Mathematics

What is statistics?

Statistics is a branch of mathematics that involves collecting, analysing, interpreting and presenting data. It provides tools and methods to understand and make sense of large amounts of data and to draw conclusion and make decision based on the data.

In practice, statistics is used in a wide range of fields, such as business, economics, social science, medicine and engineering. It is used to conduct research studies, analyse market trends, evaluate the effectiveness of treatments and interventions, and make forecasts and predictions

Examples

1. Business – Data Analysis (Identifying customer behavior) and Demand Forecasting.
2. Medical – Identify efficacy of new medicines (clinical trials); Identifying risk factor for diseases (Epidemiology)
3. Government & Politics – conducting surveys, polling
4. Environmental Science – Climate research

Statistics

Descriptive

Data is given, and the descriptive (summary detail) are generated with the help of statistics.

Inferential

Inferential Statistics use to prediction.

Population Vs Sample

Population refers to the entire group of individual or objects that we are interested in studying. It is the complete set of observation that we want to make inferences about. For example, the population might be all the students in a particular school or all the cars in a particular city.

A sample, on the other hand, is a subset of the population. It is a smaller group of individuals or objects that we select from the population to study. Samples are used to estimate characteristics of the population such as the mean or the poportion with a certain attribute. For example, we might randomly select 100 students

Example

1. All cricket fans vs fans who were present in the stadium.
2. All student vs who visit college for lectures

Things to be careful about which creating samples

1. Sample Size
2. Random
3. Representative

Parameter Vs Statistics

A parameter is a characteristic of a population, while a statistics is a characteristics of a sample. Parameters are generally unknown and are estimated using Statistics. The goal of statistical inference is to use the information obtained from the sample to make inference about the population parameter.

Inferential Statistics

Inferential statistics is a branch of statistics that deals with making inferences or predictions about a larger population based on a sample of data. It involves using statistical techniques to test hypotheses and draw conclusion from data. Some of the topics that come under inferential statistics are:

1. Hypothesis testing: This involves testing a hypothesis about a population parameter based on a sample of data. For example, testing whether the mean height of a population is different from a given value.

2. Confidence Intervals: The involves estimating the range of values that a population parameter could take based on a sample of data. For example, estimating the population mean height within a given confidence level.

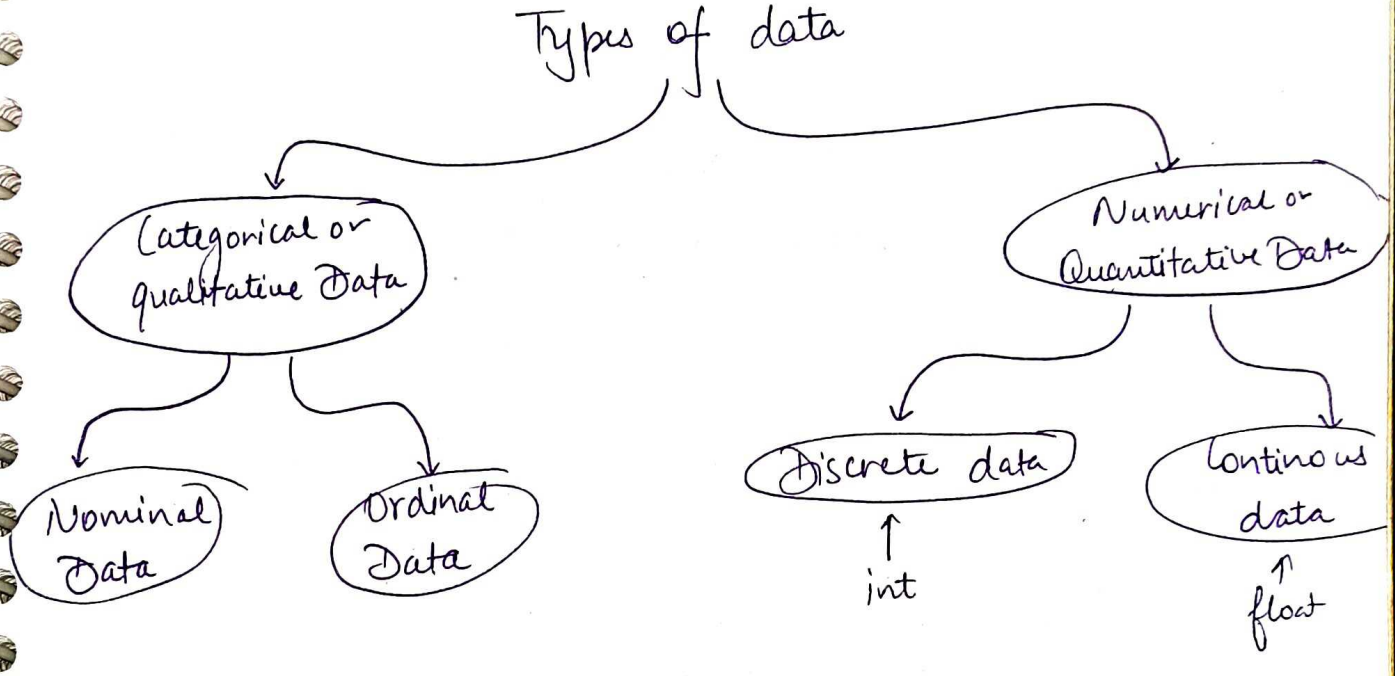
3. Analysis of Variance (ANOVA): This involves comparing means across multiple groups to determine if there are any significant differences. For example, comparing the mean height of individuals from different regions.

4. Regression analysis: This involves modeling the relationship between a dependent variable and one or more independent variables. For example, predicting the sales of a product based on ~~advantage~~ advertising expenditure.

5. Chi-square tests: This involves testing the independence or association between two categorical variables. For example, testing whether gender and occupation are independent variables.

6. Sampling techniques : This involves ensuring that the sample of data is representation the population. For example, using random sampling to select individuals from a population.

7. Bayesian statistics



Measure of Central Tendency

A measure of central tendency is a statistical measure that represents a typical or central value for a dataset. It provides a summary of the data by identifying a single value that is most representative of the dataset as a whole.

eg. Age

Find the centeralized

Central Tendency

Mean Median Mode

1. Mean

The mean is the sum of all values in the dataset divided by the number of values.

Population Mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Outlier → create issue

2. Median

The median is the middle value in the dataset when the dataset is arranged.

3. Mode

The mode is the value that appearance most frequently in the dataset.

4 Weighted Mean

The weighted mean is the sum of the products of each value and its weighted, divided by the sum of the weights. It is used to calculate a mean when the values in the datasets have different importance or frequency.

LR 0.2 → 10L

RF 0.3 → 15L

xgboost 0.5 → 12L

0.5 ← weight

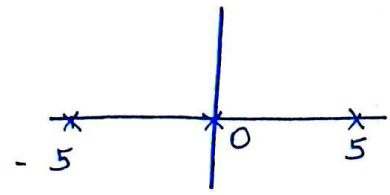
$$= \frac{0.2 \times 10 + 0.3 \times 15 \times 0.5}{0.2 + 0.3 + 0.5}$$

5. Trimmed Mean

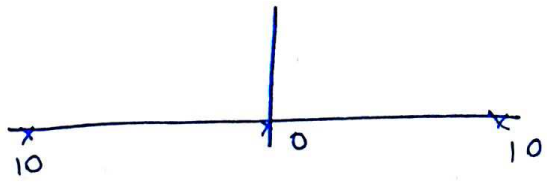
A trimmed mean is calculated by removing a certain percentage of the smallest and largest values from the dataset and then taking the mean of the remaining values. The percentage of values removed is called the trimming percentage.

Measure of Dispersion

A measure of dispersion is a statistical measure that describe the spread or variability of a dataset. It provides information about how the data is distributed around the central tendency (mean, median or mode) of the dataset.



$$\frac{-5+0+5}{3} = 0$$



$$\frac{-10+0+10}{3} = 0$$

1. Range

The Range is the diff^{n} betn the maximum and minimum values in the dataset. It is a simple measure of dispersion that is easy to calculate but can be affected by outliers.

$$\text{range} = 5 - (-5) = 10 \text{ range}$$

$$10 - (-10) = 20$$

Range affected by outliers.

1. Range

The range is the difference between the maximum values in the dataset. It is a simple measure of dispersion that is easy to calculate but can be affected by outliers.

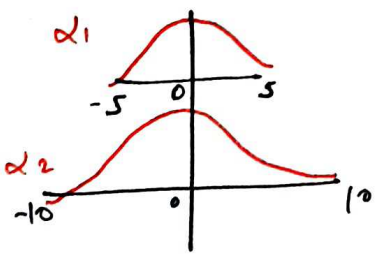
2. Variance

The variance is the average of the squad differences between each data point and the mean. It measures the average distance of each data point from the mean and is useful in comparing the dispersion of datasets with different means.

x	X-mean	(x-mean)^2
3	3-3	0
2	2-3	1
1	1-3	4
5	5-3	4
4	4-3	1

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Variance is not giving exact spread value but Variance is directly propotional to spread.



$\alpha_1 < \alpha_2$

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N}$$ Population Variance

$$S^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$ Sample Variance

We can also use (Mean Absolute error)

$$MAD = \frac{\sum |x-\bar{x}|}{N}$$

↳ Not inference → pop data ka nikal pate. variance nhi

↳ good for outlier

↳ Not use for finding variance

3. Standard Deviation

The standard deviation is the square roots of the variance. It is a widely used measured of dispersion that is useful in describing the shape of a distribution.

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N}$$ Population Variance

$$S^2 = \frac{\sum (x-\bar{x})^2}{N}$$ Sample Variance

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$ Population SD

$$\sigma = \sqrt{\frac{\sum (x-\bar{x})^2}{N}}$$ Sample SD

Standard Deviation → Same Unit bcz of under root

$$\text{Variance} = (5\,\text{lpa})^2 = 5\,\text{lpa}^2$$

$$SD = \sqrt{5\,\text{lpa}^2} = 5\,\text{lpa}$$

4. Coefficient of Variance

Coefficient of variation (CV): The CV is the ratio of the standard deviation to the mean expressed as a percentage. It is used to compare the variability of datasets with different means and is commonly used in field such as bilogical, chemistry, and eyjineu.

The coefficient of variation (CV) is a statistical measure that expresses the amount of their of variability in a dataset relative to the mean ~~of~~ data. It an dimensionless quantity that is expressed as a percentage.

$$CV = \left(\left(\text{"Standard devision"}\right) \times (\text{mean})\right) \times 100\%$$

Graph of Univariate Analysis

Categorical

Numerical

1. Categorical – Frequency Distribution Table & cumulative frequency

A frequency Distribution table is a table that summarizes the number of times (or frequency) that each value occurs in a dataset.

Let's say we have a survey of 200 people and we ask them about their favourite type of vacation, which could be one of six categories; Beach, City, Adventure, Nature, Cruise, or Other.

Types of Vacation	Frequency
Beach	60
City	40
Adventure	30
Nature	35
Cruise	20
Other	15

Bar chart

Relative Frequency is the proportion or percentage of a category in a dataset or sample. It is calculated by dividing the frequency of a category by the total number of observation in the dataset or sample.

Types of Vacation	Frequency	Relative Frequency
Beach	60	0.3
City	40	0.2
Adventure	30	0.15
Nature	35	0.175
Cruise	20	0.1
Other	15	0.075

Pie chart

Percentage

Cumulative Frequency is the running total of frequencies of a variable or category in a dataset or sample. It is calculated by adding up the frequencies of the current category and all previous categories in the dataset or sample.

Type of Vacation	Frequency	Relative Frequency	Cumulative Freq
Beach	60	0.3	60
City	40	0.2	100
Adventure	30	0.15	130
Nature	35	0.175	165
Cruise	20	0.1	185
Other	15	0.075	200

lineplot

2. Numerical - Frequency Distribution Table & Histogram

Age

32

28

41

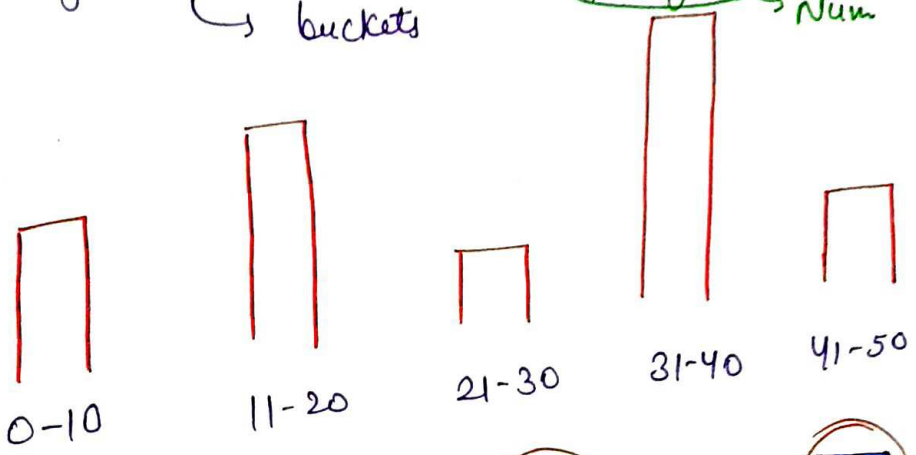
42

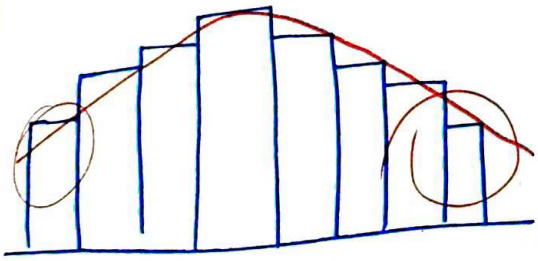
Categories → bins

↳ buckets

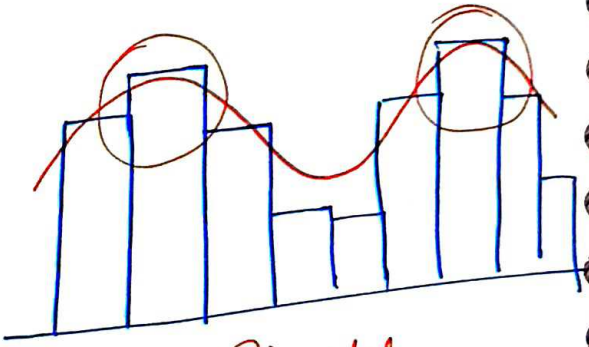
barchart → Cate

Histogram → Num

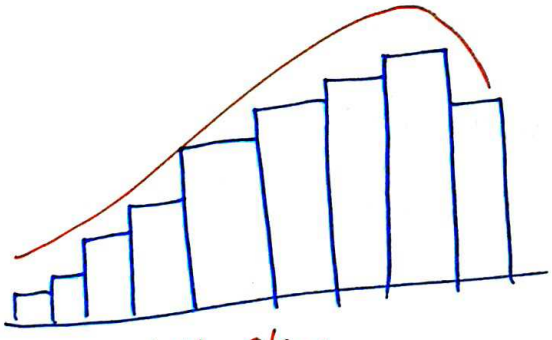




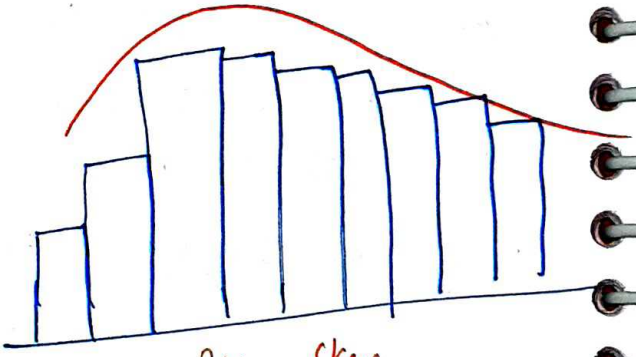
Symmetric



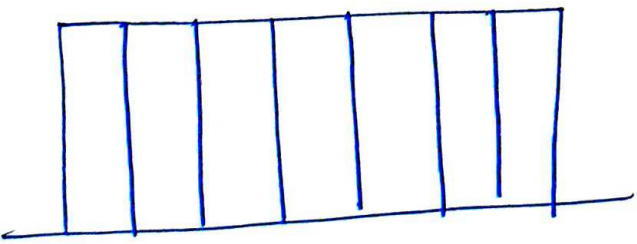
Bimodal



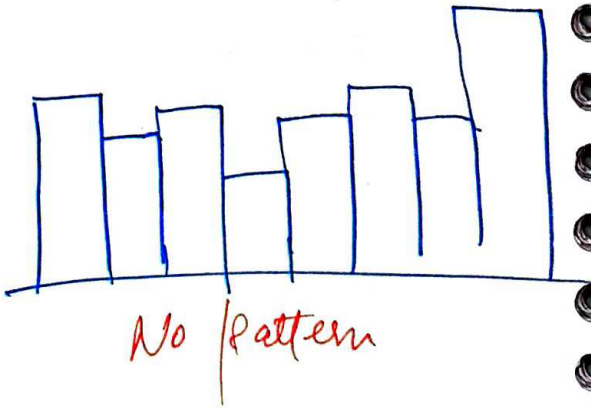
Left Skew



Right Skew

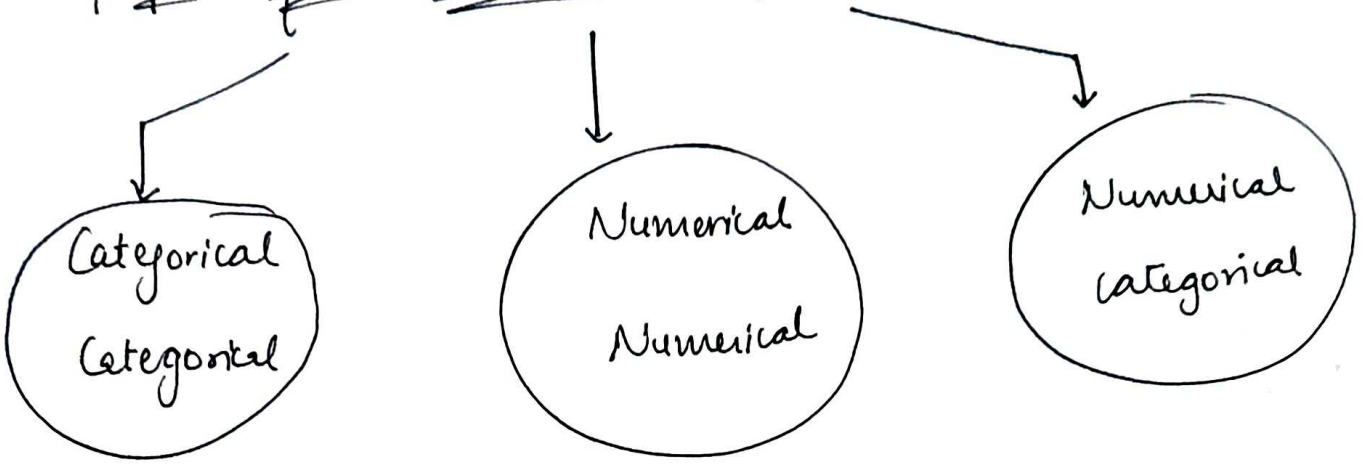


Uniform Skew



No Pattern

Graph for Bivariate Analysis



1. Categorical - Categorical

Contigency Table / Crosstab

A contigency table, also known as a cross-tabulation or crosstab is a type of table used in statistics to summarize the relationship between two categorical variables. A contigency table displays the frequencies or relative frequencies of the observed values of the two variables, organized into rows and columns.

Survived

0

1

⋮

Pclass

1

2

3

⋮

⇒

Pclass

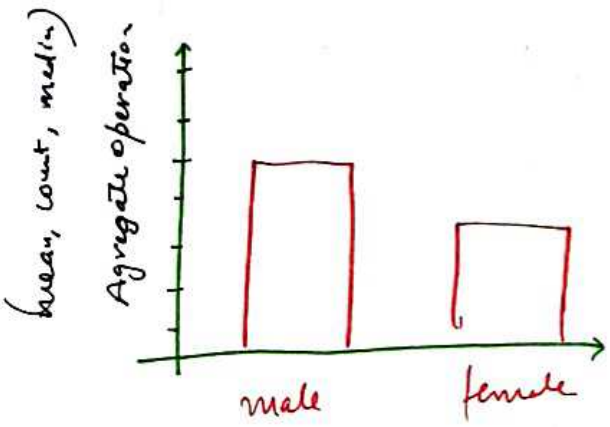
Survived	1	2	3
0	42	31	63
1	24	36	54

Contigency table

2. Numerical — Numerical

Scatter Plot

3. Categorical - Numerical



* also create Contigency table

	0-10	10-20	20-30
male	32	41	110
Female	15	18	120