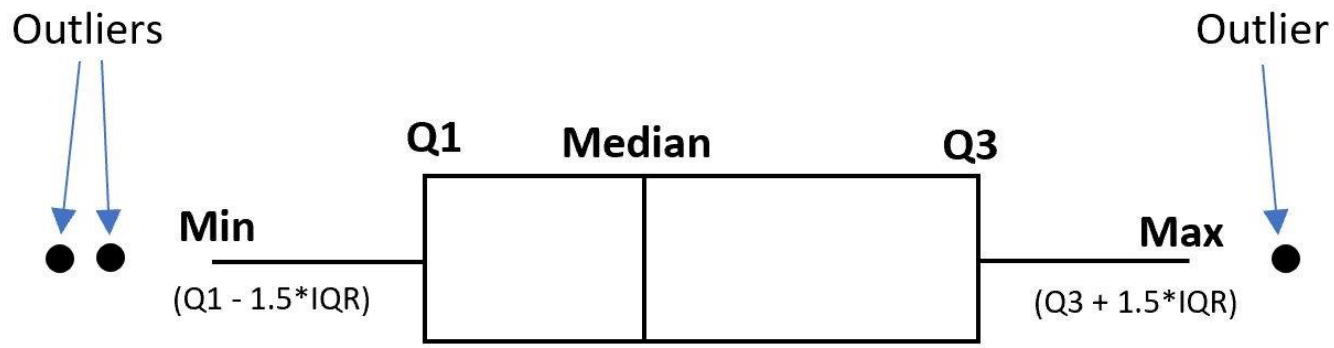
21. What is the benefit of using box plots?

Ans.

Box plot shows 5 number summary of a dataset. So we can easily identify the outliers. Boxplot also allows you to compare multiple datasets in detailed way.

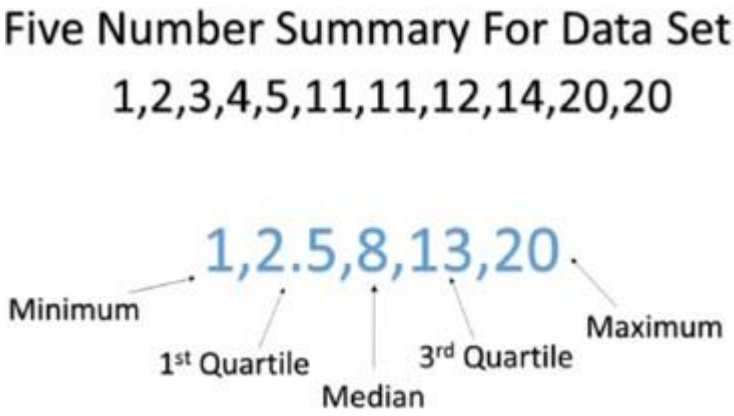


22. What is the meaning of the five-number summary in Statistics?

Ans.

5 number summary is a descriptive analysis of a large dataset where we need to find-out the

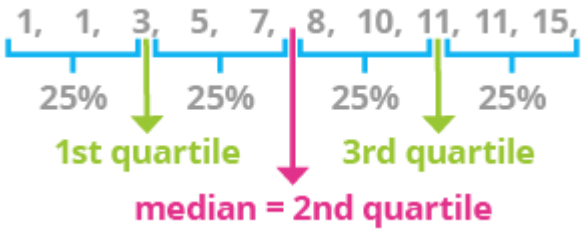
- minimum(0th percentile),
- 1st quartile(25 percentile),
- Median (2nd quartile),
- 3rd quartile(75 percentile) and
- maximum (100th percentile) point of a dataset.



23. What is the difference between the First quartile, the Second quartile, and the third quartile?

Ans.

If the data points are arranged in ascending order then first quartile means the data points under 25% (index). Second quartile means the data points under 50%(index) and third quartile means the data points under 75%(index).



24. What is the difference between percentage and percentile?

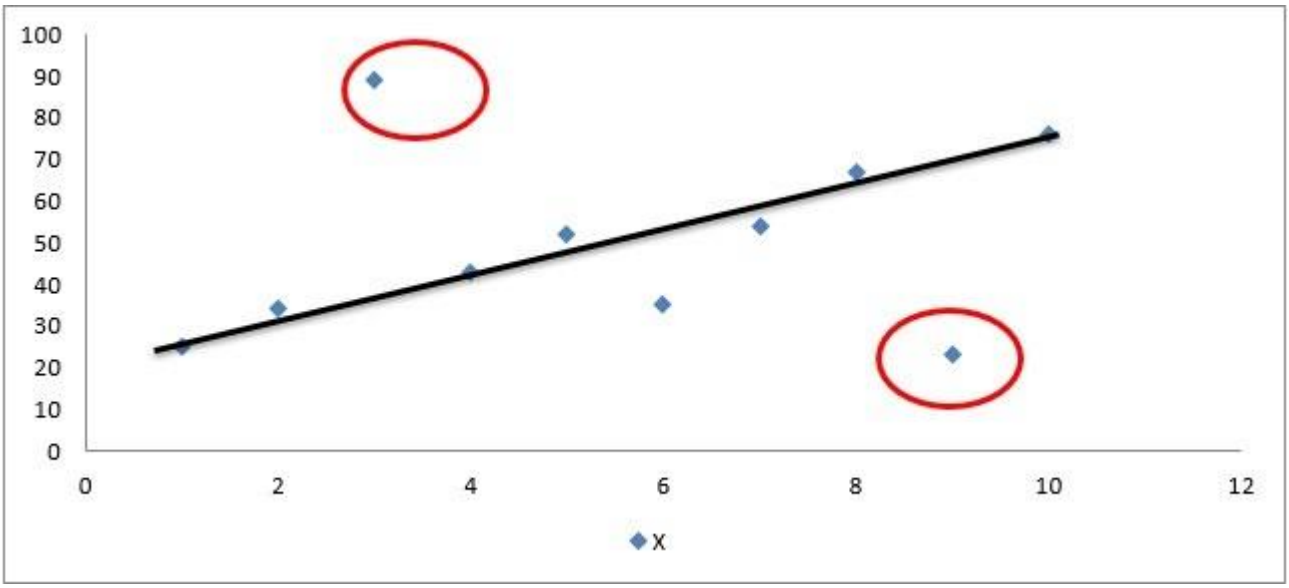
Ans.

- Percentage is a mathematical quantity which can be written out of total 100.
- Percentile is percentage of value present under a specific value.

25. What is an Outlier?

Ans.

Outlier is a data point which differs from rest of the observation. In other words the dataset which don't follow the pattern. It may occur due to variability in the dataset or due to experimental or human error.



26. What is the impact of outliers in a dataset?

Ans.

Outlier can change the result of the data analysis or model made on them.

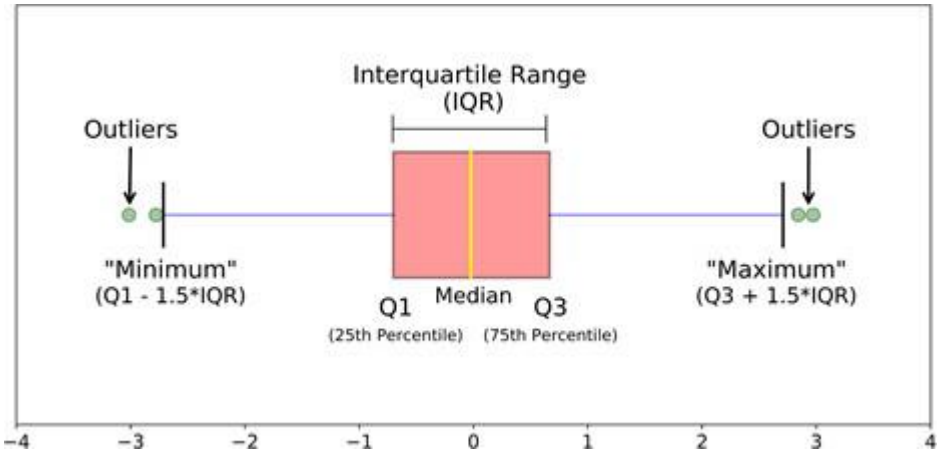
1. Outliers can decrease the normality of the distribution.
2. It affect the mean of the dataset.
3. It increases the error variance and decrease the power of statistical tests.
4. They can cause bias estimates.
5. They can also impact basic assumption of regression as well as other statistical model.

