Statistics: Percentile, Quartile and Box Plot

- Percentile
- Quartile

Percentile:

- Percentile means data divided into 100 parts
- Per cent: cent means century: 100
- 1percentile, 2p,3p....90p
- A percentile is a measure that indicates the value below which a given percentage of observations in a group of observations falls.

- **For Example**, assume that, you have written a CAT exam
- Total numbers of students appear in CAT exam is: 1000
- Total maximum marks of CAT exam: 100
- Ashutosh have written an exam he got 75 marks
- CAT exam given him a percentile: 90 percentile marks

- **Definition**: There are 90% students less than him which means out of 1000 students 900 students have got marks less than him(75) marks.

- Only 10% of students are greater than his marks
- Only 100 members got greater than 75 marks

Another Example,

Gate Exam 3rd year: 28 marks
Percentile: 91 p
Total students appeared: 1,20,000

4th year: 70 marks
Percentile: 82 pe.

- Percentage says out of 100 marks how many you got
- Percentile says how many students got better than your marks
- our percentile is 95 means there are 5% of students better than you
- Total: 1000 students
- Then 50 students are better than you

- Case 1: 75 marks, 60 percentiles
- Case 2: 35 marks, 91 percentiles

➢ Suppose we have a dataset,
$2038 $1758 $1721 $1637 $2097
$2047 $2205 $1787 $2287 $1940
$2311 $2054 $2406 $1471 $1460
Calculate 50 percentile
Calculate 25 percentile
Calculate 75 percentile

<u>50 Percentile</u>: means only 50 percentage values greater than that value

$$15 * \frac{50}{100} = 7.5$$

After 7.5 number 8 will come
So, the 8th position is = 2038

If we write the dataset in a straight line,
$1460 - !471- 1637- 1721- 1758 – 1787 – 1940 – 2038 – 2047 – 2054 – 2097 – 2205 – 2287 – 2311 – 2406

<u>25 Percentile</u>: means only 25 percentage values greater than that value

$$15 * \frac{25}{100} = 3.75$$

After 3.75 number 4 will come
So, the 4th position is = 1721

$1460 - !471- 1637- 1721- 1758 – 1787 – 1940 – 2038 – 2047 – 2054 – 2097 – 2205 – 2287 – 2311 – 2406

75 Percentile: means only 75 percentage values greater than that value

$$15 * \frac{75}{100} = 11.25$$

After 11.25 number 12 will come
So, the 12th position is = 2205

$1460 - !471- 1637- 1721- 1758 – 1787 – 1940 – 2038 – 2047 – 2054 – 2097 – 2205 – 2287 – 2311 – 2406

So, the final result will be,
$1460 - !471- 1637- 1721- 1758 – 1787 – 1940 – 2038 – 2047 – 2054 – 2097 – 2205 – 2287 – 2311 – 2406

$$50_p = 15 * \frac{50}{100} = 7.5$$

$$25_p = 15 * \frac{25}{100} = 3.75$$

$$75_p = 15 * \frac{75}{100} = 11.25$$

$$L_p = N * \frac{l_p}{100}$$

$$L_p = (N + 1) * \frac{l_p}{100}$$

$$50_p = (15 + 1) * \frac{50}{100} = 8$$

Quartile:

- Quartile means 25
- Quartiles are specific types of percentiles that divide the data into four equal parts. There are four quartiles:
    - **First quartile (Q1)**
    - **Second quartile (Q2)**
    - **Third quartile (Q3)**
    - **Fourth Quartile(Q4)**

- **Quartiles** are used to understand the spread and center of the data.

- Suppose, 100 is divided by using 25cso how many parts will come

0 – 25

25 – 50

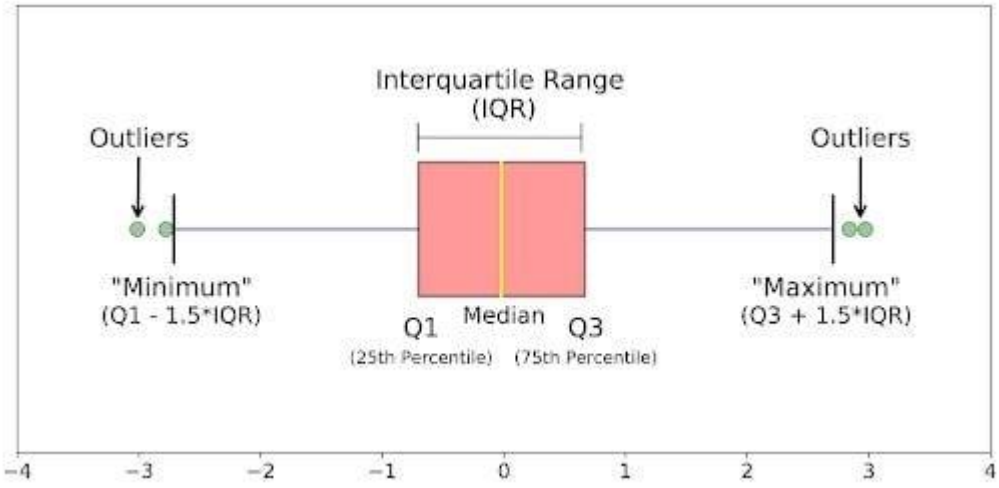
50 – 75

75 – 100

- But we know that asymptotes never touch real line
- In Statistics we can't say zero existence or 100 existences without data
- Instead of zero: we will consider as minimum point
- Instead of 100: we will consider as maximum point
    - Quartile - 1: $Q_1 = \min point\ to\ 25_p$
    - Quartile – 2: $Q_2 = 25_{\mathrm{p}}\ \ to\ 50_p$
    - Quartile – 3; $Q_3 = 50_{\mathrm{p}}\ \ to\ 75_p$
    - Quartile – 4: $Q_4 = 75_p\ to \max point$

Box Plot:

- A box plot is a standardized way of displaying the distribution of data based on a five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.
- It provides a graphical summary of the data's central tendency, variability, and shape.



- In the above diagram, Q1 = 25 p value
  Q2 = 50 p value
  Q3 = 75 p value

- Here outliers will exist after Q3 point and below Q1 point
  - Upper bound = Q3 +?
  - Lower bound = Q3 -?
- In order to find outliers, we need to travel from **Q3 to above and Q1 to below**
- The travel distance based on middle 50 % of data
- That middle 50% of data is called as **IQR: - Inter Quartile Range**
- **IQR = Q3 – Q1**

- So, the updated values are,
  - Upper bound = Q3 + IQR
  - Lower bound = Q3 – IQR

- The upper bound and lower bound cut-off varies based on How many times of IQR we are using
  - Upper bound = Q3 + k*IQR
  - Lower bound = Q3 - k*IQR

- Generally, we will use k = 1.5 and k = 3,
- When k = 1.5: - Mild outlier
  - Upper bound = Q3 + 1.5*IQR
  - Lower bound = Q3 – 1.5*IQR

- When k = 3: - High outlier
  - Upper bound = Q3 + 3*IQR
  - Lower bound = Q3 – 3*IQR

- In Python we use by default k =1.5 value only
- Middle line is called median = 50p of data

- Skewness of a Box Plot: