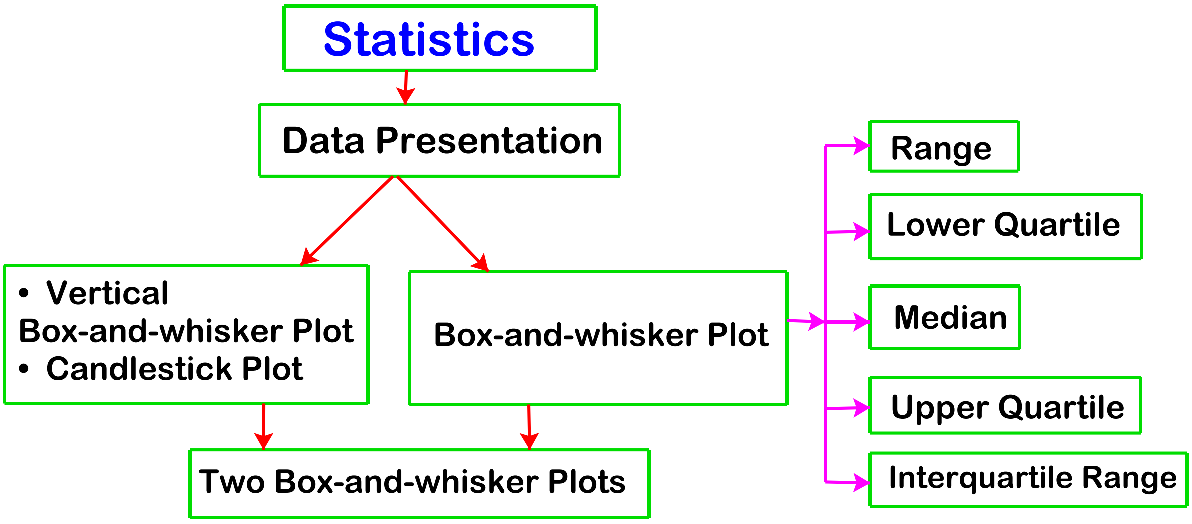
Box-and-Whisker Plot



The descriptive measures of another defining characteristic of a data set: how dispersed or spread out the data are. This measures are measures of dispersion, or measures of variation, or measures of variability. The most frequently used measures of dispersions are:

- Range
- Percentile,
- Quartiles
- Box-and-Whisker Plot

1 Range

The sample range is obtained by computing the difference between the largest value and smallest values in the data set.

Range = Maximum value – Minimum value

The sample range of the variable is quite easy to compute. However, in using the range, many data values are ignored. This is because only the largest and smallest values of the variable are considered. The other observed values are disregarded.

2 Percentiles

Percentile

A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.
Percentile rank refers to the percentage of scores that are equal to or less than a given score.

Assume that the set of data values are rank ordered from smallest to the largest. The values that divide a rank-ordered set of data into 100 equal parts are called **percentiles.**

For example, the 25th percentiles is the value below which 25% of the observations may be found.

The Median is the 50th percentile.

The Pth percentile of a group of numbers is that value below which lie P% of the numbers in group.

The position of the P_{th} percentile $= \frac{(n+1)P}{100}$

Example 1

A department store collects data on sales made by each of its salespeople. The number of sales of 20 salespeople are as follows:

9, 6, 12, 10, 13, 15, 16,14 14, 16, 17, 16, 24,
21, 22, 18, 19, 18, 20 and 17.
Find the 50^{th} percentile of this data set.

Solution

Find the 50 percentile (P50):

Arranging the data in ascending order, we have

Position	Data
1	6
2	9
3	10
4	12
5	13
6	14
7	14
8	15
9	16
10	16
11	16
12	17
13	17
14	18
15	18
16	19
17	20
18	21
19	22
20	24

The position P_{50} $= \frac{(n+1)P}{100}$

$= \frac{(20+1)50}{100}$

$= 10.5$

The data point at position 10.5 $= \frac{16+16}{2} = 16$

3 Quartiles

Quartiles

The quartiles of a ranked data set in descriptive statistics refer to the **three** points, which divide the set of data values into **four equal groups**. Each group consists of a quarter of the data. Quartiles are the percentage points that break down the data set into quarters. The three quartiles are described as follows:

1. **Lower Quartile** (**Q_1**):
The first quartile is the 25th percentile. The lower quartile is the middle number between the median and the smallest number of a data set.

2. **Median (Middle) Quartile** (**Q_2**):
The second quartile is the 50th percentile. It is also known as the median of data, it is the middle observation.

3. **Upper Quartile (Q$_3$):**
The third quartile is the 75th percentile. It is that point below which lie 75 % of the data. It is the middle value between the median and the highest value of a data set.

25% of the data are $\leq Q_1$

50% of the data are $\leq Q_2$

75% of the data are $\leq Q_3$

The relationship between quartiles and percentiles is

Q1 (the 1st quartile) = 25th percentile
Q2 (the 2nd quartile) = 50th percentile (***Median***)
Q3 (the 3rd quartile) = 75th percentile
Q4 (the 4th quartile) = 100th percentile

Interquartile Range

The Interquartile Range is used to indicate the variability around the Median (which shows central tendency). It is the difference between the 1st and 3rd quartile and covers the scores falling within the middle 2 quartiles (center 50%) of the distribution.

Interquartile Range = $Q_3 - Q_1$

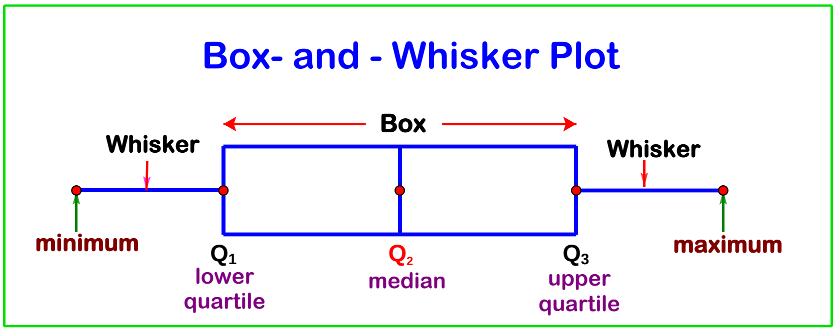
4 Box- and - Whisker Plot

A box-and - whisker plot is a graph that summarizes a set of data by displaying it along a number line. It consists of three parts: *a box* and *two whiskers*.

John Turkey created the box-and -whisker plot to represent the spread of data and its variation. Box-and whisker plot is useful when large numbers of observations are involved and when two or more data set are being compared. Box-and -whisker plot needed five special values of data to present the visual representation. The special value of the data are as follows:

1) The minimum value,
2) The maximum value,
3) The median,
4) The lower quartile and
5) The upper quartile.

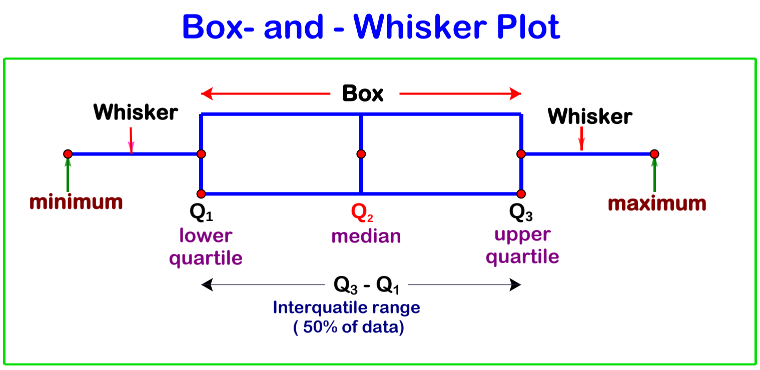
The box-and-whisker plot is as shown below:



- The left whisker extends from the minimum to the first quartile. If represents about 25% of the data.
- The box extends from the first quartile to the third quartile and has a vertical line through the median. The length of the box represents the interquartile range. It contains about 50% of the data.
- The right whisker extends from the third quartile to the maximum. If represents about 25% of the data.

Key element of Box-and-Whisker Plot

A Box-and-Whisker Plot is constructed using the minimum, maximum, quartiles (Q1 and Q3) and the median of a data set (Q2).



- The horizontal axis covers all possible data values.
- The box part of a Box-and-Whisker Plot covers the middle 50% of the values in the data set. This 50% of the data lies between Q1 and Q3, and its range is the interquartile range:

  Interquartile range = Q3 – Q1
- The lower whiskers covers all the values from the minimum value up to Q1 = 25% of the data values
- The upper whiskers covers all the values between Q3 and the maximum value = 25% of the data values
- The median sits within the box and represents the center of the data. 50% of the data values lie above the median and 50% lie below the median.

Outliers: the outliers are data outside the whiskers. Data points have to go above or below the box pretty far to count as outliers. How far? It is 1.5 times the size of the box.

