What is Statistics

09 March 2023   14:56

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 conclusions and make decisions based on the data.

In practice, statistics is used in a wide range of fields, such as business, economics, social sciences, 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

Types of Statistics

Statistics

Descriptive

Inferential
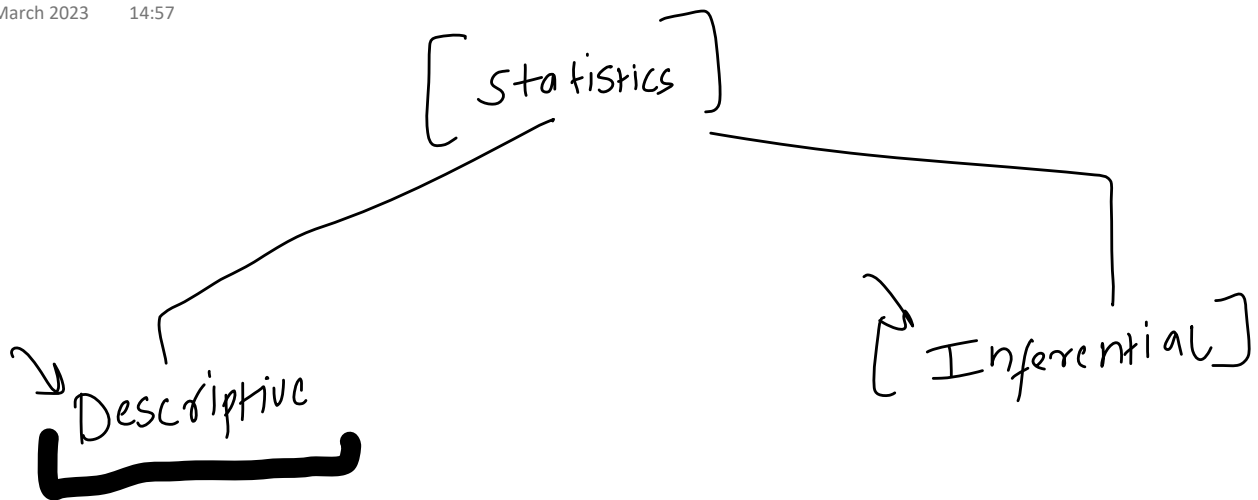
Descriptive statistics deals with the collection, organization, analysis, interpretation, and presentation of data. It focuses on summarizing and describing the main features of a set of data, without making inferences or predictions about the larger population.

Inferential statistics deals with making conclusions and predictions about a population based on a sample. It involves the use of probability theory to estimate the likelihood of certain events occurring, hypothesis testing to determine if a certain claim about a population is supported by the data, and regression analysis to examine the relationships between variables

Population Vs Sample

09 March 2023   14:57

Population refers to the entire group of individuals or objects that we are interested in studying. It is the complete set of observations 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 proportion with a certain attribute. For example, we might randomly select 100 students.

Examples

1. All cricket fans vs fans who were present in the stadium
2. All students 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 statistic is a characteristic 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 inferences about the population parameters.

