Anomaly or outlier detection

Objectives

- ❑ What is Statistics?
- ❑ What is Population?
- ❑ What is Parameter?
- ❑ What is Sample?
- ❑ What is Mean?
- ❑ What is Median?
- ❑ What is Mode?
- ❑ What is Normal Distribution?
- ❑ Types of analysis in statistics
- ❑ What is an Outlier?
- ❑ What is Interquartile Range IQR?
- ❑ What are upper and lower limits in interquartile range?
- ❑ Demo: How to handle Outlier in the dataset?

What is Statistics?

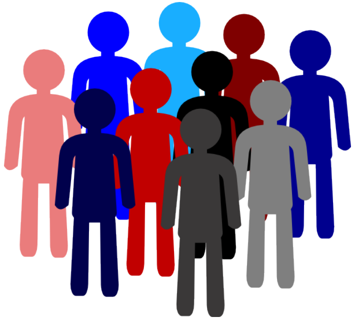
Statistics is a part of integrated applied mathematics which deals with data.

- It helps to collect data and analyze them properly.
- With the help of statistics we can read the data and organize them in order to get the hidden information from them.
- In data science domain statistics concepts are used to process the complex data to get the insights from them using mathematical computations.

What is Population?

The terms population in statistics is used to refer to the total set of observations.

Suppose,
We want to study a diabetes dataset to understand the symptoms and the other factors then the whole dataset is referred to as population.



What is Parameter?

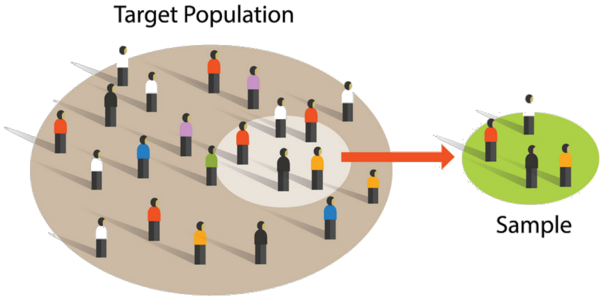
Parameters are referred to as characteristics which describe the population.

- Parameters are like average or percentage which help to describe the entire population.
- Mean and the standard deviation are two common parameters of population.
- Example: Average age for being diabetic is the parameter for whole diabetes dataset population.

What is Sample?

Sample is basically a small part or portion of the large population.

Suppose,
From the whole diabetes dataset you picked 100 rows of information to do the analysis, that 100 rows of information will be referred as Sample.

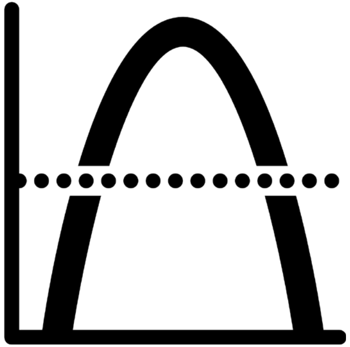


What is Mean?

The term Mean is referred to as the average value of the whole population.

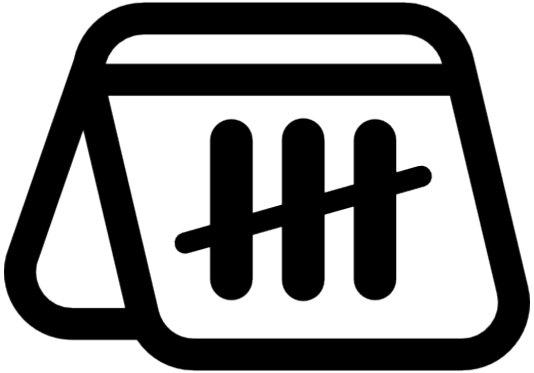
What is Median?

Median is the middle value of the data when your data is sorted in manner.



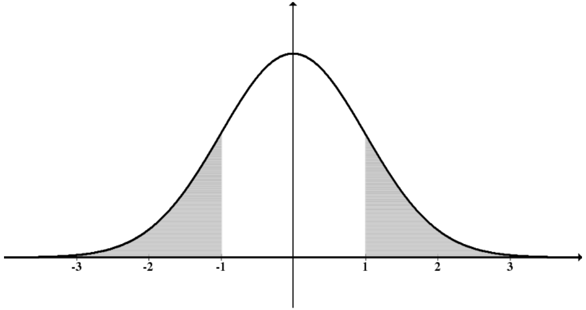
What is Mode?

Mode stands for the most occurring element in the dataset.



What is Normal Distribution?

The normal distribution is a probability function which describes how the values of a variable are distributed.



Properties of Normal distribution

- The mean, median and mode all are equal.
- The curve is symmetric at the center.
- This is also referred to as Gaussian or Gauss distribution.

Parameters of normal distribution

- Mean
- Standard deviation

Types of analysis in Statistics

Descriptive Statistics

It helps to describe the data in mathematical or graphical way.

Inferential Statistics

Inferential statistics split the data into samples and applies probability to arrive to the conclusion.

What is an Outlier?

Outliers in the dataset are referred to as the unusual value(s) which can distort and violate statistical analysis.

- Outliers are basically experimental errors in the data.
- Some outliers are good for the dataset to detect anomaly like: detecting fraud transaction
- It effects the mean and the standard deviation of the data and most of the machine learning technique does not perform good with outliers.

What is Interquartile Range IQR?

Interquartile range divides the dataset into quartiles to measure the variability and the spread of the dataset.

- Splits the data into 4 equal part in sorted manner
- Q1, Q2, Q3 are called first, second and third quartiles:
- Q1 → 25th percentile of the dataset
- Q2→ 50th percentile of the dataset
- Q3→ 75th percentile of the dataset
- Formula: IQR→ Q3 – Q1

What are upper and lower limits in interquartile range?

Lower and upper limits in the interquartile are basically the ranges where data points lie.

- Formula to find the lower limit:
- Lower_limit = Q1 - 1.5 IQR
- Formula to find the upper limit:
- Upper_limit = Q3 + 1.5 * IQR

Demo: How to handle Outlier in the dataset?

Thank you