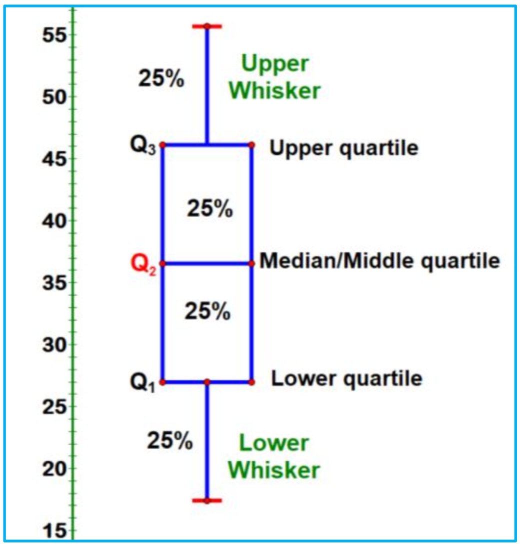
Quartiles:

Quartiles are values that divide a data set into **4 parts.**

- The *median* or *second quartile* (Q_2)
- The *first quartile* (Q_1) is the middle value of the lower half of the ordered set of numbers.
- The *third quartile* (Q_3) is the middle value of the upper half of the ordered set of numbers.
- The *interquartile* range is the difference between the third and the first quartile.

Vertical Box-and-Whisker Plot

The box-and-whisker plot can also be drawn vertically as shown below.



Example 2

What are the minimum, first quartile, median, third quartile, interquartile range and maximum of the data below?

11, 12, 7, 3, 4, 9,

14, 13, 25, 23, 21

Solution

1: Arrange the data in order from least to greatest.

3,	4,	7,	9,	11,	12,	13,	14,	21,	23,	25
minimum		**Q_1**			**median (Q_2)**			**Q_3**		**maximum**

2: Minimum value = 3

Q_1 = 7

Median (Q_2) = 12

Q_3 = 21

Maximum value = 25

Interquartile range = $Q_3 - Q_1$

= 21 – 7

= 14

Example 3

The scores of 15 students in a mathematics test are as follows:

37 60 24 55 81 67 62 82

95 32 53 80 77 91 88

Solution

Step 1: Draw a box-and-whisker plot for the data above.
Rearrange the data in ascending order

← lower half → ← upper half →

24 32 37 53 55 60 62 67 77 80 81 82 88 91 95

lower quartile (53) median (67) upper quartile (82)

Step 2: Find the **median**
The median is the value exactly in the middle of an order set of numbers.
∴ Median **Q_2** = 67

Step 3: Find the lower quartile, **Q_1**
The lower quartile is the middle value of the lower half of the ordered set of numbers.
∴ Lower quartile, Q_1 = 53

Step 4: Find the upper quartile, **Q_3**

The upper quartile is the middle value of the upper half of the ordered set of numbers.

∴ Upper quartile, Q_3 = 82

Step 5: Find the minimum (or lower extreme) and maximum values (or the upper extreme)

The minimum value = 24

The maximum value = 95

Range = 95 – 24

= 71

Step 6: Interquartile range = $Q_3 - Q_1$

= 82 – 53

= 29

Step 7:

- Draw the box-and-whisker plot

  It is important to start a box plot with a **scaled number line**. Otherwise the box plot may not be useful.

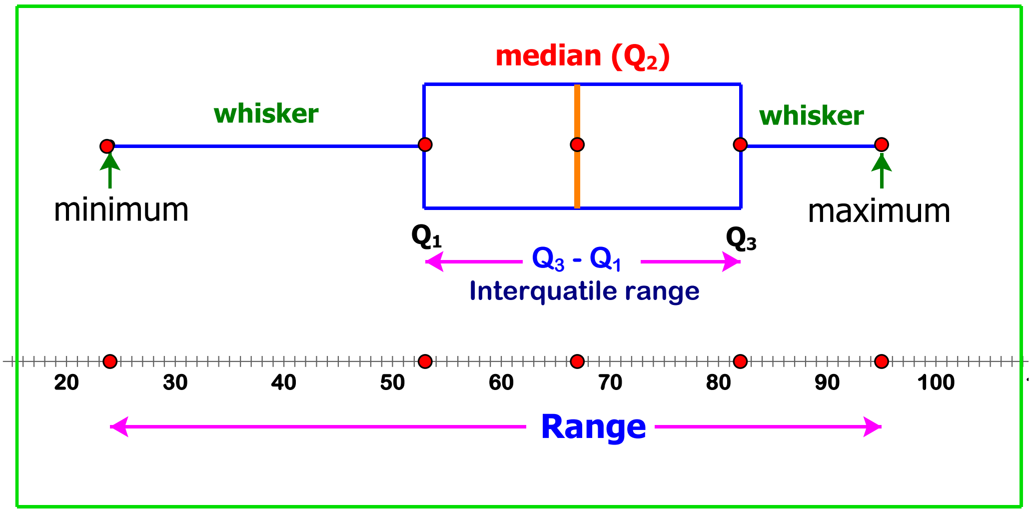
- Draw a horizontal number line to cover the two extremes, the lower and upper quartiles, and the median.

- Draw a rectangular box with the lower and upper quartiles indicated by the respective ends of the box.

- Draw a vertical line in the box to mark the median.

- Add in the whiskers to join the lower extreme to the lower quartile and the upper quartile to upper extreme.

A box-and-whisker plot for the data above is shown below:



Example 4

The length of time (minutes) taken by students to complete one mathematics question is given as follows.

4.4	6.0	5.6	6.1	8.9	8.4	4.8	4.9
4.6	5.5	6.6	8.5	5.3	5.5	5.7	5.8
4.6	6.7	6.5	5.2	8.9	6.4	5.6	6.8
7.2	4.7	7.7	6.0	5.5	9.7	5.6	6.9
5.1	6.1	5.2	6.2	6.3	5.5	6.6	6.6

a) Construct a stem and leaf plot of the set of data

b) Draw a box-and-whisker plot of the set of data including the lower extreme value and the upper extreme value. What we can conclude from this?

Solution

a) A stem and leaf plot of the set of data is shown as follows:

Stem	Leaf
4	4, 6, 6, 7, 8, 9
5	1, 2, 2, 3, 5, 5, 5, 5, 6, 6, 6, 7, 8
6	0, 0, 1, 1, 2, 3, 4, 5, 6, 6, 6, 7, 8, 9
7	2, 7
8	4, 5, 9, 9
9	7

b) From the stem and leaf plot we have:

1) Minimum value or Lower extreme value = 4.4 min.
2) Maximum value or Upper extreme value = 9.7 min.

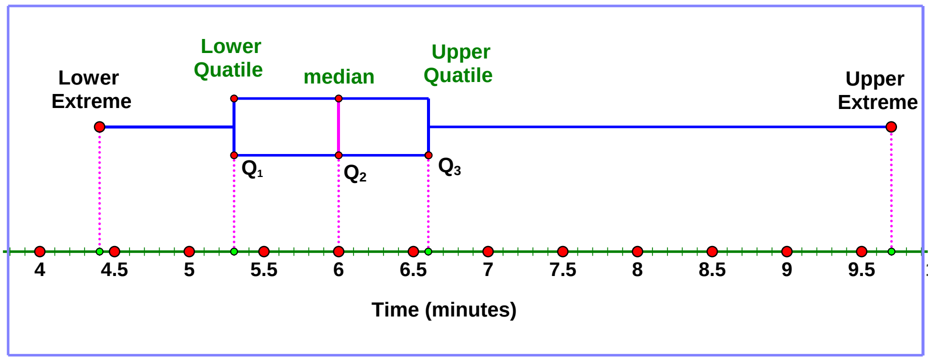
Range = Upper extreme – Lower extreme

= 9.7 – 4.4 = 5.3 min.

3) Median value (Q_2) = 6.0 min
4) The Lower quartile (Q_1) = 5.3 min.
5) The Upper quartile (Q_3) = 6.6 min.

Interquartile range = $Q_3 - Q_1 = 6.6 - 5.3$ = 1.3 min.

Draw a box-and-whisker:



From the Box-and -whisker plot, we can concluded that:

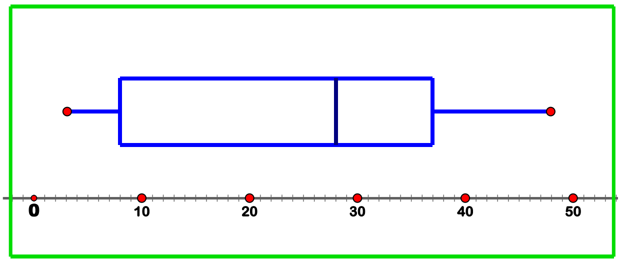
1) The median is near the center of the box. The whisker to the right of the box is much longer than the whisker to the left of the box.
   Thus, we can conclude that although most of the student's time spent close estimates for the time spent. There were a few students who spent unusually large values for the time spent in solving one mathematics question.

2) Although the range of the time spent is 5.3 min, the interquartile range which represents the middle half of the distribution is only 1.3 min.
   The wide range obtained here is due to the presence of a few unusually large values.

3) This shows us that the range might not always be a good indication of the spread of a distribution. A wide range does not always mean that the values in a distribution are well spread. A wide range could be due to the presence of a few unusually small or large values in a distribution.

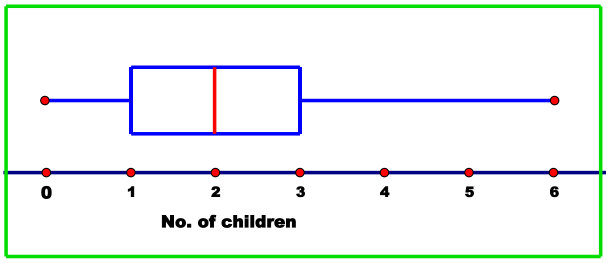
Exercise 1

1. A group of 13 students were asked for their one-way travel times (in minutes) to school. The data are presented in the box-and-whisker plot shown below.



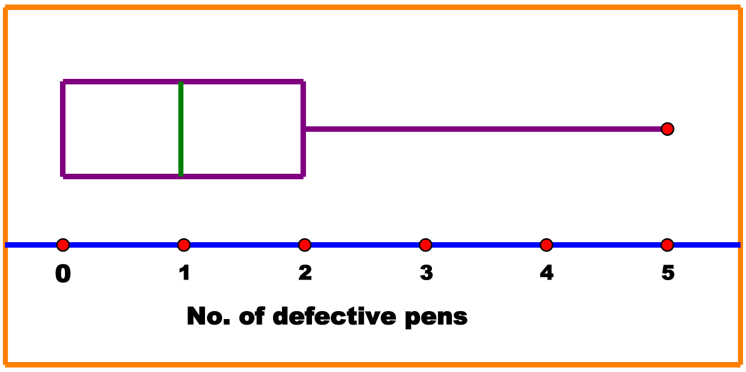
a) State the median.
b) Find the range.
c) Find the interquartile range
d) Notice that the median is closer to the upper quartile. What can you conclude from this?

2. The number of children per family in a block of condominium are represented in the box-and-whisker plot shown below.



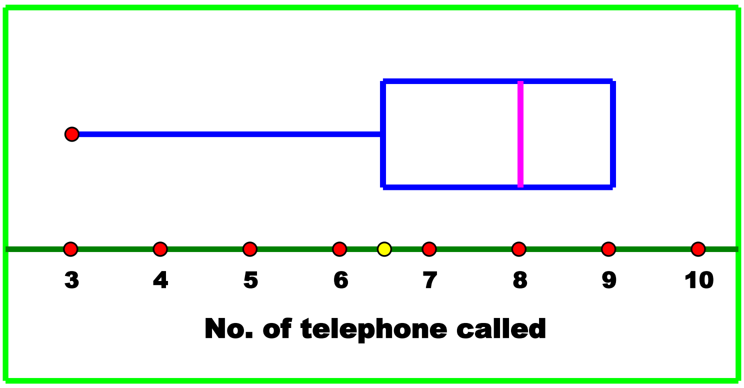
a) State the median number of children per family.
b) Find the range.
c) Find the interquartile range
d) What conclusion can you make from the box-and-whisker plot?

3. In order to find out the number of defective ballpoint pens in each box, the workers carried out the observation on the batch of boxes that he received. The results are summarized in the following box-and-whisker plot.



a) State the median number of defective pens.
b) State a special feature of the box-and-whisker plot.
c) State the lower quartile. What do you notice about the values of the lower quartile and the lower extreme?
d) Find the interquartile range.
e) Find the range.

4. The box-and-whisker plot given below shows the number of telephone called in one hour of ABC Shop.



a) State the median number of defective pens.

b) State a special feature of the box-and-whisker plot.

c) State the upper quartile. What do you notice about the values of the upper quartile and the upper extreme?

d) Find the interquartile range.

e) Find the range.

Why is the interquartile range useful?
Because it represents the middle of the data set, so it is not affected by the minimum, maximum, or any outliers.

Candlestick Chart

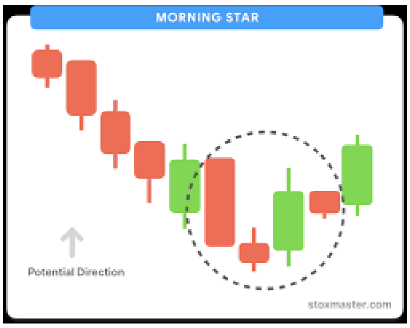
A candlestick chart (also called Japanese candlestick chart) by Munehisa Homma, a Japanese rice trader in the 18th century.

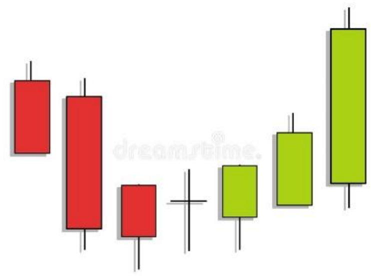
A candlestick chart is a visualization of a financial day's open, high, close, low. The real body is the widest part of the candlestick.

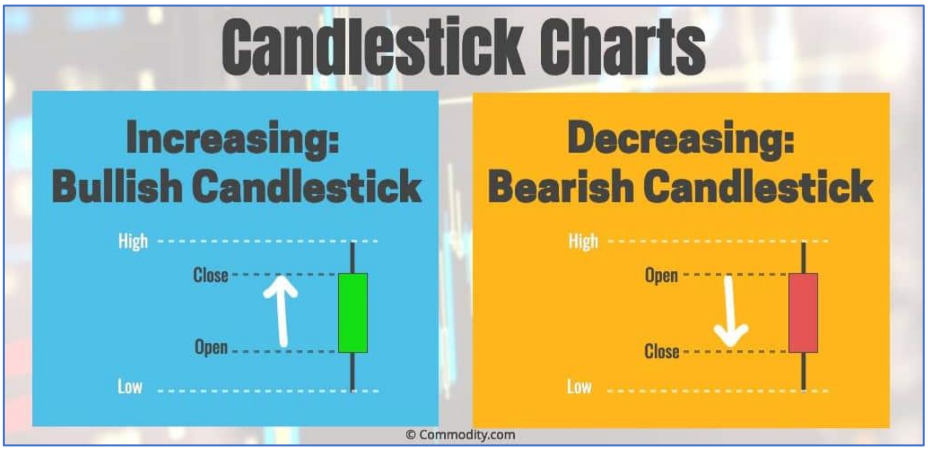
Candlestick charts are a style of financial charts. Candlestick patterns is a style of financial chart, suitable for forex stock market investment trading concept and used to describe price movements of a security, derivative, or currency.

Each candlestick element typically shows one day. A one-month chart may show the 20 trading days as 20 candlesticks elements. Candlestick charts are most often used in the

- analysis of equity,
- stock, foreign exchange, commodity, and option trading.
- currency price patterns and are similar to box plots.







Two Box-and-Whisker Plots

When comparing the distributions of two data sets on the same measurement using box-and-whisker plots, we can compare the ***shape***, ***average***, and ***spread*** of the data sets as follows:

- **Shape:** The shape of a data set refers to whether or not it is symmetric or skewed.

  If a data set is distributed **symmetrically** about the center, the box should be evenly split by the median and the whiskers should approximately equal length.

  If a data set is ***skewed*** then one of the whiskers will be longer than other.

  - If a data set is **right** or **positively skewed**, then the *right* whisker will be longer than the *left* and the boxes on the *right* side will be longer than those on the *left*.
  - If a data set is **left** or **negatively skewed**, then the *left* whisker will be longer than the *right* and the boxes on the *left* side will be longer than those on the *right*.

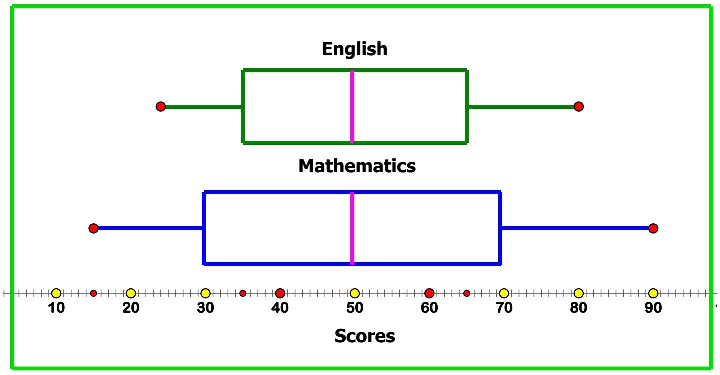
- **Average**: In a box plot, the measure of an average used the **median.** Comparing the medians of two data sets, we can determine in which data set the values are "on average" higher or lower than in the other, or there is no difference on average.

- **Spread**: The spread of the data sets can be compared by looking at the ranges (maximum value and minimum value) and the interquartile ranges (Q3 – Q1)

  - The **range** tells the overall spread of each data set.
  - The **interquartile range** tells the *spread* of the middle *50% of the data*, that is how far the middle 50% of the data deviates from the center.

Example 5

A class of students took an English test and Mathematics tests are represented in the box-and-whisker plots given below.



a) State the median score for the English test and mathematics test.

b) Find the interquartile range of the score for the English test and mathematics test.

c) Compare the two box-and-whisker plots. Comment on the spread of the scores for each of the tests.

Solution

a) Median score of English = 50 score

Median score of Mathematics = 50 score

b) The interquartile range of the score for the English test

= 65 – 35 = 30 scores

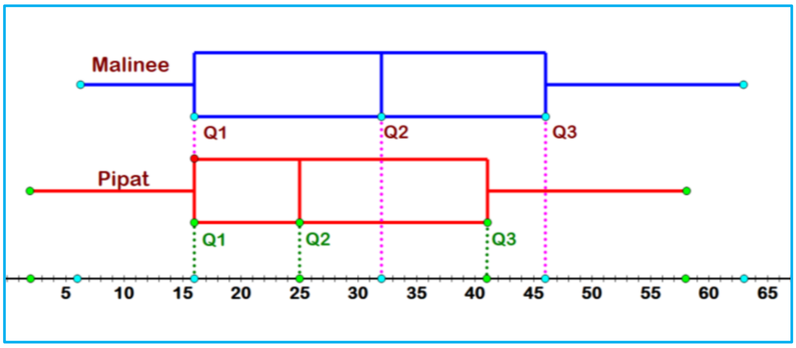
The interquartile range of the score for the mathematics test

= 70 – 30 = 40 scores

c) Although the median scores for both test are the same, the interquartile range of the score for the mathematics test is larger than that for the English test. These observations indicate that the scores for the Mathematics test are more widely spread than those for the English test.

Example 6

The following box plots show the number of cars sold by two sales managers working in the same company, Miss Malinee and Mr. Pipat.



a) Find the interquartile range for each of the box-and-whisker plots.

b) By inferring for the box-and-whisker plots, make comparisons between Miss Malinee's and Mr. Pipat's.

Solution From the given two box-and- whisker plots, we can use the data in the box-and-whisker plots above to compare the results from the two sales managers Miss Malinee and Mr. Pipat as follows.

	Malinee's	Pipat's
Minimum number of cars sold	6	2
Maximum number of cars sold	63	58
Range	63 – 6 = 57	58 – 2 = 56
Lower Quartile (Q1)	16	16
Median number of cars sold (Q2)	32	25
Upper Quartile (Q3)	46	41
Interquartile for the range (Q3 – Q1) number of cars sold	46 – 16 = 30	41 – 16 = 25

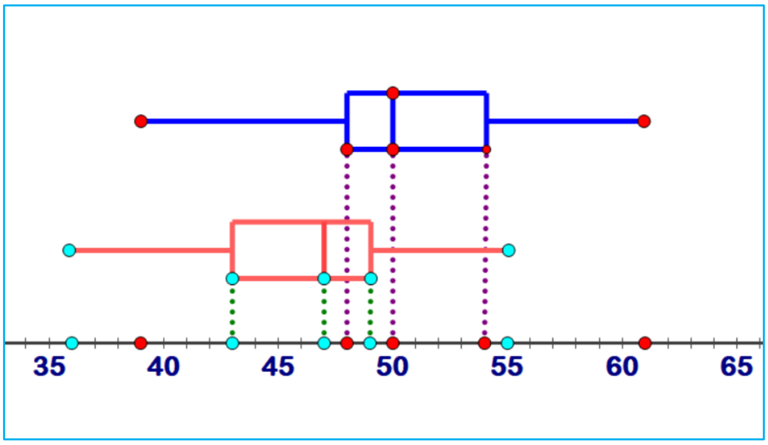
Comparisons:

1) Interquartile for the range number of cars sold by Malinee = 46 – 16 = 30
   Interquartile for the range number of cars sold by Pipat = 41 – 16 = 25
2) The median number of cars sold by Malinee (Q2)= 32
   The median number of cars sold by Pipat (Q2) = 25
   The median number of cars sold by Malinee is higher than that sold by Pipat.
3) Minimum number of cars sold by Malinee are both higher than that sold by Pipat.
4) The upper whisker in Malinee's box-and-whisker plot is longer than the lower whisker.
5) The median number of cars sold by Malinee (Q2) is near the center of the box, we can say that Malinee's sales are **more consistent**.

⇨ All of these suggest that Miss Malinee sold more cars than Mr.Pipat.

Exercise 2

1. The box-and-whisker plots given below show the masses (in kg) of the boys and girls in a school respectively.

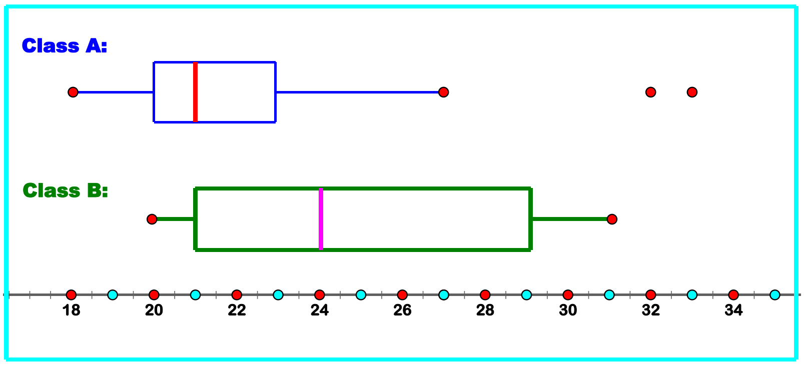


a) Find the median, the lower quartile and upper quartile of the masses of the boys.

b) Find the median, the lower quartile and upper quartile of the masses of the girls.

c) What conclusion can you make by comparing the two box-and-whisker plots?

2. The box-and-whisker plots given below shows the scores of the students in a statistics class ***A*** and in a statistics Class ***B***.



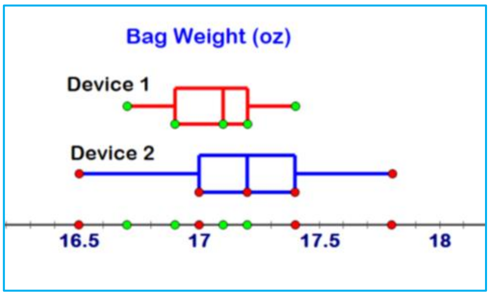
a) Find the median, the lower quartile and upper quartile of the scores of students of Class A and Class B.

b) Which class has the most variability, if you use range?

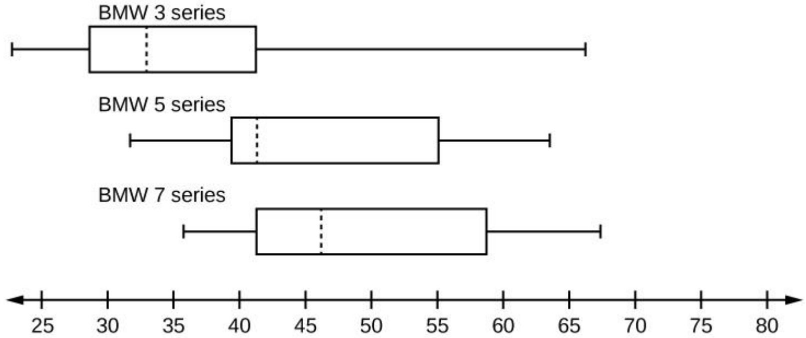
c) Which class has the most variability, if you use Interquartile range to measure variability?

d) What conclusion can you make by comparing the two box-and-whisker plots?

3. A cereal company is choosing between two devices to package their cereal into bags. The box-and-whisker plots below show the weights of the bags packed by each device. Which device produces packages with a more consistent weight? Explain.



4. A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.



a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.

b. Which group is most likely to have an outlier? Explain how you determined that.

c. Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?

d. Look at the BMW 5 series. Which quarter has the smallest spread of data? What is the spread?

e. Look at the BMW 5 series. Which quarter has the largest spread of data? What is the spread?

f. Look at the BMW 5 series. Estimate the interquartile range (IQR).

g. Look at the BMW 5 series. Are there more data in the interval 31 to 38 or in the interval 45 to 55? How do you know this?

h. Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?

   a. 31–35
   b. 38–41
   c. 41–64