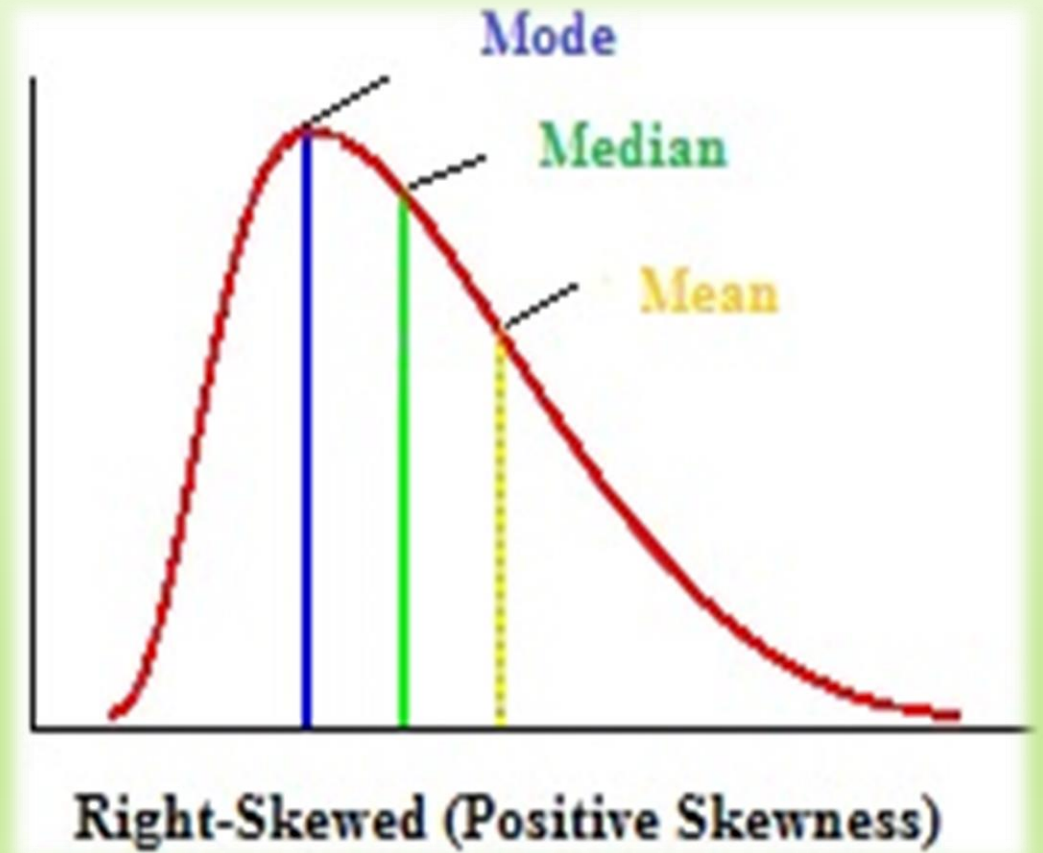
~~Mode = 0~~

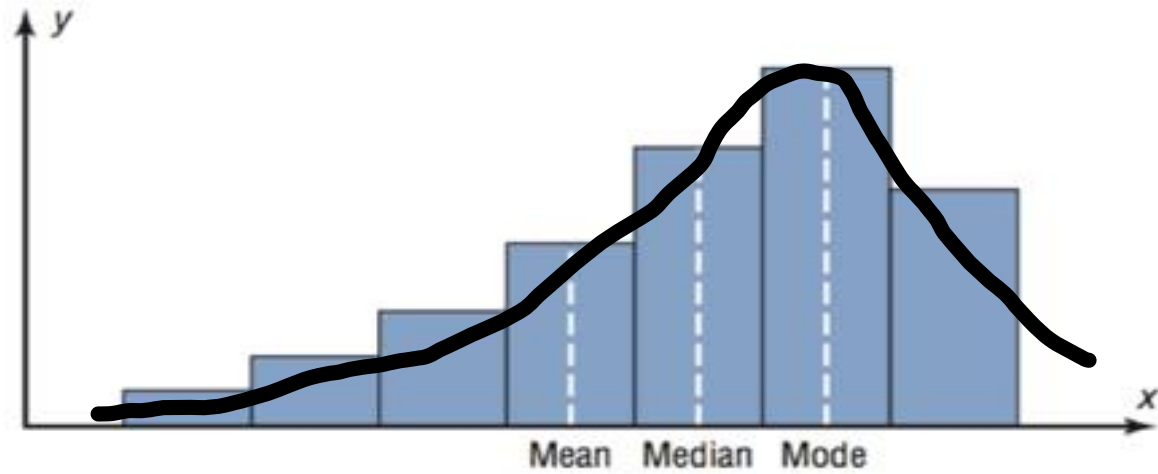
No mode

- If the data have multiple values that are tied for occurring the most frequently, you have a multimodal distribution.
- If no value repeats, the data do not have a mode.



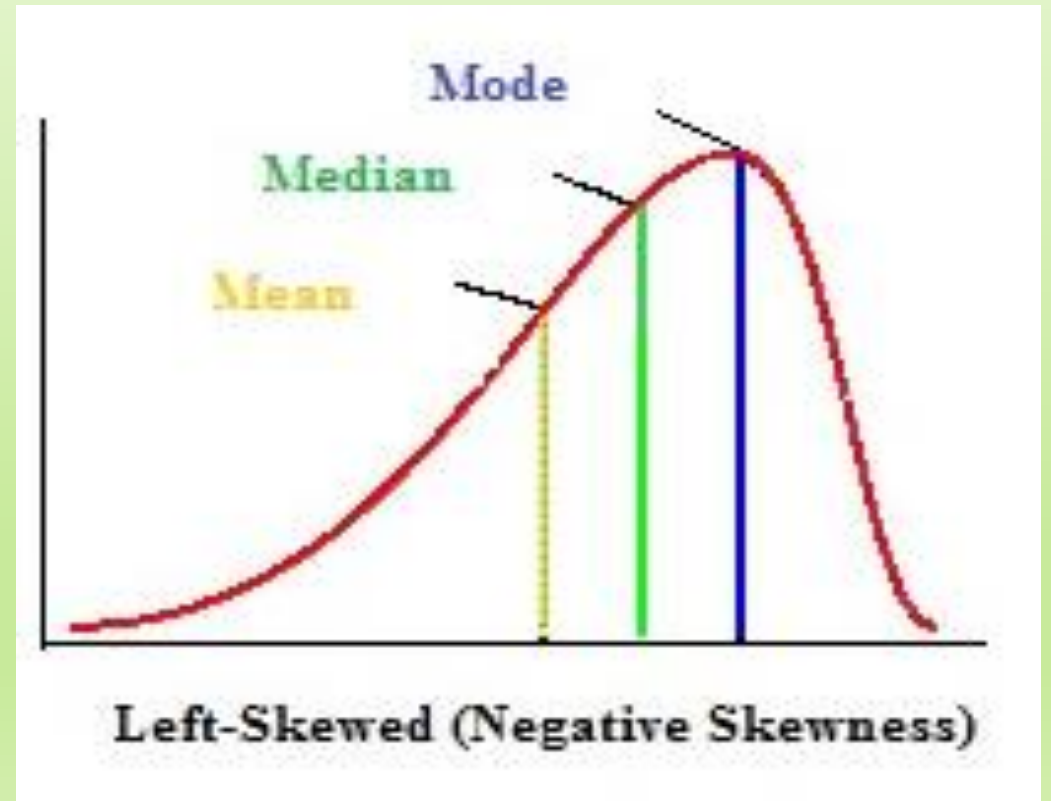
Mode < Median < Mean

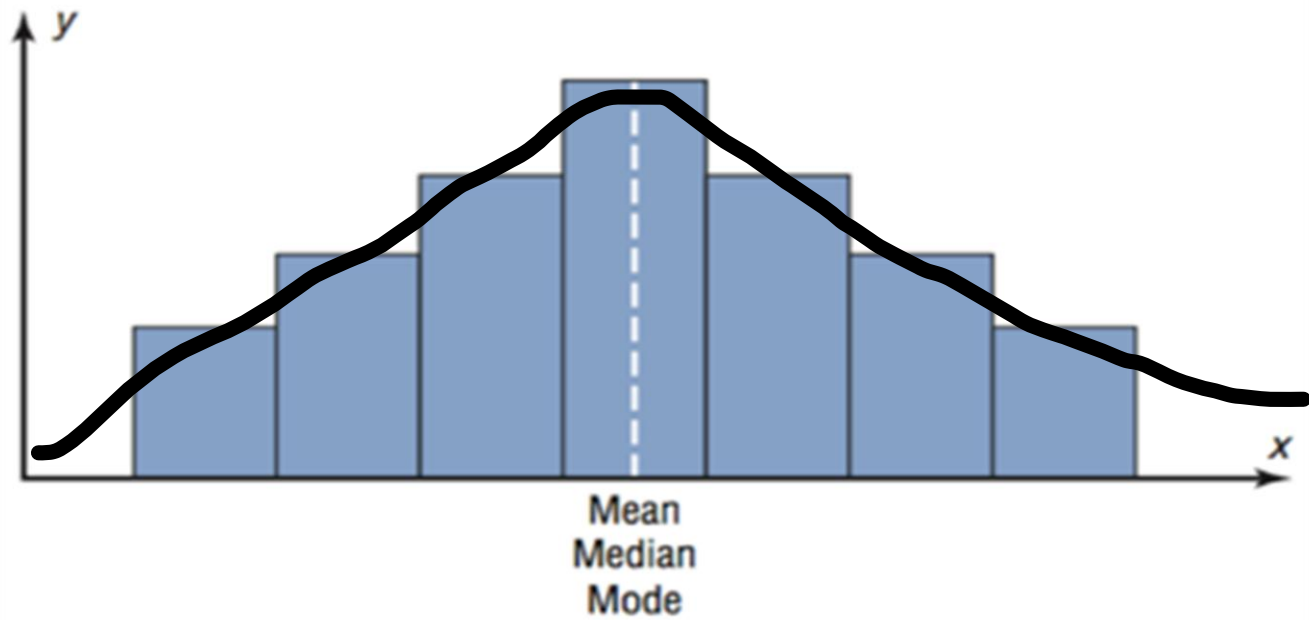




(c) Negatively skewed or left-skewed

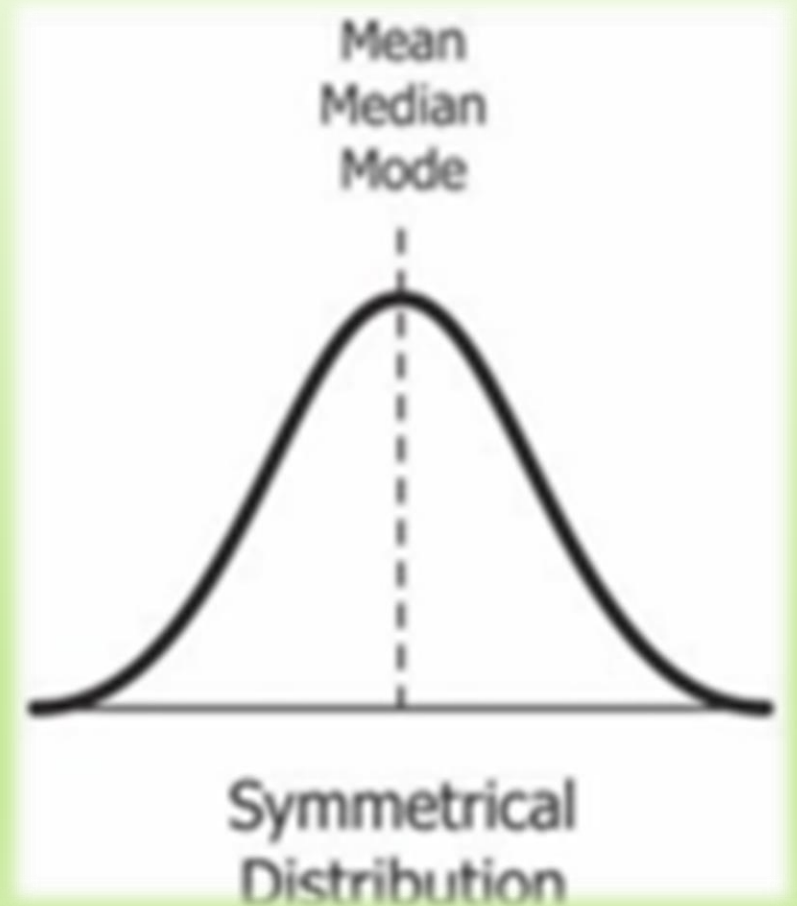
$$\text{Mode} > \text{Median} > \text{Mean}$$





(b) Symmetric

$$\text{Mode} = \text{Median} = \text{Mean}$$



Note that:

- **In a symmetric distribution, the mean locates the center accurately.** However, in a skewed distribution, the mean can miss the mark. In the histogram above, it is starting to fall outside the central area. This problem occurs because outliers have a substantial impact on the mean. **Extreme values in an extended tail pull the mean away from the center. As the distribution becomes more skewed, the mean is drawn further away from the center.**
- **When to use the mean: Symmetric distribution.**
- **When you have a symmetrical distribution for continuous data, the mean, median, and mode are equal. In this case, analysts tend to use the mean because it includes all of the data in the calculations.**
- **When you have a skewed distribution, the median is a better measure of central tendency than the mean.**
- **For categorical data, you must use the mode.**

SHEET (2)

9. The lengths of time, in minutes, that 10 patients waited in a doctor's office before receiving treatment were recorded as follows: 5, 11, 9, 5, 10, 15, 6, 10, 5, and 10. Treating the data as a random sample, find (a) the mean; (b) the median; (c) the mode.

Answer

(a) Mean = $\frac{5+11+9+\dots+10}{10} = 8.6$

(b) 5 5 5 6 9 10 10 10 11 15
 $n = 10$ " even "

Median = $\frac{x_5+x_6}{2} = \frac{9+10}{2} = 9.5$

