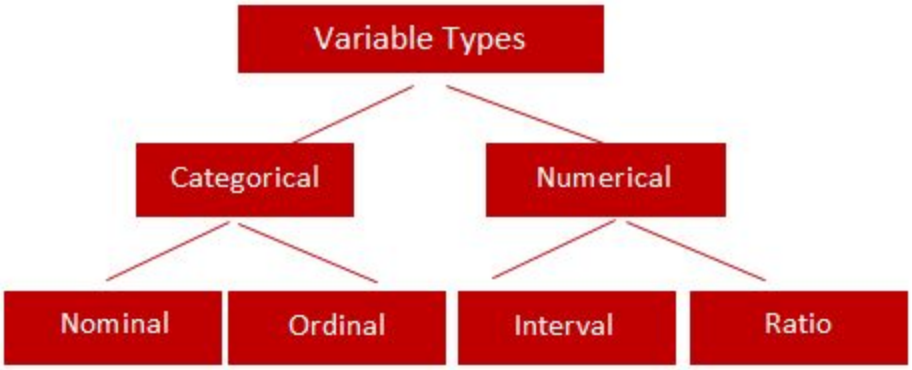
Statistics Needed for Data Science

What is statistics:
Definition: Study of the collection, analysis, interpretation, presentation, and organization of data.

Key concepts: include probability distributions, statistical significance, hypothesis testing, and regression.

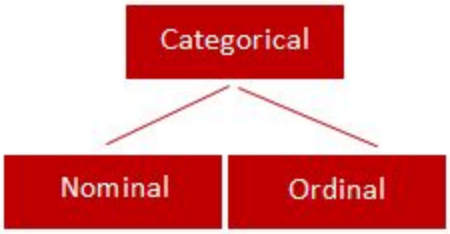
1. Basics Statistics:

Types of Variables



Categorical

Qualitative data are often termed **categorical data**. Data that can be added into **categories** according to their characteristics.



Nominal Variable (Unordered list)

A variable that has two or more categories, without any implied ordering.
Examples :

- Gender - Male, Female
- Marital Status - Unmarried, Married, Divorcee
- State - New Delhi, Haryana, Illinois, Michigan

Ordinal Variable (Ordered list)

A variable that has two or more categories, with clear ordering.

Examples :

- Scale - Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree
- Rating - Very low, Low, Medium, Great, Very great

Interval

An interval variable is similar to an ordinal variable, except that the intervals between the values of the interval variable are equally spaced. In other words, it has order and equal intervals.

Examples :

- Temperature in Celsius - Temperature of 30°C is higher than 20°C, and temperature of 20°C is higher than 10°C. The size of these intervals is the same.
- Annual Income in Dollars - Three people who make $5,000, $10,000 and $15,000. The second person makes $5,000 more than the first person and $5,000 less than the third person, and the size of these intervals is the same.

Ratio It is interval data with a natural zero point. When the variable equals 0.0, there is none of that variable.

Examples :

- Height
- Weight
- Temperature in Kelvin - It is a ratio variable, as 0.0 Kelvin really does mean 'no temperature.

2. Descriptive statistics

It provides information on summary statistics that includes *Mean, Standard Error, Median, Mode, Standard Deviation, Variance, Kurtosis, Skewness, Range, Minimum, Maximum, Sum, and Count*.

Measure of Central Tendency

It describes a whole set of data with a single value that represents the centre of its distribution. *There are three main measures of central tendency: the mode, the median and the mean.*

Mean, Median and Mode

Mean	Average value
Median	Middle value
Mode	Most frequent value

Mean *It is the sum of the observations divided by the sample size.*
The mean of the values 5,6,6,8,9,9,9,9,10,10 is
(5+6+6+8+9+9+9+9+10+10)/10 = 8.1

Limitation : *It is affected by extreme values. Very large or very small numbers can distort the answer.*

Median *It is the middle value. It splits the data in half. Half of the data are above the median; half of the data are below the median.*

Advantage : *It is* ***NOT*** *affected by extreme values. Very large or very small numbers does not affect it*

Mode *It is the value that occurs most frequently in a dataset*
Advantage : *It can be used when the data is not numerical.*
Disadvantage :
1. There may be no mode at all if none of the data is the same
2. There may be more than one mode

When to use mean, median and mode?

Mean – When your data is not skewed i.e normally distributed. In other words, there are no extreme values present in the data set (Outliers).
Median – When your data is skewed or you are dealing with ordinal (ordered categories) data (e.g. likert scale 1. Strongly dislike 2. Dislike 3.Neutral 4. Like 5. Strongly like)
Mode - When dealing with nominal (unordered categories) data.

Measure of Dispersion

It refers to the spread or dispersion of scores. There are four main measures of variability: *Range, Standard deviation and Variance.*

Range	Difference between max and min in a distribution
Standard Deviation	Average distance of scores in a distribution from their mean
Variance	Square of the standard deviation
Skewness	Degree to which scores in a distribution are spread out.
Kurtosis	Flatness or peakness of the curve

Range *It is simply the largest observation minus the smallest observation.*
Advantage: *It is easy to calculate.*
Disadvantage: *It is very sensitive to outliers and does not use all the observations in a data set.*

Standard Deviation *It is a measure of spread of data about the mean.*
Advantage : *It gives a better picture of your data than just the mean alone.*
Disadvantage :
1. It doesn't give a clear picture about the whole range of the data.
2. It can give a skewed picture if data contain outliers.

Skewness *It is a measure of symmetry. A distribution is symmetric if it looks the same to the left and right of the center point.*

Kurtosis *It is a measure of whether the data are peaked or flat relative to the rest of the data. Higher values indicate a higher, sharper peak; lower values indicate a lower, less distinct peak.*

Examples:
Example 1: Suppose you are asked to provide a figure that best describes the annual salary offered to students in ABC College. You would answer this question with a measure of central tendency and variability.

3. Population and Sample:

Population: A population includes all of the elements from a set of data.
Sample: A sample consists of one or more observations drawn from the population.
Population study means census study and sample study means go with some units.

4. Data representation:

Pie Chart:
A pie chart shows the relative proportions of data in different categories.

Bar Chart
A bar chart displays frequencies of categories of data.

Histogram:
A histogram displays frequencies of quantitative data that has been sorted into intervals.

Boxplot:

A boxplot (also known as a box and whiskers plot) is another way to display quantitative data. It displays the five 5 number summary (minimum, Q1, median, Q3, maximum).

5. Why we standardise data:

The concept of standardization comes into the picture when continuous independent variables are measured at different scales. It means these variables do not give equal contribution to the analysis.
Example: income of the person it varies from $100 to $10000.

6. Skewness and Kurtosis:

Skewness: Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.
Kurtosis: Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers.