Empirical Relations of dispersion

The Measures- For symmetrical, bell-shaped frequency distribution (also called Normal Curve), the range within which a given percentage of values of the distribution are likely to fall within a specified number of the standard deviation of the mean is determined as follows:

$\mu \pm Q.D$ covers approximately 50% of values in the data set

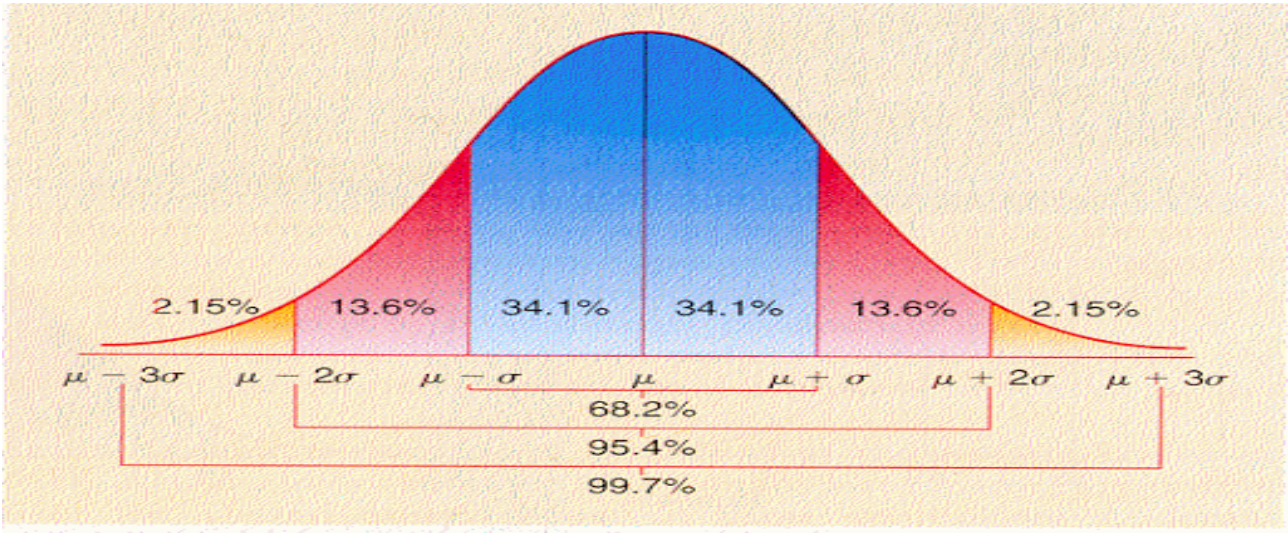
$\mu \pm M.D$ covers approximately 57.5% of values in the data set

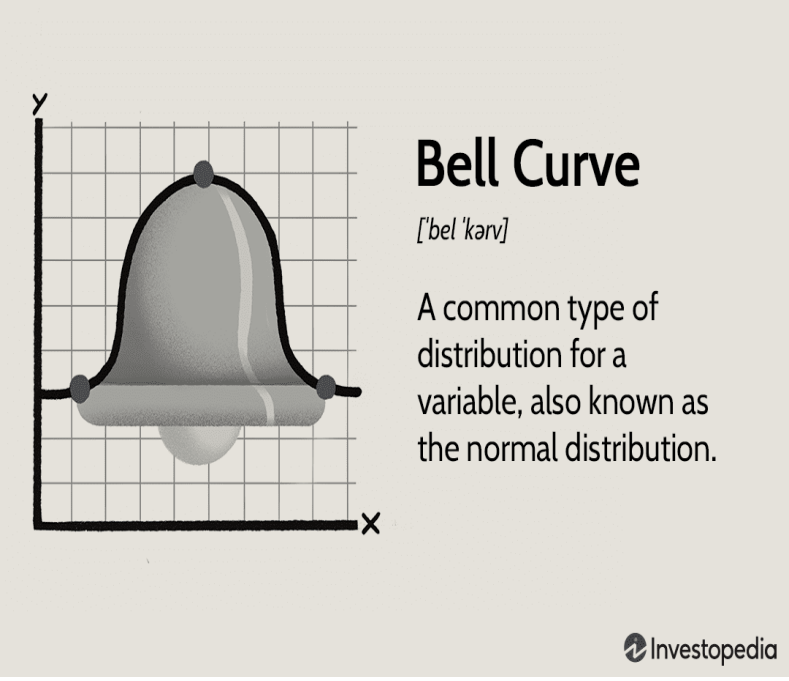
$\mu \pm S.D\ (\sigma)$ covers approximately 68.27% of values in the data set

$\mu \pm 2\sigma$ covers approximately 95.45% of values in the data set

$\mu \pm 3\sigma$ covers approximately 99.73% of values in the data set

Figure





Test Yourself

The following data give the number of passengers travelling by airplane from one city to another in one week.

115 112 129 113 119 124 132 120 110 116

Calculate the mean and standard deviation and determine the percentage of class that lie between

$(i)\mu \pm \sigma$; $(ii)\mu \pm 2\sigma$; and $(iii)\mu \pm 3\sigma$.

What percentage of cases lie outside these limits?

Answer: The calculation for mean and standard deviation are given in the following table-

x	$x - \mu$	$(x - \mu)^2$
115	-5	25
122	2	4
129	9	81
113	-7	49
119	-1	1
124	4	16
132	12	144
120	0	0
110	-10	100
116	-4	16

$$\mu = \frac{\sum x}{N} = \frac{???}{?} = 120 \text{ and}$$

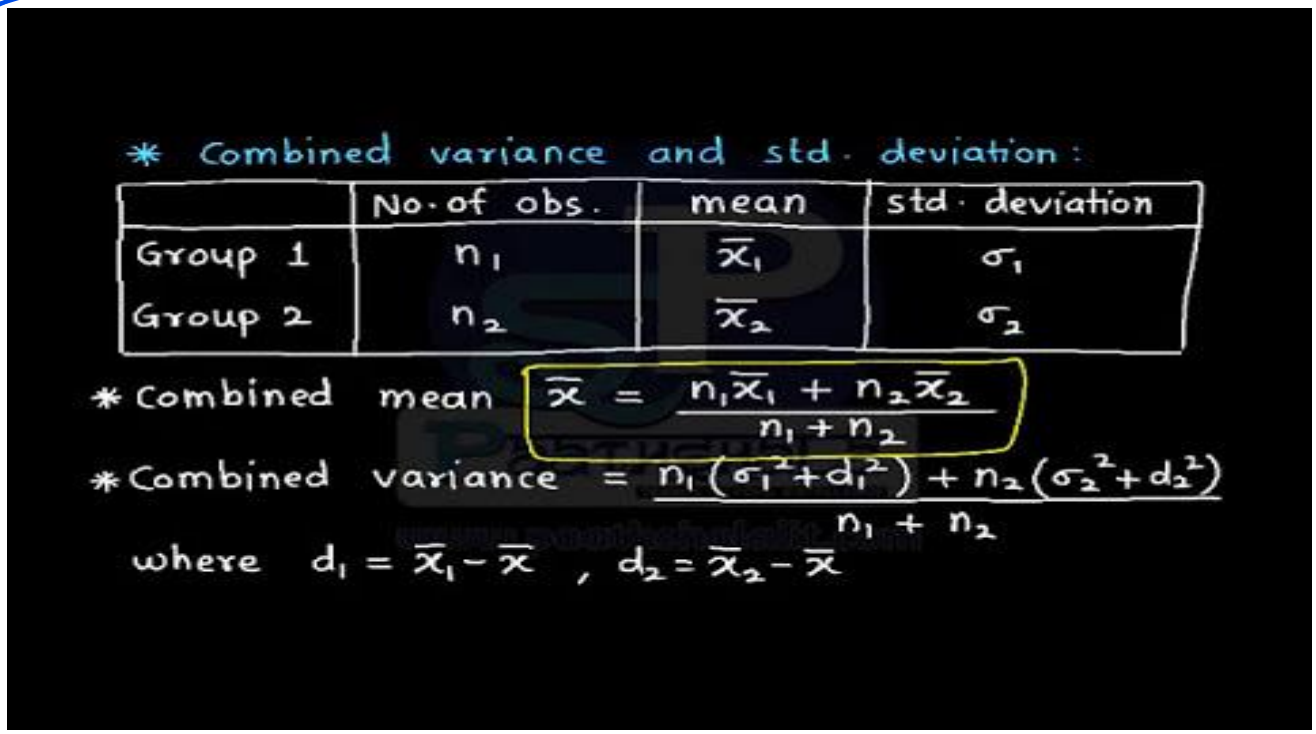
$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} = 43.6$$