(c) Mode = 5, 10.

Sheet (2): 8, 10.

(B) MEASURE OF DISPERSION :

- The measures of central tendency are not adequate to describe data. Two data sets can **have the same mean, but they can be entirely different.** Thus to describe data, one needs to know the extent of variability. This is given by the measures of dispersion.
- In statistics, **dispersion** (also called **variability, scatter, or spread**) is the extent to which a distribution is stretched or squeezed.
- The three most common measures of dispersion are the Range, interquartile range, and (variance , standard deviation).

(1) Range (R):

- The range is the difference between the largest and the smallest observation in the data.

Range ‘ R ’ = largest value – smallest value

- Advantage : Is that it is easy to calculate.
- Disadvantages : It is very sensitive to outliers and does not use all the observations in a data set.

Example :- **Statistics grade (A) : 20, 30, 35, 45** $\longrightarrow R_A = 45 - 20 = 25$

Statistics grade (B) : 1, 30, 35, 45 $\longrightarrow R_B = 45 - 1 = 44$

Statistics grade (C) : 20, 30, 35, 100 $\longrightarrow R_C = 100 - 20 = 80$

Who had the greater spread of marks?

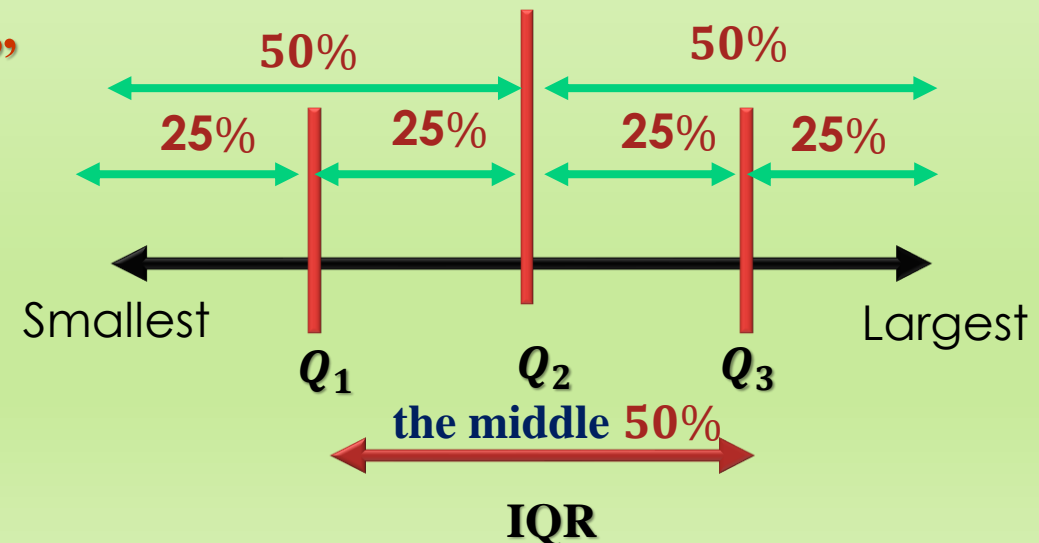
c has a great spread of marks, because $R_C > R_B > R_A$

(2) Interquartile Range (IQR):

- The interquartile range describes the middle 50% of observations.
- If the interquartile range is large it means that the middle 50% of observations are spaced wide apart.
- Advantage : Is that it is not affected by extreme values.
- Disadvantages : It does not use all the observations in a data set.

□ To get the value of the interquartile range, follow these steps:

- put the data in order “ from smallest to largest”
- Find the median of all data : Q_2
- Find the median of the data before Q_2 : Q_1
- Find the median of the data after Q_2 : Q_3
- $IQR = Q_3 - Q_1$



- Example (1) :- Statistics grade : 60, 35, 25, 45, 30, 37, 45, 10

(1) 10 25 30 35 37 45 45 60

(2) $n = 8$ " even $Q_2 = \frac{x_4 + x_5}{2} = \frac{35 + 37}{2} = 36$

(3) 10 25 30 35

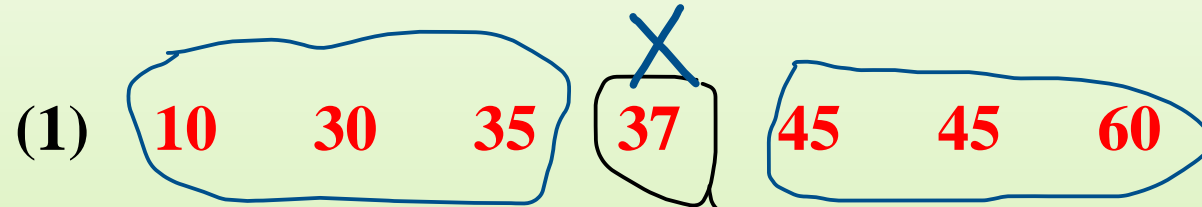
$$Q_1 = \frac{25 + 30}{2} = 27.5$$

(4) 37 45 45 60

$$Q_3 = \frac{45 + 45}{2} = 45$$

(5) $IQR = Q_3 - Q_1 = 45 - 27.5 = 17.5$

- Example (2):- Statistics grade : 60, 35, 45, 30, 37, 45, 10



(2) $n = 7$ " odd, $Q_2 = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{8}{2}\right)} = x_{(4)} = 37$

(3) 10 **30** 35

$Q_1 = 30$

(5) $IQR = Q_3 - Q_1 = 45 - 30 = 15$

(4) 45 **45** 60


$Q_3 = 45$

(3) Variance (S^2) and Standard deviation (S):

Variance (S^2):

- It measures how far a set of numbers is spread out from their average value. The Variance is the average of the squared differences from the Mean.
- Let x_1, x_2, \dots, x_n be a sample of size n . To calculate the variance follow these steps:
 - Work out the Mean (the simple average of the numbers) $\rightarrow \bar{X}$
 - Then for each number: subtract the Mean and square the result (the squared difference) $\rightarrow (x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \dots + (x_n - \bar{X})^2 = \sum_{i=1}^n (x_i - \bar{X})^2$.
 - Then work out the average of those squared differences.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n - 1}$$

 the data is a **Sample** (a selection taken from a bigger Population).

Standard deviation (S^2):

- The Standard Deviation is a measure of how spread out numbers. It is a measure of how far typical value tends to be from the mean.
- It is the square root of the Variance $\rightarrow S = \sqrt{S^2}$.

Example : 9, 2, 5, 4, 12, 7.

Find the Standard Deviation?

$$\square n = 6.$$

$$\square \bar{X} = \frac{\sum_{i=1}^6 x_i}{6} = \frac{9+2+\dots+7}{6} = 6.5$$

$$\square \sum_{i=1}^6 (x_i - \bar{X})^2 = (9 - 6.5)^2 + (2 - 6.5)^2 + \dots + (7 - 6.5)^2 = 65.5$$

$$\square S^2 = \frac{\sum_{i=1}^6 (x_i - \bar{X})^2}{6-1} = \frac{65.5}{5} = 13.1$$

$$\square S = \sqrt{S^2} = \sqrt{13.1} = 3.619$$

SHEET (2)

11. The following measurements were recorded for the drying time, in hours, of a certain brand of latex paint.

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

- (a) Calculate the sample mean for these data.
(b) Calculate the sample median.
(c) Compute the sample variance and sample standard deviation.

Answer

$$n = 5 \times 3 = 15$$

(a) Sample Mean $\bar{X} = \frac{\sum_{i=1}^{15} x_i}{15} = \frac{3.4+2.5+\dots+4.8}{15} = 3.787$

