The Five-Number summary

A five-number summary consists of minimum observation, lower quartile, median, upper quartile, maximum observation and provides a way of determine the shape of the distribution, that is, to see if there is symmetry or not in data.

If the data are perfectly symmetrical, the relationship among the various measures in the five-number summary will be as expressed below:

- The distance from the smallest value, i.e. $X_{smallest}$ to median is equal to the distance from median to the largest value, i.e., $X_{largest}$.
- The distance from the smallest value i.e. $X_{smallest}$ to first quartile (Q_1) is equal to the distance from upper quartile (Q_3) to the largest value, i.e., $X_{largest}$.

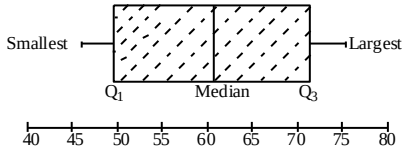
For asymmetrical distributions, the relationship among the various measures in the five-number summary will be as expressed below:

- In right-skewed distributions, the distance from median to the smallest value, i.e. $X_{smallest}$ is smaller than the distance from the largest value, i.e., $X_{largest}$ to median.
- In right-skewed distributions, the distance from upper quartile (Q_3) to the largest value, i.e., $X_{largest}$ is greater than the distance from the smallest value i.e. $X_{smallest}$ to first quartile (Q_1).
- In left-skewed distributions, the distance from the smallest value, i.e. $X_{smallest}$ to median is greater than the distance from the largest value, i.e., $X_{largest}$ to median.
- In right-skewed distributions, the distance from upper quartile (Q_3) to the largest value, i.e., $X_{largest}$ is greater than the distance from the smallest value i.e. $X_{smallest}$ to first quartile (Q_1).

Box-and-whisker plot

A box and whisker plot is a graphical presentation of the data that displays a five number summary of a data set based on the minimum, lower quartile, median, upper quartile, and maximum. The vertical line drawn within the box represents the median. The vertical line at the left side of the box represents the location of Q_1 and the vertical line in the right side of the box represents the location of Q_3. Hence the box contains the middle 50% of the values in the distribution of the given data set. The lower 25% of the data are represented by the line (known as whisker) connecting the left side of the box to the location of the smallest value. Similarly, the

upper 25% of the data are represented by the line (known as whisker) connecting the right side of the box to the location of the largest value as shown in the following box and whisker plot.



Example: The data represent the amount of grams of carbohydrates in a serving of breakfast cereal 1, 15, 23, 29, 19, 22, 21, 20, 15, 25, 17. Construct a box and whisker plot for the carbohydrate amounts.

Solution: For the calculation of Q_1, median, and Q_3 arrange the data in ascending order to their magnitude as follows

11 15 15 17 19 20 21 22 23 25 29

Five number summary (Smallest, Q_1, median, Q_3 and the largest) values can be represented through box - and - whisker plot.

The smallest value is 11 and the largest value is 29.

For first quartile (Q_1),

$$Q_1 = \text{value of } \left(\frac{n+1}{4}\right)^{th} \text{ item} = \text{value of } \left(\frac{11+1}{4}\right) \text{item} = \text{value of } 3^{rd} \text{ item}$$

$\therefore$ $Q_1 = 15$, Where the number of observations (n) = 11

For Median,

$$\text{Median } (M_d) = \text{Value of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = \text{Value of } \left(\frac{n+1}{4}\right)^{th} \text{ item} = \text{Value of } 6^{th} \text{ item}$$

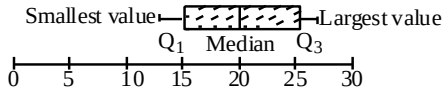
$\therefore$ $M_d = 20$

Similarly, for 3^{rd} quartile (Q_3)

$$Q_3 = \text{value of } \left[\frac{3(n+1)}{4}\right]^{th} \text{ item} = \text{value of } \left[\frac{3(11+1)}{4}\right]^{th} \text{ item} = \text{value of } 9^{th} \text{ item}$$

$\therefore$ $Q_3 = 23$

The five number summary is (11, 15, 20, 23, 29). Now, the box - and - whisker plot for the given data of amount of carbohydrates in a serving of breakfast is as follows:



Example: In Pokhara, saving banks are permitted to sell a form life insurance called Saving Bank Life insurance. The approval process consists of under writing, which includes a review of the application, a medical information bureau check, possible requests for additional medical information and medical exams a policy complication stage where the policy pages are generated and sent to the bank for delivery. The ability to deliver approved policies to customers in a timely manner is critical to the profitability of this service to the bank. During a period of 1 month, a random sample of 27 approved policies was selected and the total processing time in days was recorded with the following results.

73	19	16	64	28	28	31	90	60	56
22	18	45	48	17	17	7	91	63	50
51	51	31	56	69	16	17			

i) Construct a box and whisker plot.

ii) Are the data skewed? If So, how?

Solution: (i) Let arrange the data into ascending order of their magnitude as follows:

16, 16, 17, 17, 18, 19, 22, 28, 28, 31, 31, 45, 48, 50, 51, 56, 60, 63, 64, 69, 74, 90, 91, 92

In order to construct box - and - whisker plot, let us first compute the five number summary (i.e. smallest, Q_1, median, Q_3, largest) as follows:

Here, The smallest value (S) = 16

The largest value (L) = 92

For quartile, Q_1,

$$Q_1 = \text{Value of } \left(\frac{n+1}{4}\right)^{th} \text{ item} = \text{Value of } \left(\frac{27+1}{4}\right)^{th} = \text{Value of 7th item}$$

$\therefore$ $Q_1 = 18$

Median, Md,

$$Md = \text{Value of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = \text{Value of } 14^{th} \text{ item}$$

$Md = 45$

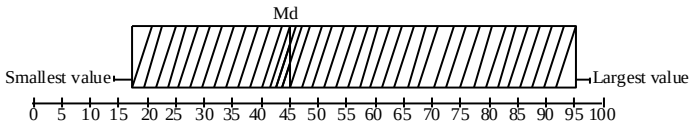
$$Q_3 = \text{Value of } \left[\frac{3(n+1)}{4}\right]^{th} \text{ item} = \text{Value of } \left[\frac{3(27+1)}{4}\right]^{th} \text{ item}$$

$= \text{Value of } 21^{th} \text{ item} = 63$

$\therefore$ $Q_3 = 63$

Hence, the five number summary is given by (16, 18, 45, 63, 92)

This five number summary can be presented in the box – and – whiser put as follows.



(ii) The distance from the smallest value to the median = 45 – 16 = 29

The distance from median to largest value = 92 – 45 = 47

The distance from the smallest value to Q_1 = 18 – 16 = 29

The distance from median to largest value = 92 – 45 = 47

The distance from the smallest value to Q_1 = 18 – 16 = 2

The distance from Q_3 to the largest value = 92 – 63 = 29

Here, the distance from the smallest value to the median is less than distance from the median to the largest value (i.e 29 < 49), Similarly, the distance from the smallest value to Q_1 is also less than the distance from Q_3 to the largest value (i.e 2 < 29). This shows that the data is right skewed.

However if we compute $(M_d - Q_1) > (Q_3 - M_d)$ i.e 27 > 18 which is indicative of left skewed distribution. This contradictory information clearly indicates the given data set is not uniformly distributed.