Therefore, $\sigma = \sqrt{\sigma^2} = \sqrt{43.3} = 6.60$

Noted that: We can also find out variance as, $\sigma^2 = \frac{\sum x^2}{N} - \bar{x}^2$ [for ungroup data]

$$\sigma^2 = \frac{\sum f x^2}{N} - \bar{x}^2 \text{ [for group data]}$$

Combined mean & standard deviation for two or more group:

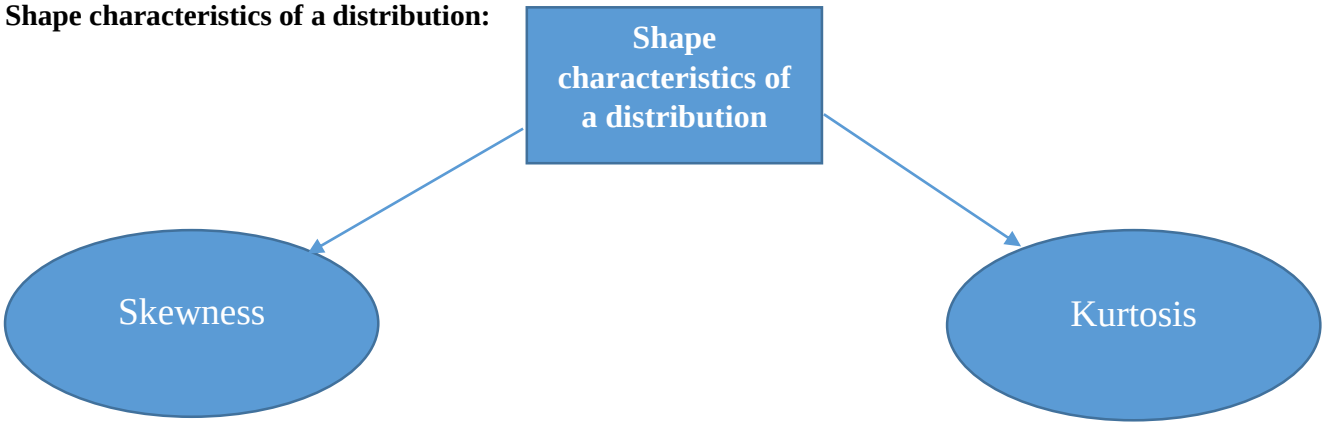


Empirical Relationship between Quartile Deviation (Q.D), Mean Deviation (M.D), and Standard deviation (S.D):

4 S.D = 5 M.D = 6 Q.D

- For two unequal observations, Mean Deviation=Standard Deviation=$\frac{\text{Range}}{2}$
- For "n" natural number mean = $\frac{n+1}{2}$, and variance = $\frac{n^2-1}{12}$
- Coefficient of the first "n" natural number is, C.V= $\frac{1}{\sqrt{3}} \times \sqrt{\frac{n-1}{n+1}} \times 100\%$

Shape characteristics of a distribution:



- The study of shape characteristics of a distribution is of crucial importance in comparing a distribution with other distributions.

- By shape characteristic of a distribution we refer to the extent of its **asymmetry and peakedness** relative to an agreed upon standard.
- The study of these two characteristics (i.e. asymmetry and peakedness) is accomplished through what is known as the measures of **skewness and kurtosis,** respectively.

Skewness: The term skewness means the lack of symmetry; it may be either positive or negative. When the skewness is positive the associated distribution is called positively skewed. When the skewness is negative the associated distribution is negatively skewed. There are 4 methods of measuring skewness-

i) Skewness according to Karl Pearson
ii) Skewness according to Bowley
iii) Skewness according to Kelly
iv) Skewness based on moments

There are 2 types of skewness-

i) Absolute measures of skewness.
ii) Relative measure of skewness.

Absolute measures of skewness- The measure of skewness which depend on unit of variable is known as absolute measure. Karl Pearson suggested absolute measure of skewness on the basis of mean, median, & mode.

1. **Pearson's Skewness-** Simply, for a distribution, If

- Mean = Median= Mode: The distribution is symmetric

If, Mean≠Median≠ Mode: The distribution is Asymmetric

- Mean > Median > Mode: The distribution is positively skewed
- Mean < Median < Mode: The distribution is negatively skewed

SKp = Mean − Mode = Mean – Median

Then if,

- Sk_p = 0 : The distribution is symmetric

If, $Sk_p \neq 0$: The distribution is Asymmetric

- Sk_p > 0 : The distribution is positively skewed
- Sk_p < 0 : The distribution is negatively skewed

2. **Bowley suggested another absolute measure of skewness. According to Bowley-**

$$Sk_b = (Q_3 - Q_2) - (Q_2 - Q_1) = (Q_3 - Median) - (Median - Q_1)$$

3. **Kelly suggested another absolute measure of skewness. According to Kelly-**

$$Sk_k = D_{90} - 2D_5 + D_1 = P_{90} - 2P_{50} + P_{10}$$

The comment will be same as SKp for Sk_b & Sk_k.

Relative measures of skewness: The relative measure of skewness which is free from unit of variable. The relative measure is also known as coefficient of skewness. This measure is used to compare the asymmetry of two or more curve.

1. **Pearson's coefficient for Skewness,** $\text{PCS} = \frac{\text{Mean} - \text{Mode}}{\text{SD}} = \frac{3\,(\text{Mean} - \text{Median})}{\text{SD}}$

Then if,

- $PCS > 0$: The distribution is positively skewed
- $PCS < 0$: The distribution is negatively skewed
- $PCS = 0$: The distribution is symmetric

2. **Bowley's coefficient for Skewness** i.e. quartile skewness coefficient, $\text{BCS} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$

$$BCS = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

Then if,

- $BCS > 0$: The distribution is positively skewed
- $BCS < 0$: The distribution is negatively skewed
- $BCS = 0$: The distribution is symmetric

Note: $\mathbf{-1 \leq PCS, BCS \leq 1}$

3. **Kelly's coefficient for Skewness** i.e. Decile skewness coefficient, $KCS = \frac{D_9 - 2D_5 + D_1}{D_9 - D_1}$