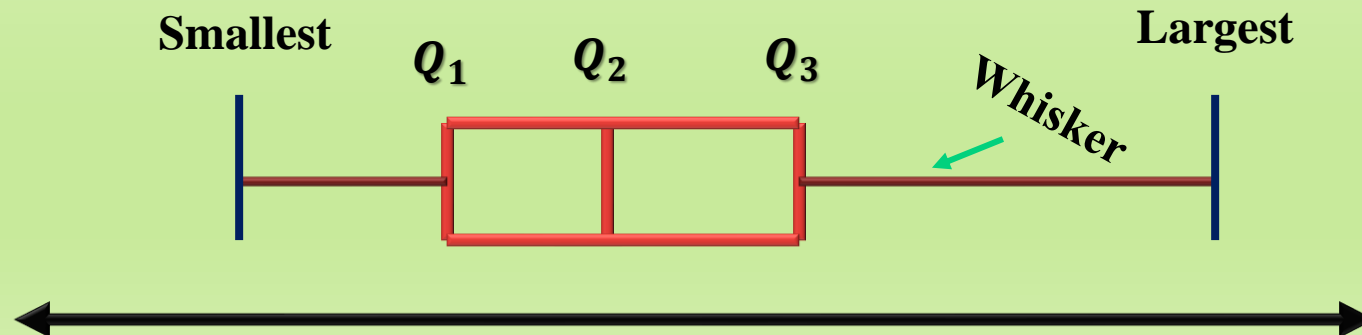
(b) **2.5 2.8 2.8 2.9 3.0 3.3 3.4 3.6 3.7 4.0 4.4 4.8 4.8 5.2 5.6**
n : odd \rightarrow Median = $x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{16}{2}\right)} = x_{(8)} = 3.6$

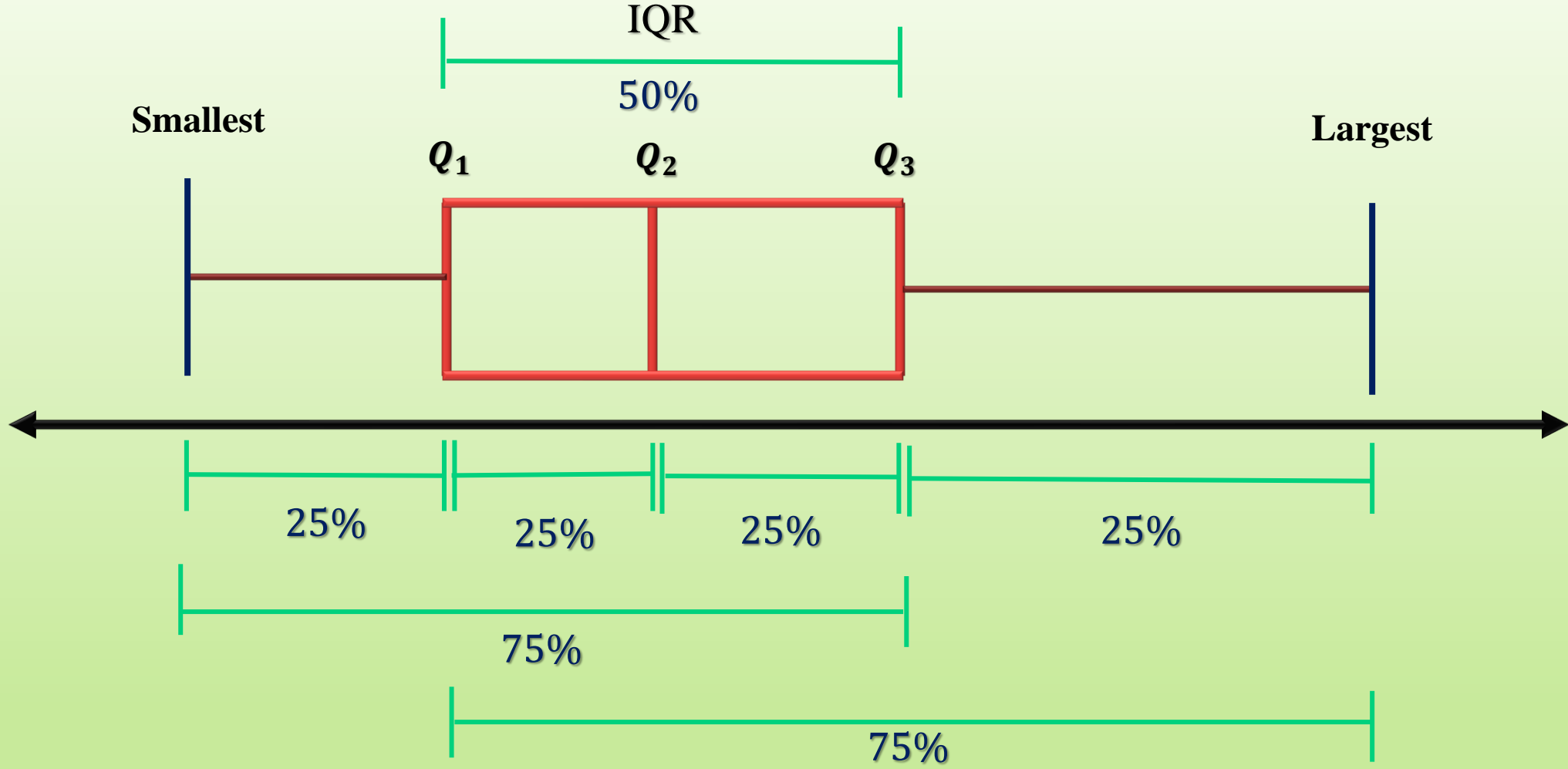
(c) $\sum_{i=1}^{15} (x_i - \bar{X})^2 = (3.4 - 3.787)^2 + \dots + (4.8 - 3.787)^2 = 13.19733$

$$S^2 = \frac{\sum_{i=1}^{15} (x_i - \bar{X})^2}{15-1} = \frac{13.19733}{14} = 0.943 \rightarrow S = \sqrt{S^2} = \sqrt{0.943} = 0.971$$

BOX PLOT

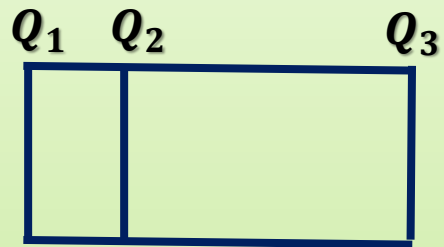
- A boxplot, also called a box and whisker plot, is a way to show the spread and skewness of a data set.
- **Steps:**
 - put the data in order “ from smallest to largest”
 - Find the median of all data : Q_2 “ *Median* “
 - Find the median of the data before Q_2 : Q_1 “ *Lower Quartile* “
 - Find the median of the data after Q_2 : Q_3 “ *upper Quartile* “



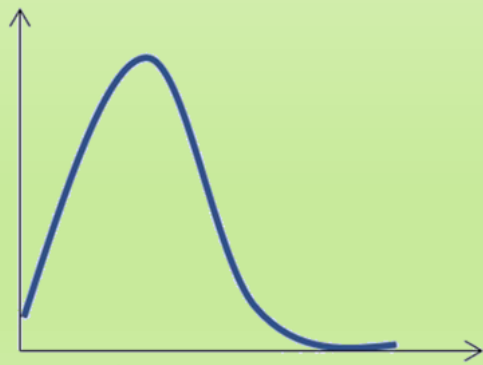


- The lines extending parallel from the boxes are known as the “**whiskers**”, which are used to indicate variability outside the upper and lower quartiles.

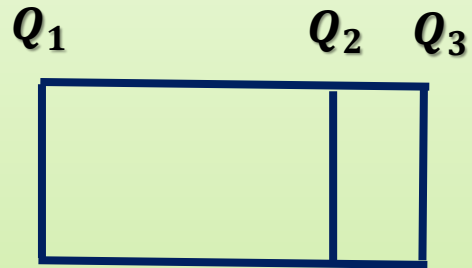
- Skewness:**



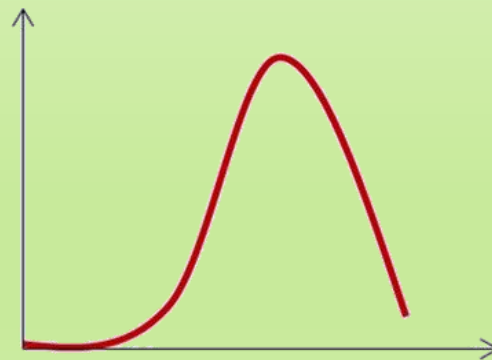
$$Q_3 - Q_2 > Q_2 - Q_1$$



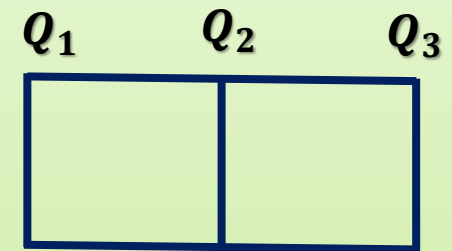
+ve, right skew



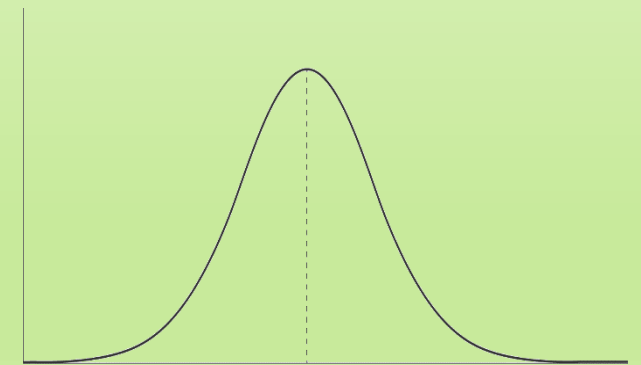
$$Q_3 - Q_2 < Q_2 - Q_1$$



-ve, left skew



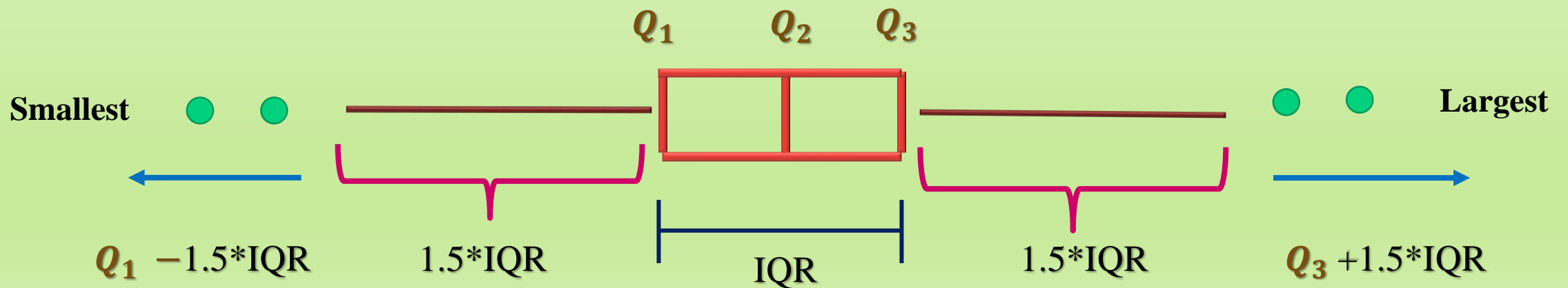
$$Q_3 - Q_2 \cong Q_2 - Q_1$$



Symmetric, Normal

Outliers:

- **Outliers are data points that are far from other data points. In other words, they're unusual values in a dataset.**
- **Calculating the Outlier Fences Using the Interquartile Range :**
 - **any value $> Q_3 + 1.5 * IQR$**
 - **any value $< Q_1 - 1.5 * IQR$**



SHEET (2)

2. Draw a box plot for the following set of data.

4.7, 3.8, 3.9, 3.9, 4.6, 4.5, 5

Answer

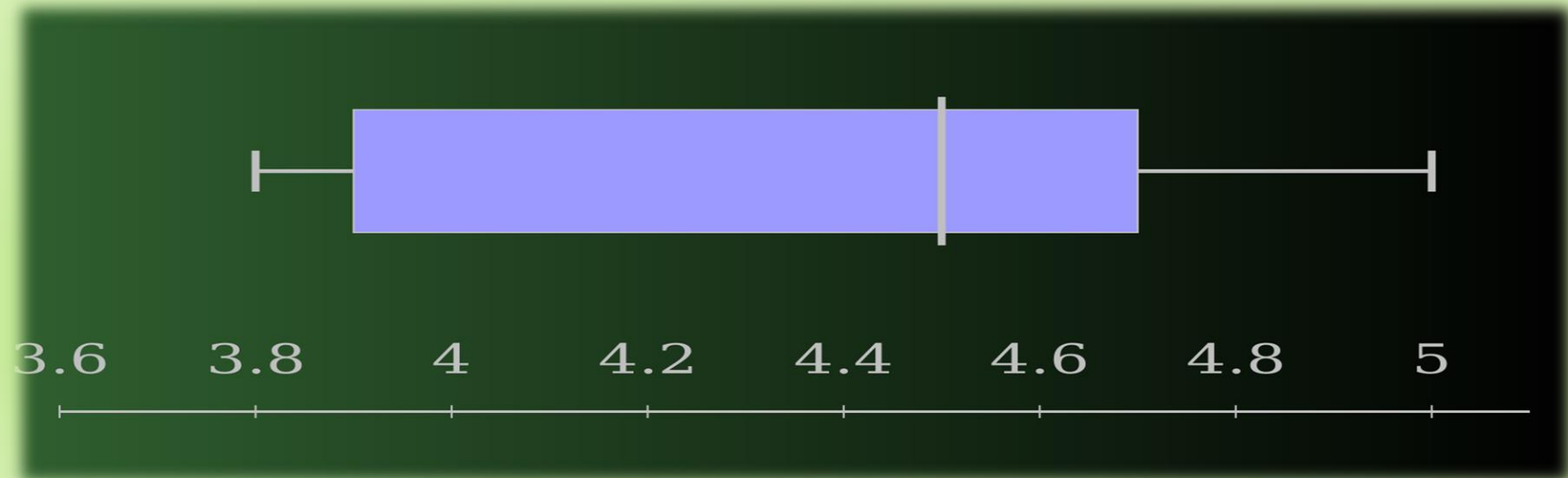
- Organize the data from the smallest to largest:

3.8 (3.9) 3.9 (~~4.5~~) 4.6 (4.7) 5

- Smallest = 3.8

Largest = 5

- $Q_2 = 4.5$
- $Q_1 = 3.9$
- $Q_3 = 4.7$



SHEET (2)

6. For the following list of numbers, draw a box and whisker plot showing outliers values:

-6, 4.5, 5.5, 6.5, 8.5, 10, 12, 30

Answer

• **Smallest = -6** ,

Largest = 30

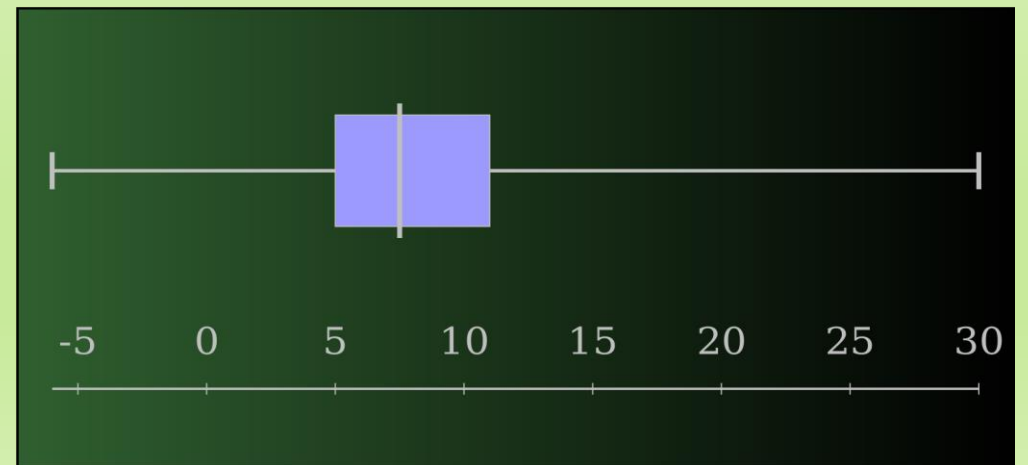
-6 4.5 5.5 6.5 8.5 10 12 30

← 7.5 →

• $Q_2 = \frac{6.5+8.5}{2} = 7.5$

• $Q_1 = \frac{4.5+5.5}{2} = 5$

• $Q_3 = \frac{10+12}{2} = 11$



• **Outliers:**

➤ $IQR = Q_3 - Q_1 = 11 - 5 = 6$

➤ $1.5 * IQR = 1.5 * 6 = 9$

➤ any value $> Q_3 + 1.5 * IQR$
 $> 11 + 9$
 $> 20 \longrightarrow 30$

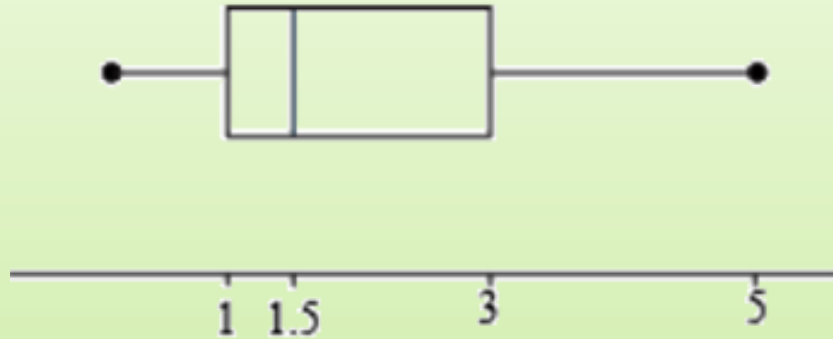
➤ any value $< Q_1 - 1.5 * IQR$
 $< 5 - 9$
 $< -4 \longrightarrow -6$

