Kanmani Thambidurai

**Statistics**

Table of Contents

[What is Statistics? 2](#_Toc49656449)

[Purpose of Statistics 2](#_Toc49656450)

[Terminologies used in Statistics 2](#_Toc49656451)

[Types of Statistics 2](#_Toc49656452)

[Variables and its Types 2](#_Toc49656453)

[Characteristics of a Frequency distribution 3](#_Toc49656454)

[Modality 3](#_Toc49656455)

[Symmetry 3](#_Toc49656456)

[Central Tendency 3](#_Toc49656457)

[Variability 3](#_Toc49656458)

[What is Outliers 4](#_Toc49656459)

[Mean,Mode,Median Imputation 4](#_Toc49656460)

[Dispersion measures 4](#_Toc49656461)

[Percentile 4](#_Toc49656462)

[Box Plot 4](#_Toc49656463)

[Method to identify Outlier 5](#_Toc49656464)

# What is Statistics?

Statistics is a branch of mathematics that helps to extract information from Data which helps in decision making. It involves collection,analysis,interpretation and presentation of data in the same order.

## Purpose of Statistics

* Present and describe the Data
* Draw conclusions from large population of data using sample data
* Make reliable forecasts with the information acquired from studying sample data
* To improve the business process

## Terminologies used in Statistics

1. Population – Data that is grouped together by one or more common features,properties
2. Sample – subset of population

## Types of Statistics

Statistics is classified into two types

* Descriptive Statistics – Descriptive Statistics presents, describes the population in a meaningful way using graphs, charts and tables.
* Inferential Statistics - Inferential Statistics make predictions about a population based on a sample of data taken from that population.

## Variables and its Types

Variable is a feature or properties of the population that can be measured or counted or categorised. There are two types of variables namely Quantitative and Qualitative Variables.

* Qualitative Variable: It is categorical data that describes the qualities or categorises the data.
* Nominal – Non numerical data that cannot be ordered or ranked logically

Eg: Pin code(even though it is a number doesn’t measure or counts),Gender,Religion

* Ordinal – Non-Numerical data that can be ordered or ranked logically

Eg: Movie ratings,School ranks

* Quantitative Variable: It is numerical data that measures or counts the data the quantities
* Discrete – Numerical values that are countable. It will not take any decimal values.

Eg: Number of kids in the family,Number of cars in a apartment

* Continuous – Numerical values that are measurable.It takes decimal values.

Eg: Height , Weight,Temperature

Independent Variable: Input variables that are fed in. Represented on X axis.

Dependent Variable : Output variables that are dependent on the input variables. Represented on Y axis.

(In terms of machine learning, columns that has to be predicted is dependent variable as it is depended on the features of the input variables that are fed in.Columns that are used to predict are independent variables )

Univariate Analysis : Analysis done on single variable or column is Univariate Analysis

Bivariate Analysis : Analysis done on two variables or columns is Bivariate Analysis

## Characteristics of a Frequency distribution

### Modality

* Unimodal – has a single mode
* Bimodal – has two modes i.e. two values in the population are found most frequently with equal number of occurrences

### Symmetry

* Symmetric – when the distribution in the graph is divided at the centre,it divides into two halves,mirroring each side.

In a symmetric distribution, the [mean, mode and median](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/)all fall at the same point.

* Asymmetric- the distribution cannot be divided into equal halves
* Positive Skewness – When the tail is longer on the right side of the graph and the distribution is more concentrated on the left side.
* Negative Skewness-– When the tail is longer on the left side of the graph and the distribution is more concentrated on the right side.

### Central Tendency

Central Tendency identifies the central position around which the data is distributed for that set of data. Central Tendency can be computed in any of the below three method.

* Mode

A number that appears most often in the set of data is the mode of the set.

* Median

Median is found by sorting the set in ascending order and finding the exact middle value.

If ‘n’ is the number of values in a set,

* When ‘n’ is odd , Median is [(n + 1) ÷ 2]th value i.e. exact middle number of the sequence
* When ‘n’ is even,Median is[(n/2)th value + {(n/2)+1}th value]/2

i.e. mean of the two middle numbers of the sequence

* Mean

Mean is the average of the data in the set. It is found by the sum of the items in the set divided by the number of items in the set.

If the data is a [population](https://www.statisticshowto.com/what-is-a-population/), then the mean is called a [population mean](https://www.statisticshowto.com/population-mean/), denoted by ‘μ’. If the data is a [sample](https://www.statisticshowto.com/sample/), then the mean is called a [sample mean, denoted by](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/sample-mean/)x̄.

### Variability

* Range is the measure of the spread of the data.

It can be calculated using Maximum value of the set – Minimum value of the set.

* Standard Deviation is also the measure of the spread of data but in a more detailed way.

It measures how the values are spread around the mean in a set.

Standard deviation is the square root of Variance.

Variance is the average of squared distance of the mean and each value.

A large standard deviation indicates that the values are far from the mean and a small standard deviation indicates that the values are clustered closely around the mean.

## What is Outliers

An outlier is an value of the population that lies in an abnormal distance from other values in the

population.

So when there is outliers in a column, Median should be considered as the measure of central of tendency instead of mean.

Also outliers should be imputed with median instead of mean, as imputing with mean will cause the data to be skewed towards the outliers.

## Mean,Mode,Median Imputation

When the column of analysis have Null values ,it is not advisable to remove those columns as it impacts the calculation of centre of tendency.Instead substitute the null values with any of mean,median or mode.This is called Mean,Mode,Median Imputation.

## Dispersion measures

Dispersion measures tells us how far the data is spread from the centre. Below are the methods to measure the dispersion.

* Range
* Standard Deviation
* Mean Absolute Deviation
* Interquartile Range

## Percentile

Percentile is a number where a certain percentage of values fall below that number.

Eg: If you scored 90th percentile in a test, that means you scored greater than 90% of people who took the test.

Formula to calculate Percentile is R = (P/100) x N, where P is the desired percentile,N is the number of values,R is the rank assigned by sorting the data is ascending order

## Box Plot

Box plot is a 5 number summary of a data.Below are the 5 numbers that are used to summarise the data.

1. Min
2. 25th percentile – Quartile 1 – Q1
3. 50th percentile – Quartile 2 – Q2 (Ideally is the mean)
4. 75th percentile – Quartile 3 – Q3
5. Max

Inter quartile range(IQR) is calculated by Q3-Q1

Sample Box Plot

50% Percentile(Q2)

min max

25

25% percentile(Q1) 75% percentile(Q3)

## Method to identify Outlier

@@@@

\*\*\*\* 25

Lower limit Upper Limit

Lower limit or Lower Whisker is Q1-1.5(Q3-Q1), any values lower than this lower limit is outliers

Upper limit or Upper Whisker is Q3+1.5(Q3-Q1), any values greater than this upper limit is outliers

where 1.5 is a constant, (Q3-Q1) is inter quartile range

From the above Box plot, \*\*\*\*\* is lower outliers, @@@@ is upper outliers