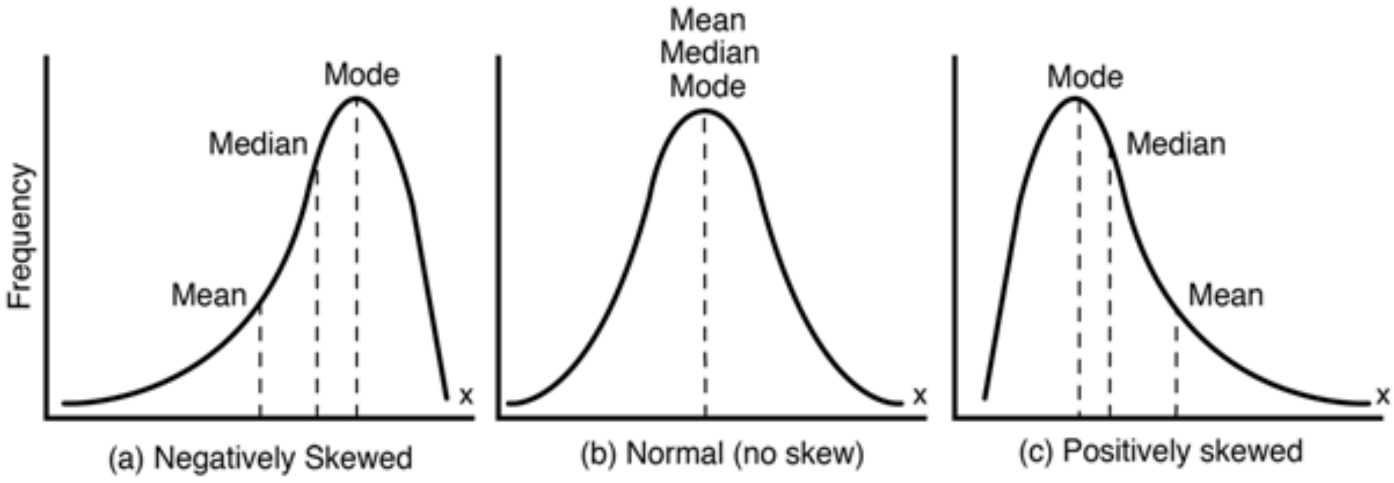
Percentiles skewness coefficient, $KCS = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$

Note: $\mathbf{[Q_2 = D_5 = P_{50} = Median]}$

Same comment will be for KCS.

- $KCS > 0$: The distribution is positively skewed
- $KCS < 0$: The distribution is negatively skewed
- $KCS = 0$: The distribution is symmetric

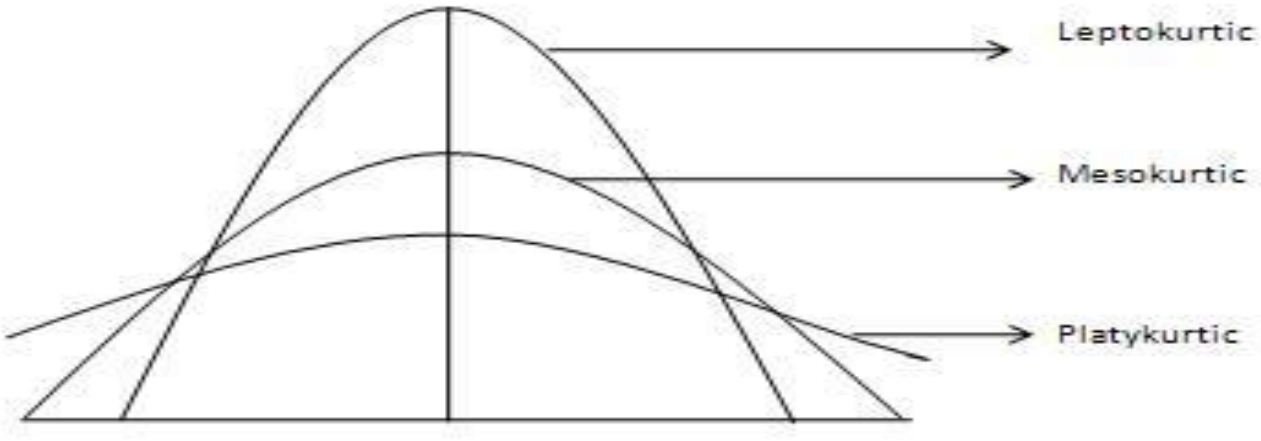
Skewness of Distribution



Kurtosis

- There is considerable variation among symmetrical distributions.
- For instance, they can differ markedly in terms of peakedness. This is what we call kurtosis.
- Kurtosis, as defined by Spiegel, is the degree of peakedness of a distribution, usually taken in relation to a normal distribution.

Kurtosis of Distribution



Summary:

- Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.
- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. It helps to determine if the curve is more or less high as compared to the normal curve.

Test Yourself

1. If for a distribution Mean =18, Median= 32, and Mode= 36, the distribution is _____ skewed.

a. Positively, b. Symmetrically, c. None, d. Negatively

2. If for a distribution Mean = 20, Median= 26.4, and SD= 3.3, the distribution is _____ skewed.

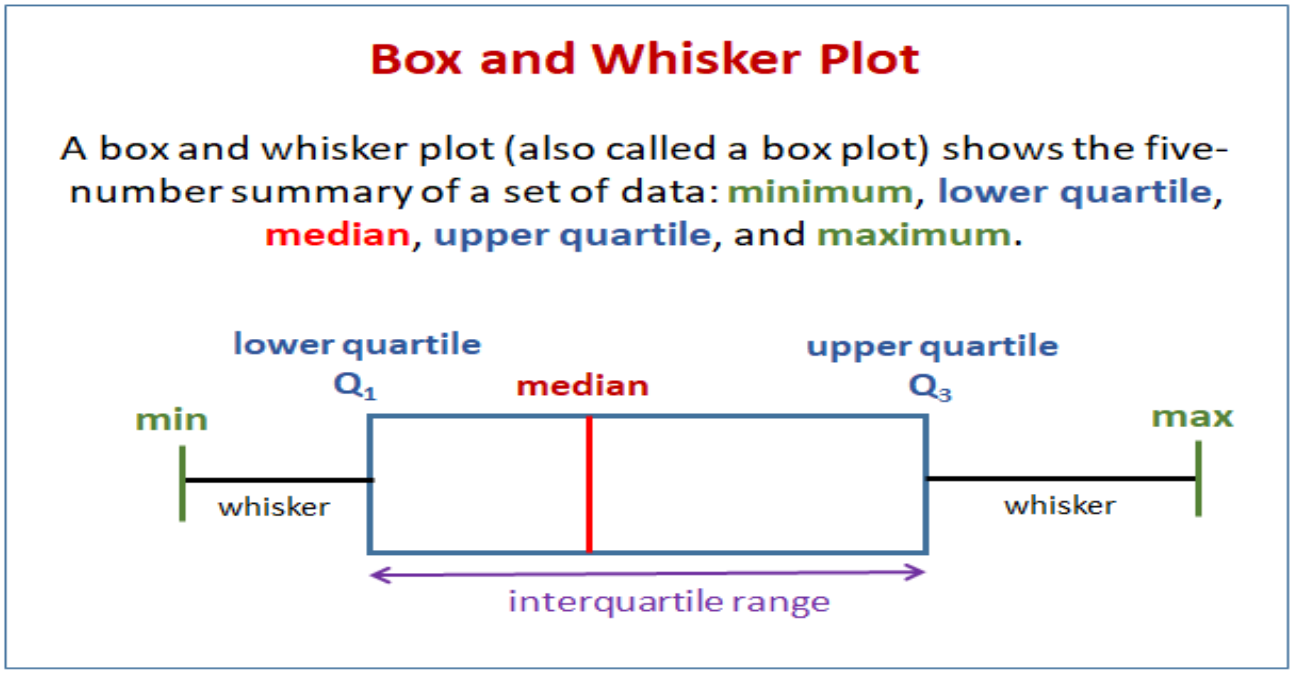
a. Positively, b. Symmetrically, c. None, d. Negatively

3. If for a distribution, Mean = 35.6, Mode = 24, and SD= 5.2, what is the skewness coefficient of the distribution?

Box & Whisker Plot: A box and whisker plot is defined as a graphical method of displaying variation in a set of data. In most cases, a histogram analysis provides a sufficient display, but a box and whisker plot can provide additional detail while allowing multiple sets of data to be displayed in the same graph.

A box plot is a graphic display that shows the general shape of a variable's distribution. It is based on five descriptive statistics:

1. The minimum value,
2. The first quartile (Q_1),
3. Median, Q_2
4. Third quartile (Q_3), and
5. The maximum value

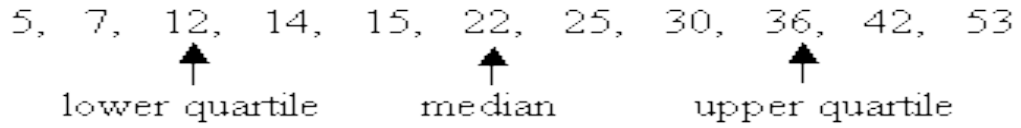


Drawing A Box & Whisker Plot

Example: Construct a box plot for the following data: 12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25

Solution: Step 1: Arrange the data in ascending order.

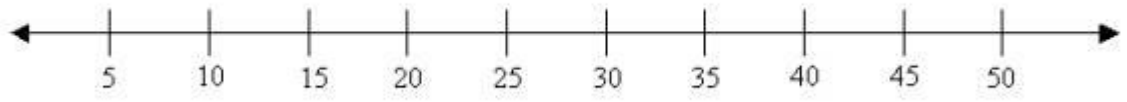
Step 2: Find the median, lower and upper quartile.



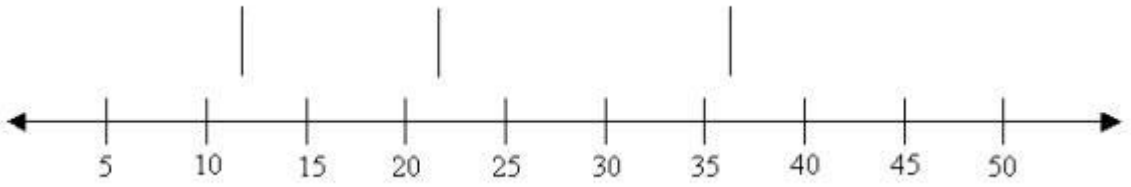
Median (middle value) = 22
Lower quartile (middle value of the lower half) = 12
Upper quartile (middle value of the upper half) = 36

(If there is an even number of data items, then we need to get the average of the middle numbers.)

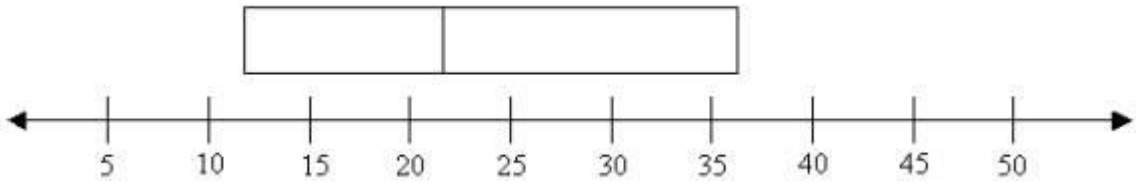
Step 3: Draw a number line that will include the smallest and the largest data.



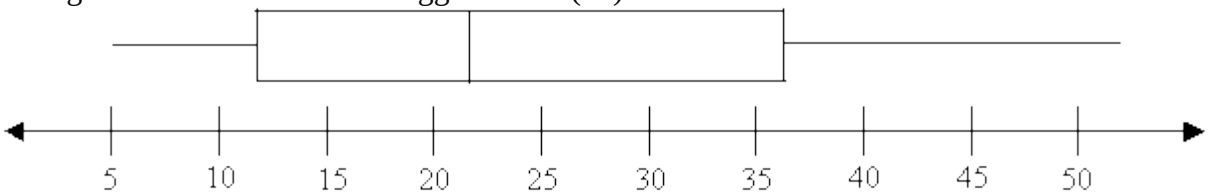
Step 4: Draw three vertical lines at the lower quartile (12), median (22) and the upper quartile (36), just above the number line.



Step 5: Join the lines for the lower quartile and the upper quartile to form a box.



Step 6: Draw a line from the smallest value (5) to the left side of the box and draw a line from the right side of the box to the biggest value (53).



How to interpret a box and whisker plot?

Box and Whisker Plots are graphs that show the distribution of data along a number line. We can construct box plots by ordering a data set to find the median of the set of data, median of the upper and lower quartiles, and upper and lower extremes. We can draw a Box and Whisker plot and use box plots to solve a real world problem. By finding the middle values of the ordered data set, you have separated the data into four equal groups called quartiles. A shorter

distance means the quartile data is bunched together. A longer distance means the quartile data is spread out.

For example, if your job is to compare the annual snowfall between two ski resorts for the past 50 years, you would need a way to summarize all the data. A box plot displays the range and distribution of data along a number line.

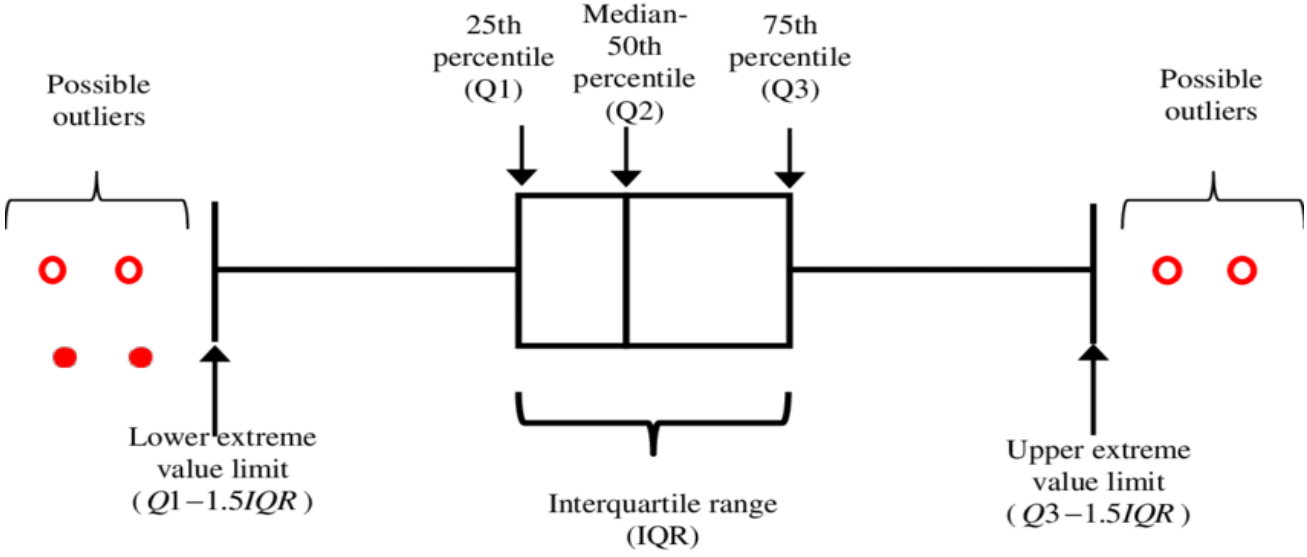
How To Make A Box Plot From A Set Of Data?

1. Order the data from least to greatest.
2. Find the median or middle value that splits the set of data into two equal groups. If there is no one middle value, use the average of the two middle values as the median.
3. Find the median for the lower half of the data set.
4. Find the median for the upper half of the data set.
5. Use these five values to construct a box plot: lower extreme, lower quartile, median, upper quartile, upper extreme.
6. Plot the points of the five values above a number line.
7. Draw vertical lines through the lower quartile, median and upper quartile.
8. Form a box by connecting the vertical lines from the lower quartile, median, and upper quartile.
9. Draw the whiskers from the extremes to the box.

Outliers In A Box And Whiskers Plot

What is an outlier?
An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.

- Inter-Quartile Range (IQR) is the distance between the first and second quartiles.
- Multiply the IQR by 1.5.
- Subtract that value from the 1st Quartile to get your lower boundary.
- Add that value to the 2nd Quartile to get your upper boundary.
- Values in the data set that fall outside of these limits are considered outliers.



Understanding & Comparing Boxplots (Box and Whisker Plots)

Box and whisker plots are graphical displays of the five number summary (minimum, quartile 1, median, quartile 3, and maximum). Compare two boxplots and see how larger spread makes predictions more difficult. Check for evidence of claim using the boxplots.

Example: Pizza Hut offers free delivery of its pizza within 15 miles. Mr. Rahman, the owner, wants some information on the time it takes for delivery e.g. how long does a typical delivery take, or within what range of times will most deliveries be completed. For a sample of 20 deliveries, he determined the following information:

Minimum value = 13 minutes

Q_1 = 15 minutes

Median = 18 minutes

Q_3 = 22 minutes

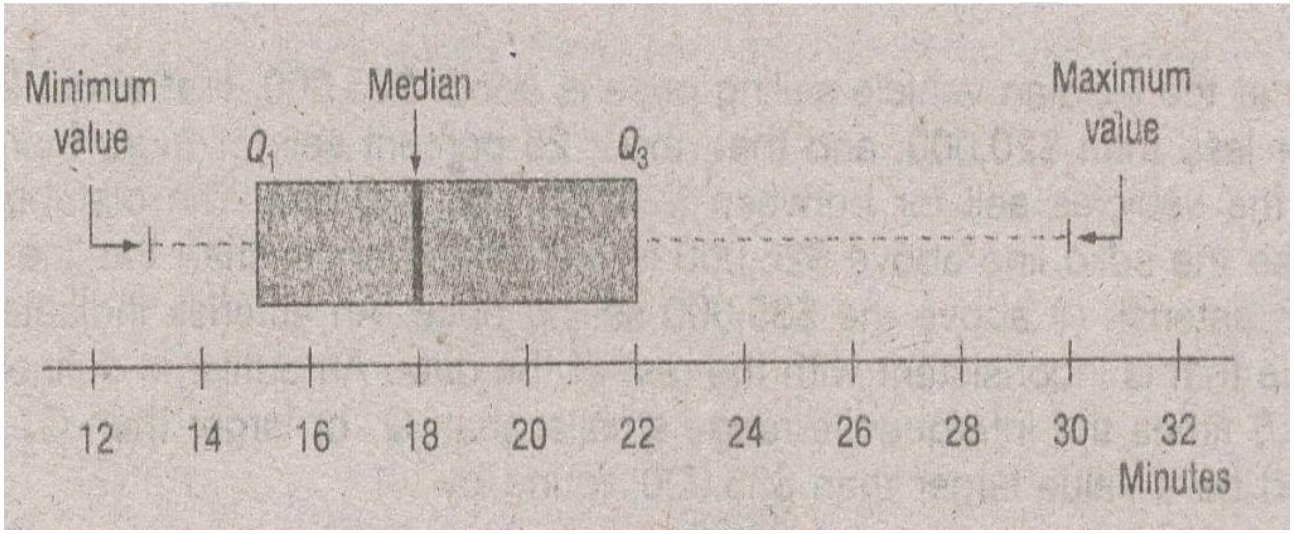
Maximum value = 30 minutes

Develop a box-plot for the delivery times. What conclusions can you make about the delivery times?

Solution: In order to draw box plot follow the steps mentioned below:

- Step 1: Create an appropriate scale along the horizontal axis.
- Step 2: Draw a box that starts at Q_1 (15 minutes) and ends at Q_3 (22 minutes)
- Step 3: Place a vertical line to represent the median (18 minutes)
- Step 4: Extend the horizontal lines from the box out to the minimum value (13 minutes) and the maximum value (30 minutes). These horizontal lines are sometimes called "whiskers" due to the resemblance with cat's whiskers.

Box Plot for Pizza Delivery



Interpretation of the Box Plot

- The box plot shows that the middle 50 percent of the deliveries take between 15 minutes and 22 minutes. The distance between the ends of the box, 7 minutes, is the inter quartile range i.e. the distance between the first and third quartile. That shows the spread or dispersion of the majority of deliveries.
- The box plot also reveals that the distribution of the delivery times is positively skewed. The guiding principle for such conclusion are
- The dashed line to the right of the box from 22 minutes (Q_3) to the maximum time of 30 minutes is longer than the dashed line from the left of 15 minutes (Q_1) to the minimum value of 13 minutes.
- The median is not in the middle in the center of the box. The distance from the first quartile to the median is smaller than the distances from the median to the third quartile.

Test Yourself

Construct a box plot for the data given below and hence comment on the skewness of the distribution:

99	75	84	33	45	66	97	69	55	61
72	91	74	93	54	76	62	91	77	68

Hints:

- Calculate Median,
- Then the median of the 1st half,
- Then the median of the 2nd half.
- Then proceed according to the instruction

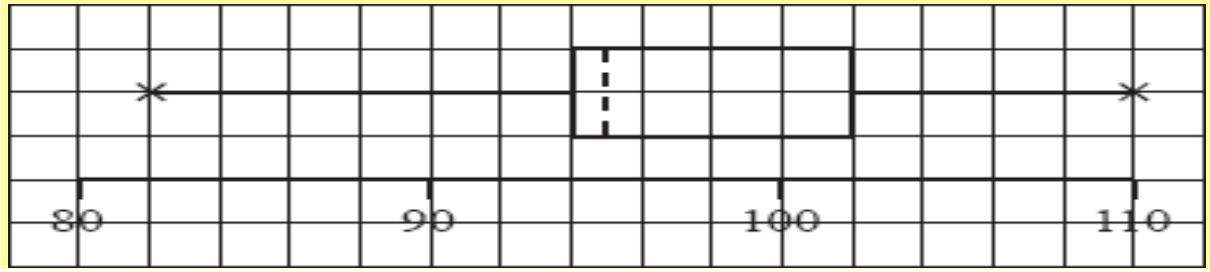
Test of MCQ-

Example 1

Given the information below, draw a box and whisker plot.

Minimum	82
Lower quartile	94
Median	95
Upper quartile	102
Maximum	110

The box and whisker plot is shown below.