Outliers = 30, -6

Sheet (2): 1, 3, 5, 7.

SHEET (2)

4. The box and whisker plot below was drawn using a list of numbers (data). Determine if each statement is definitely true, definitely false, or cannot be determined.

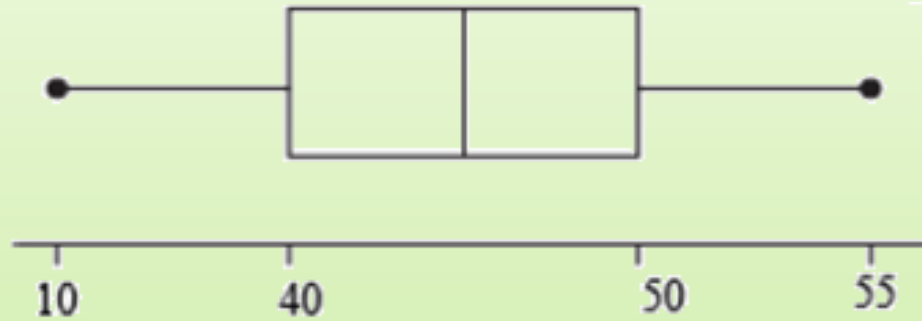


$$\begin{aligned}Q_1 &= 1 \\Q_2 &= 1.5 \\Q_3 &= 3 \\ \text{Largest} &= 5\end{aligned}$$

- (a) **Half the data falls between 1 and 3. (definitely true)**
- (b) **The number 5 must be in the list of numbers from which this plot was drawn. (definitely true)**
- (c) **The number 1.5 must be in the list of numbers from which this plot was drawn. (cannot be determined)**

SHEET (2)

Which must be in the list of numbers from which this box and whisker plot was drawn?



$$Q_1 = 40$$

$$Q_3 = 50$$

$$\text{Smallest} = 10$$

$$\text{Largest} = 55$$

(a) 10

(b) 40

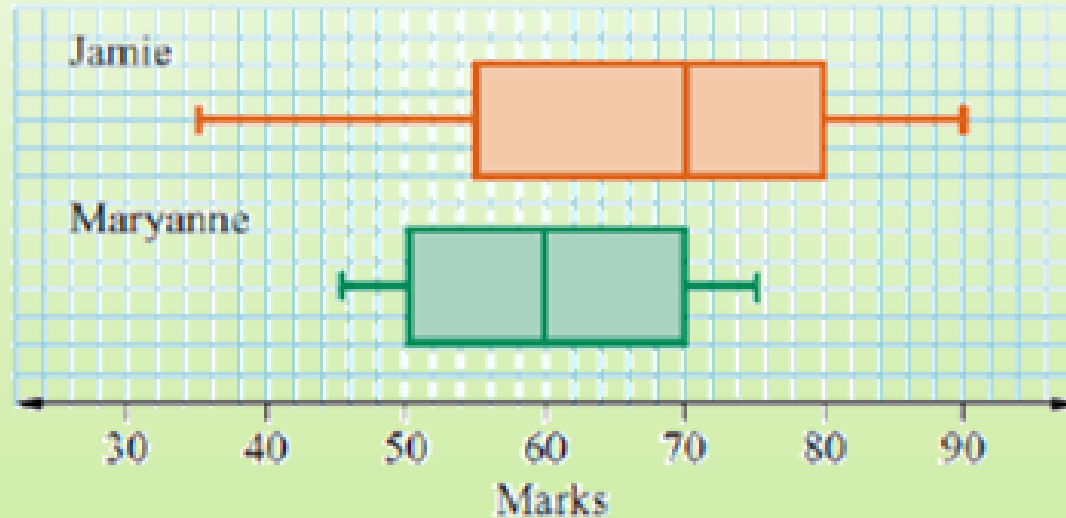
(c) 50

(d) 55

(e) 65

SHEET (2)

19. A class has eight assessment tasks over a year. The parallel box plots show the summary of the marks for the assessments for two students, Jamie and Maryanne.



Jamie
 $Q_1 = 55$
 $Q_2 = 70$
 $Q_3 = 80$
Smallest = 35
Largest = 90

Maryanne
 $Q_1 = 50$
 $Q_2 = 60$
 $Q_3 = 70$
Smallest = 45
Largest = 75

- (i) Who scored the highest mark? (Jamie)
- (ii) Who scored the lowest mark? (Jamie)

(iii) What was the range of marks for each student?

$$R_{\text{Jamie}} = 90 - 35 = 55$$

$$R_{\text{Maryanne}} = 75 - 45 = 30$$

(iv) Who had the greater spread of marks?

$$R_{\text{Jamie}} > R_{\text{Maryanne}}$$

(Jamie)

(v) What was the interquartile range for each student?

$$IQR_{\text{Jamie}} = 80 - 55 = 25$$

$$IQR_{\text{Maryanne}} = 70 - 50 = 20$$