27. Mention methods to screen for outliers in a dataset.

Ans.

There are four ways to identify outliers

1. Sorting method
2. Data visualization (Boxplot, histogram, scatter plot)
3. Statistical tests(z-score- Empirical rule)
4. Inter quartile range



28. How you can handle outliers in the datasets.

Ans.

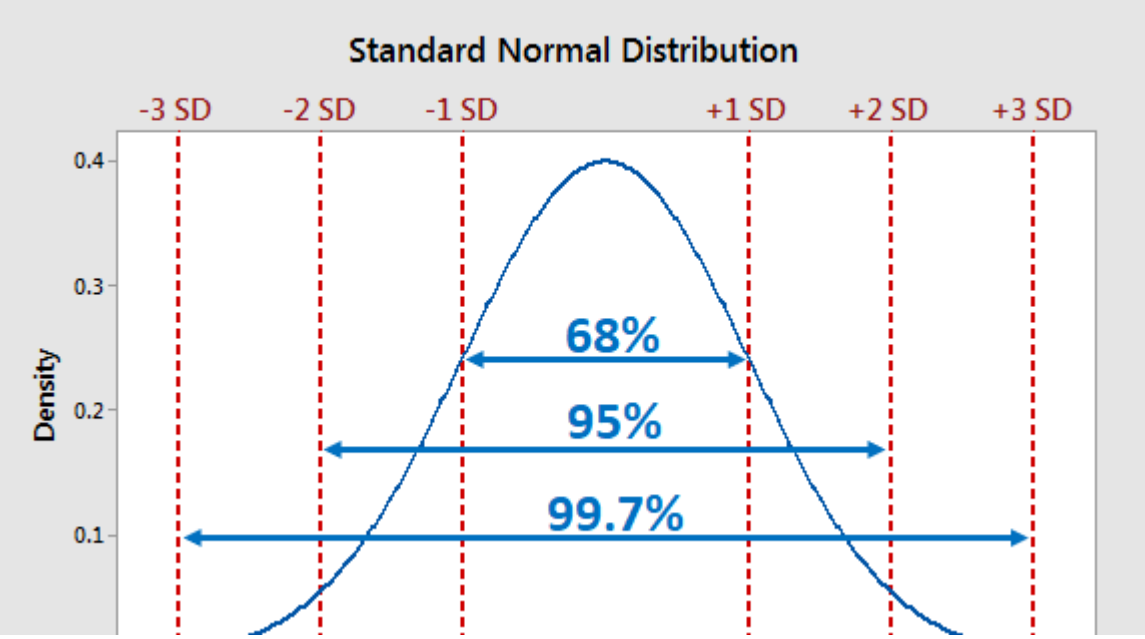
We can handle outliers by :

1. Drop outliers
2. Capping and replacing the outliers
3. Mean/median imputation
4. Transforming the dataset

29. What is the empirical rule?

Ans.

Empirical rule also known as 68-95-99.7% rule. Emperical rule estimates the percentage of dataset will lie within one, two, and three standard deviation of the mean.



30. How to calculate range and interquartile range?

Ans.

Range is the difference between minimum value from maximum value.

Interquartile range is the difference between Q3(75 percentile) and Q1(25 percentile).