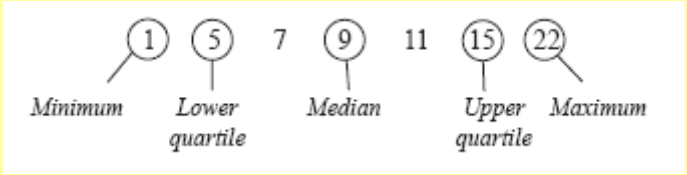


Example 2

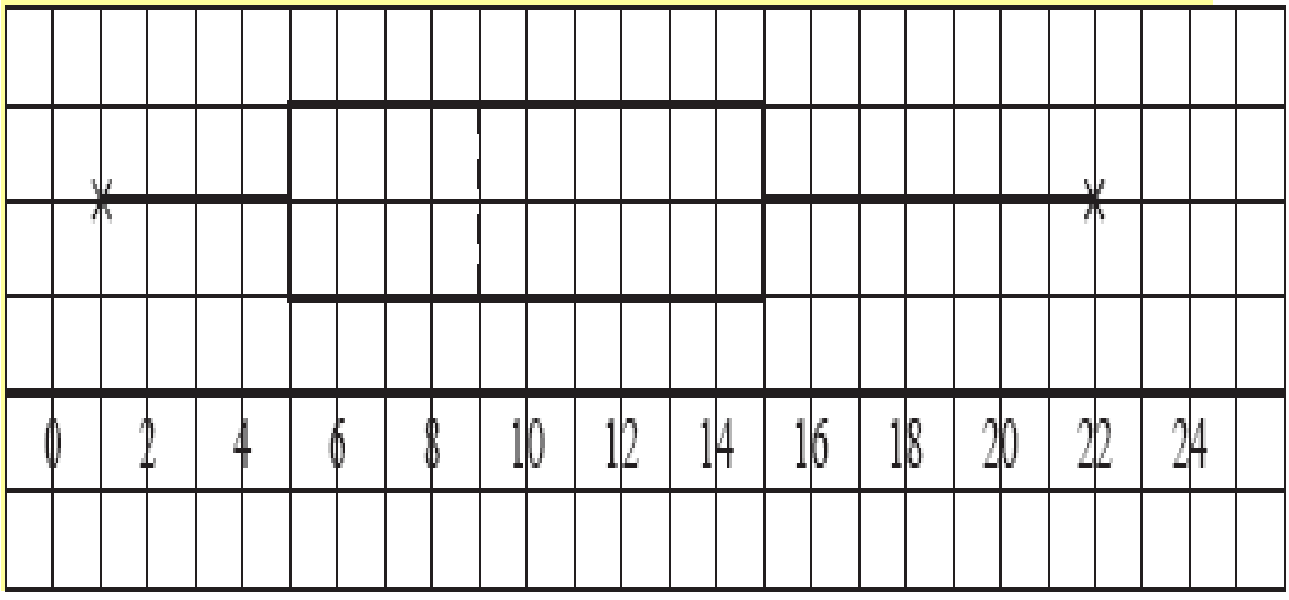
Draw a box and whisker plot for this sample:

5 7 1 9 11 22 15

First list the sample in order, to determine the median and the quartiles.

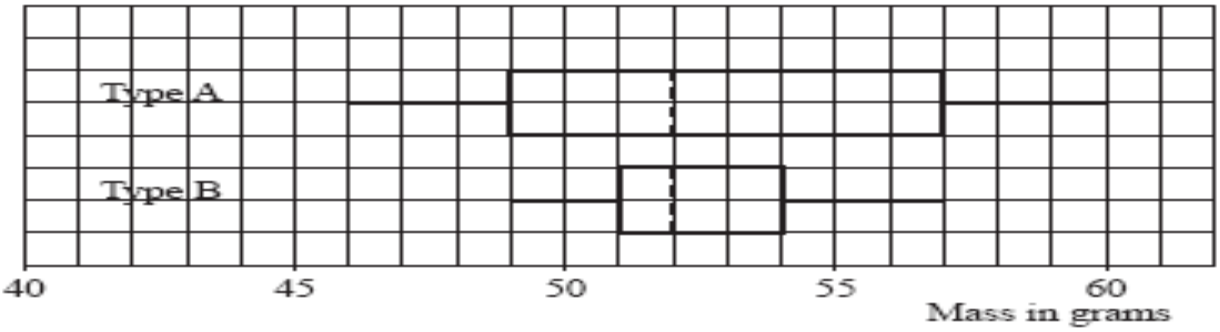


The box and whisker plot is shown below:



Example 3

A gardener collected data on two types of tomato. The box and whisker plot below shows data for the masses in grams of the tomatoes in the two samples.
Compare and contrast the two types and advise the gardener which type of tomato he should grow in future.



	Type A	*Type B*
Median	52 grams	52 grams
Lower Quartile	49 grams	51 grams
Upper Quartile	57 grams	54 grams
Range	14 grams	8 grams
Interquartile Range	8 grams	3 grams

From this table we can see that both types of tomato have the same average mass because their medians are the same.

Comparing the medians and interquartile ranges shows that there is far more variation in the masses of the type A tomatoes, which means that the masses of type B are more consistent than those of type A.

However, comparing the two box and whisker plots, and the upper quartiles, shows that type A tomatoes will generally have a larger mass than those of type B.

Nevertheless, there will be some type A tomatoes that are lighter than any of type B.

Taking all this together, the gardener would be best advised to plant type A tomatoes in future as he is likely to get a better yield from them than from type B.

Exercises

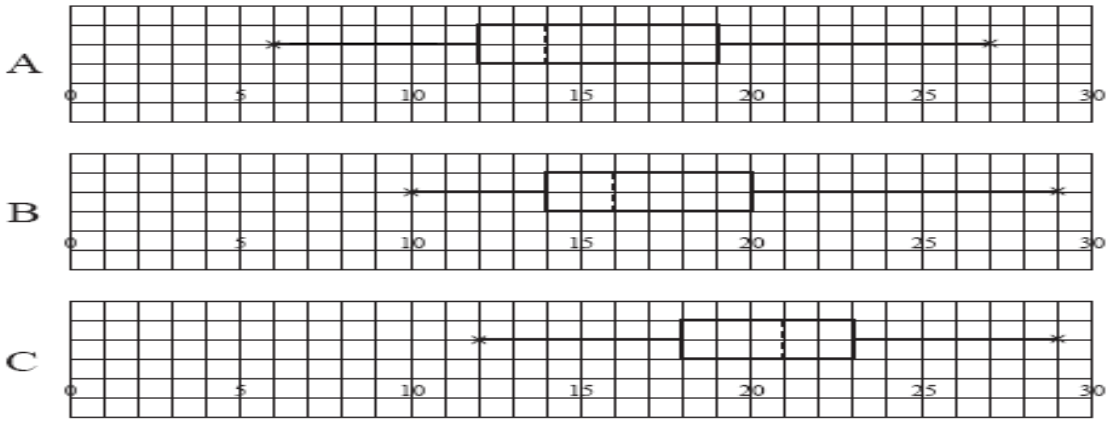
Question 1

Choose one of the box and whisker plots for a sample that has:

Minimum 10

Lower quartile 14

Median	16
Upper quartile	20
Maximum	29

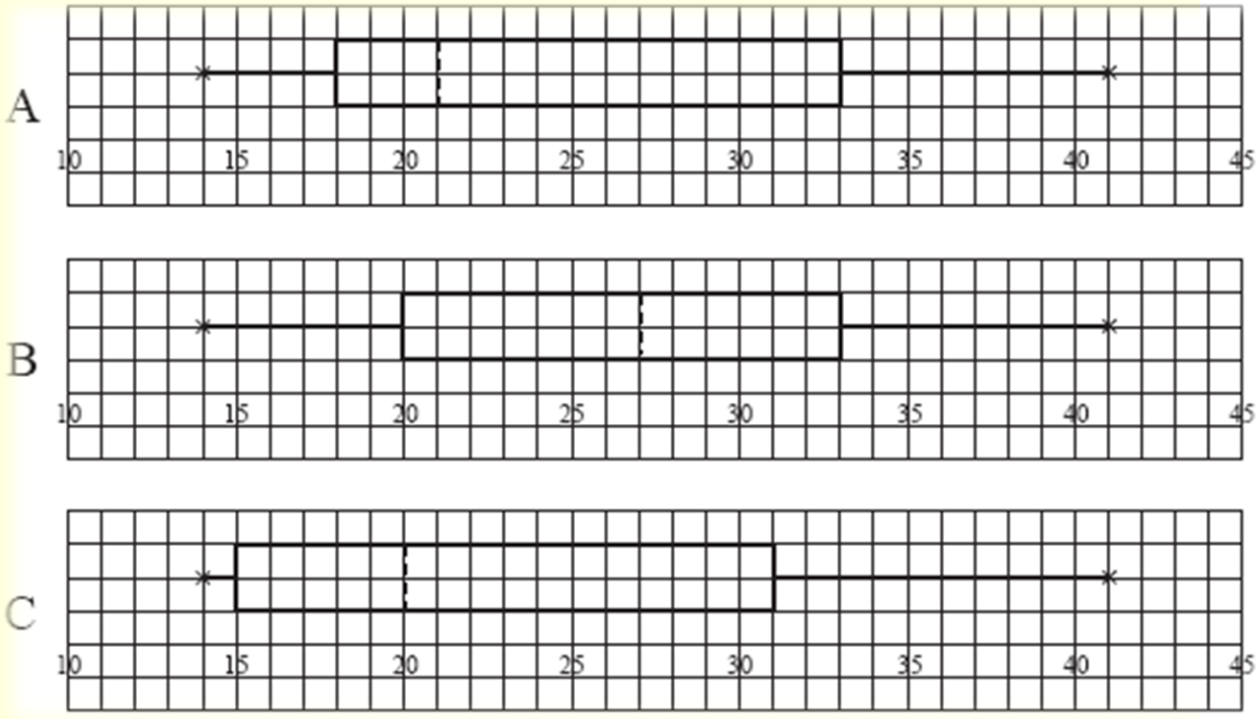


B

Question 2

Choose one of the box and whisker plots for the following sample:

17 22 18 33 14 36 39 41 25 31 18 19 16 21 21



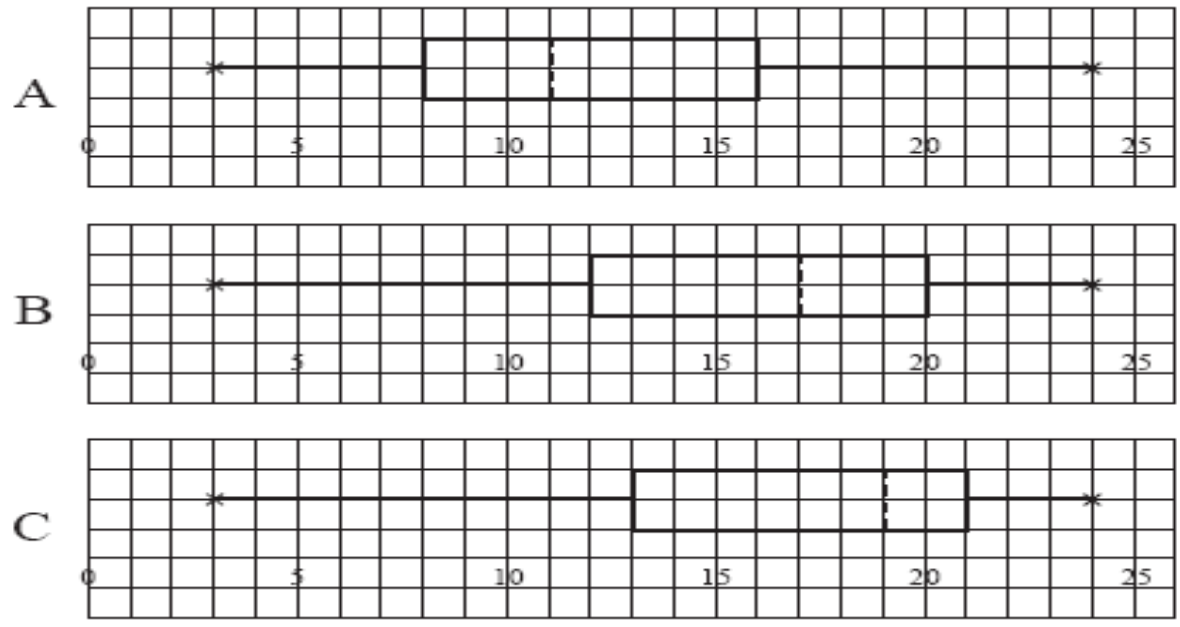
A

Question 3

A sample has:

Minimum	3
Range	21
Semi-interquartile range	4
Median	17
Upper quartile	20

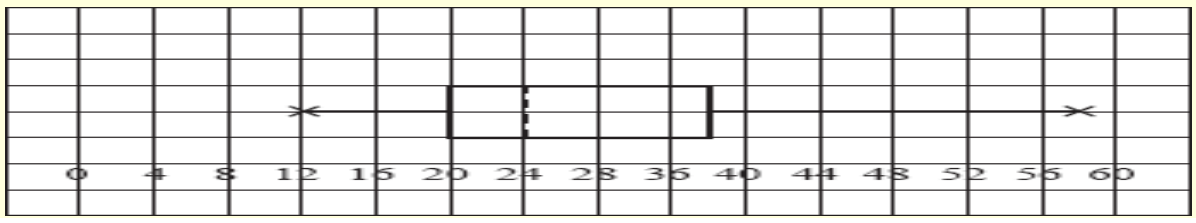
Choose a box and whisker plot for the sample.



B ✓ Correct

Question 4

For the sample illustrated in the following box and whisker plot, determine:



(a)

the range,

46 Correct

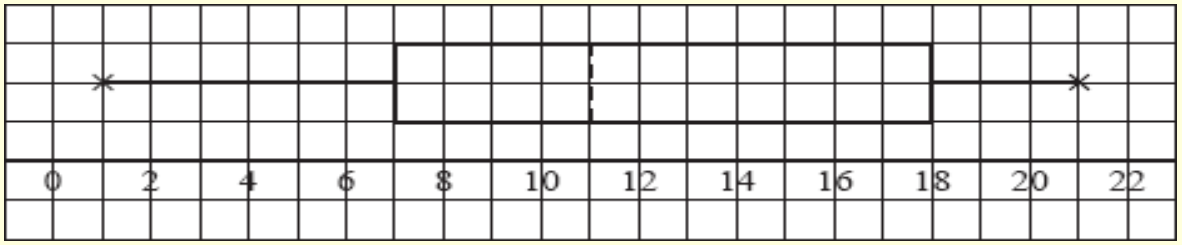
(b)

the semi-interquartile range.

9 Correct

Question 5

What are the median and the semi-interquartile range of the following sample:

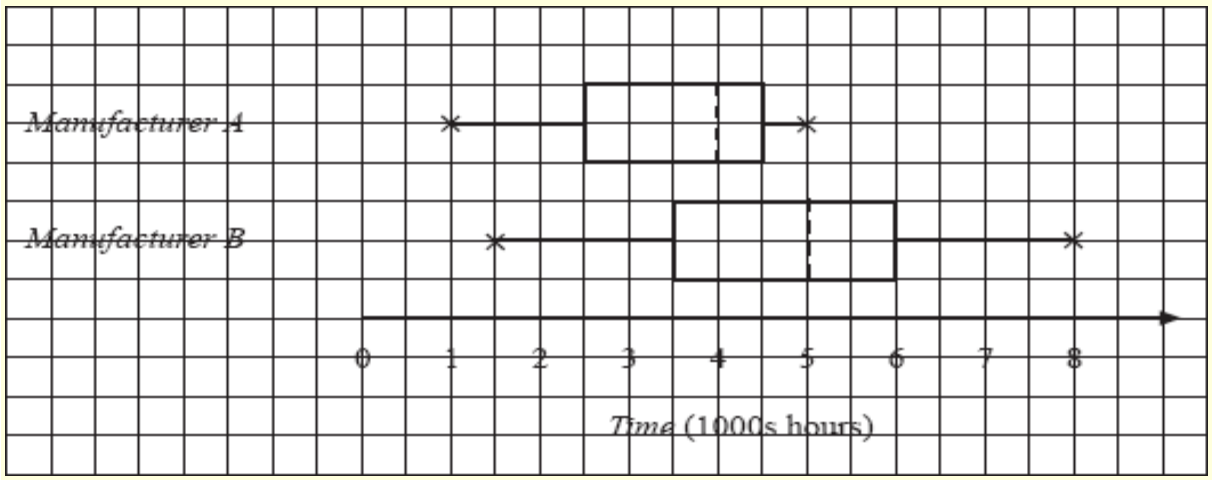


Median = 11

Semi-interquartile range = 5.5 Correct

Question 6

The two box and whisker plots show the data collected by the manufacturers on the life-span of light bulbs.



From this data, which manufacturer produces the long light bulb?

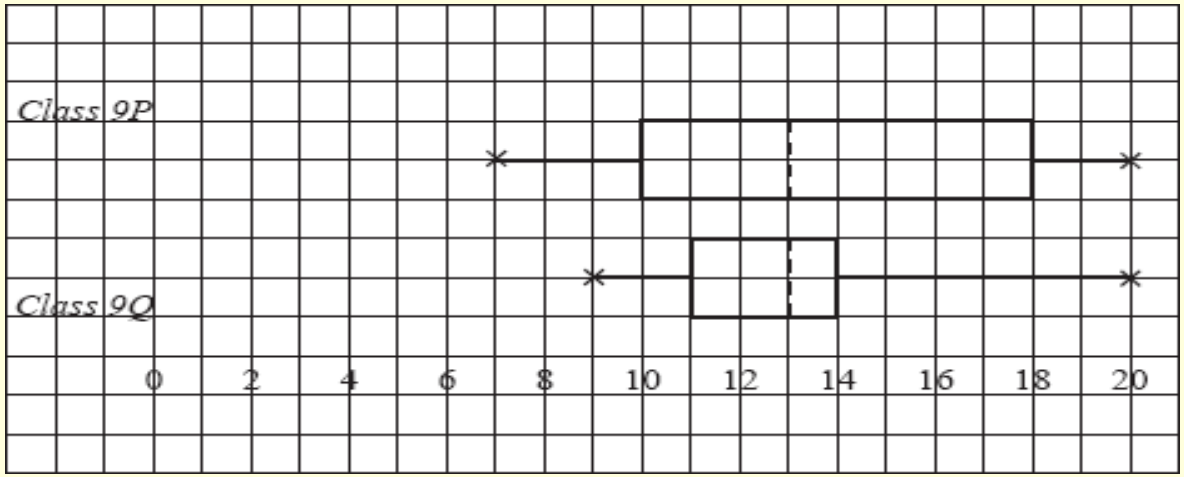
Manufacturer B Correct

	Manufacturer A	*Manufacturer B*
Median	4000 hours	5000 hours
Lower Quartile	2500 hours	3500 hours
Upper Quartile	4500 hours	6000 hours
Range	4000 hours	6500 hours
Interquartile Range	2000 hours	2500 hours

From this table we can see that, on average, the light bulbs produced by manufacturer B have the longer life-span. Although there is far more variation in the life-spans of type B, the lower quartile for type B almost the same as the median for type A, which gives further support to manufacturer B producing the better light bulb.

Question 7

A maths test is given to two classes. The results are illustrated below.
Compare the results.



	Class 9P	*Class 9Q*
Median	13	13
Lower Quartile	10	11
Upper Quartile	18	14

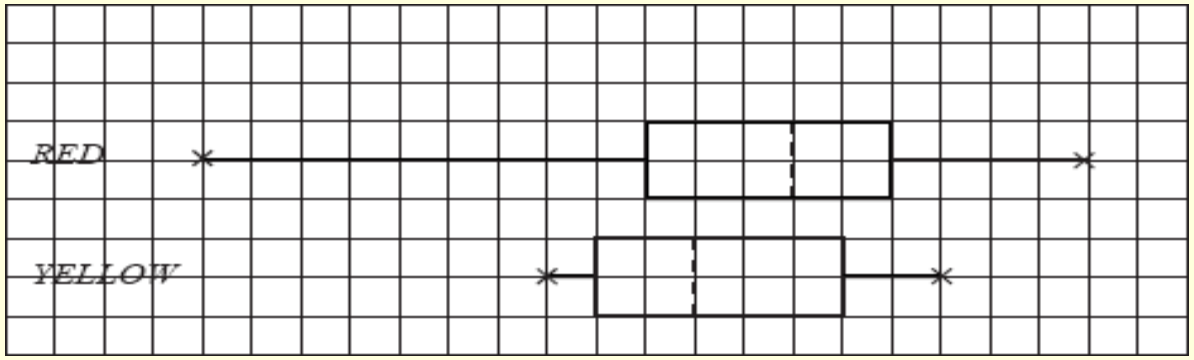
Range	13	11
Interquartile Range	8	3

In general, Class 9P tended to score better marks. ✔ Correct

From the table we can see that both classes had the same average mark. There is far more variation in the scores for class 9P. Some students in 9Q performed better than their counterparts in 9P. In general, though, class 9P tended to score better marks than class 9Q.

Question 8

A builder can choose between two different types of brick that are coloured *red* or *yellow*. The box and whisker plots below illustrate the results of tests on the strength of the bricks.



From the data illustrated in the box and whisker plots:

(a)

give one reason why the builder might prefer to use *red* bricks. Complete the sentence.
The builder might prefer to use the red bricks because that type has the higher median , i.e. they tend, on average, to be stronger than the yellow bricks. ✔ Correct

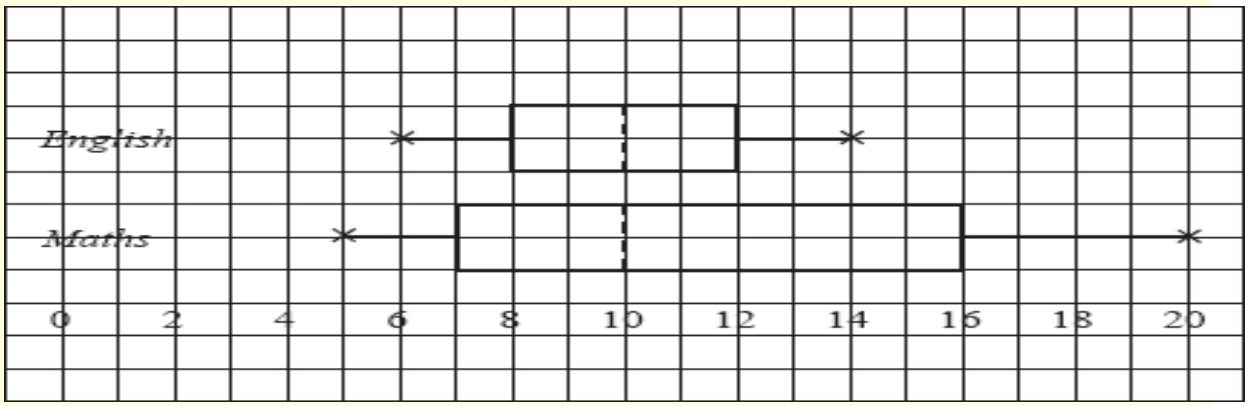
(b)

give one reason why the builder might prefer to use *yellow* bricks. Complete the sentence.

The builder might prefer to use the yellow bricks because that type has the smaller range , and the larger minimum , which means that he is less likely to get a weak brick. ✔ Correct

Question 9

A class took an English test and a Maths test. Both tests had a maximum possible mark of 25. The results are illustrated below.



Compare the results.

	English	*Mathematics*
Median	10	10
Lower Quartile	8	7
Upper Quartile	12	16
Range	8	15
Interquartile Range	4	9

In general, the class tended to score better marks on the Mathematics test. ✔ Correct

From this table we can see that the class had the same average mark in both subjects. There is far more variation in the scores for Mathematics. Few pupils in the class scored high marks in the English test. Nobody scored full marks in either test. In general, though, the class tended to score better marks on the Mathematics test.