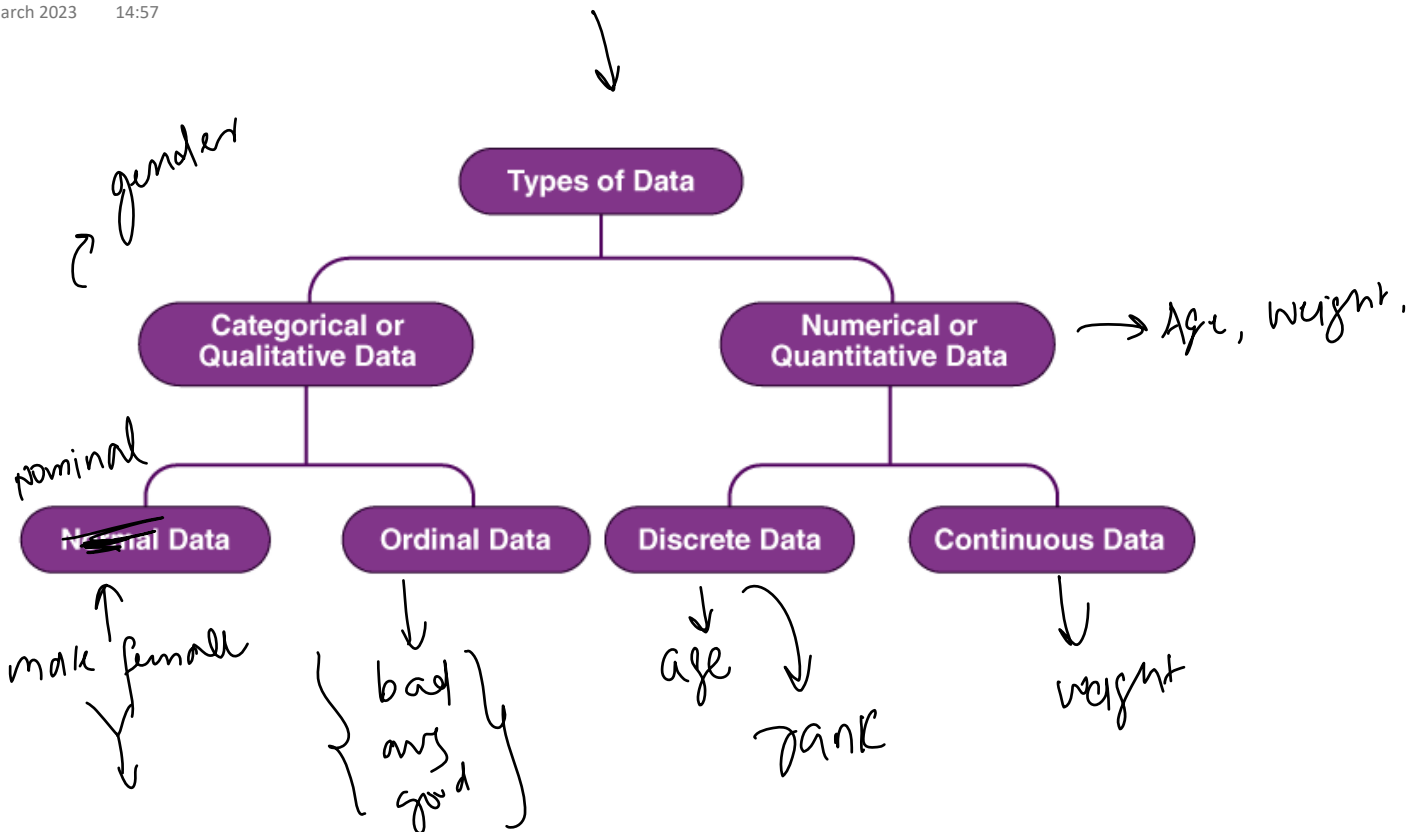# Inferential Statistics

09 March 2023   14:57

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 conclusions 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**: This 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 modelling the relationship between a dependent variable and one or more independent variables. For example, predicting the sales of a product based on 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 representative of the population. For example, using random sampling to select individuals from a population.
7. **Bayesian statistics**: This is an alternative approach to statistical inference that involves updating beliefs about the probability of an event based on new evidence. For example, updating the probability of a disease given a positive test result.

Why ML is closely associated with statistics?

Types of Data

gender

```
                          Types of Data
                                |
            +-------------------+-------------------+
            |                                       |
    Categorical or                          Numerical or          → Age, Weight.
    Qualitative Data                        Quantitative Data
            |                                       |
      +-----+-----+                          +-------+--------+
      |           |                          |                |
  Nominal      Ordinal Data            Discrete Data    Continuous Data
  Data
```

nominal

male female

bad
avg
good

age
rank

weight

Measure of Central Tendency

09 March 2023   14:58

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.

Age

mean   median   mode
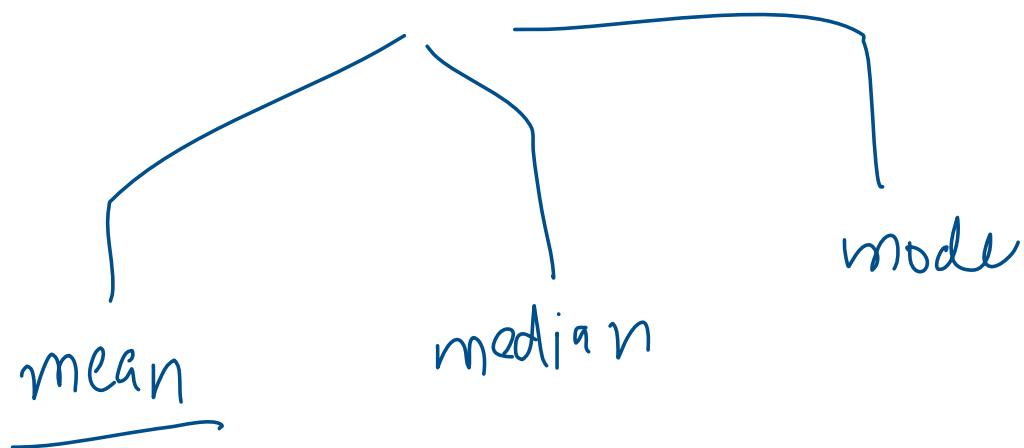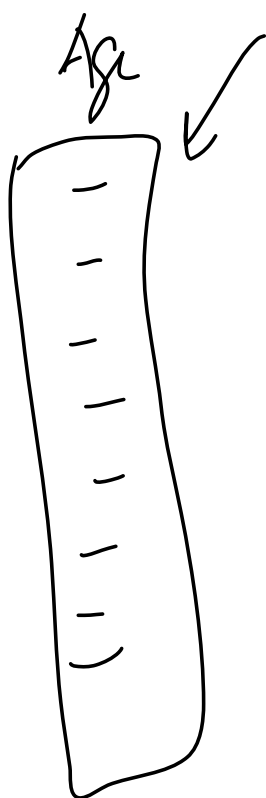
1. Mean

Mean: The mean is the sum of all values in the dataset divided by the number of values.
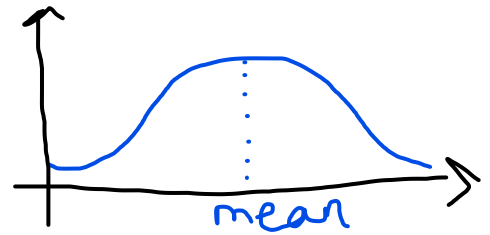
Prone to Outlier

3

4

$$= \frac{15}{5} = 3$$

1
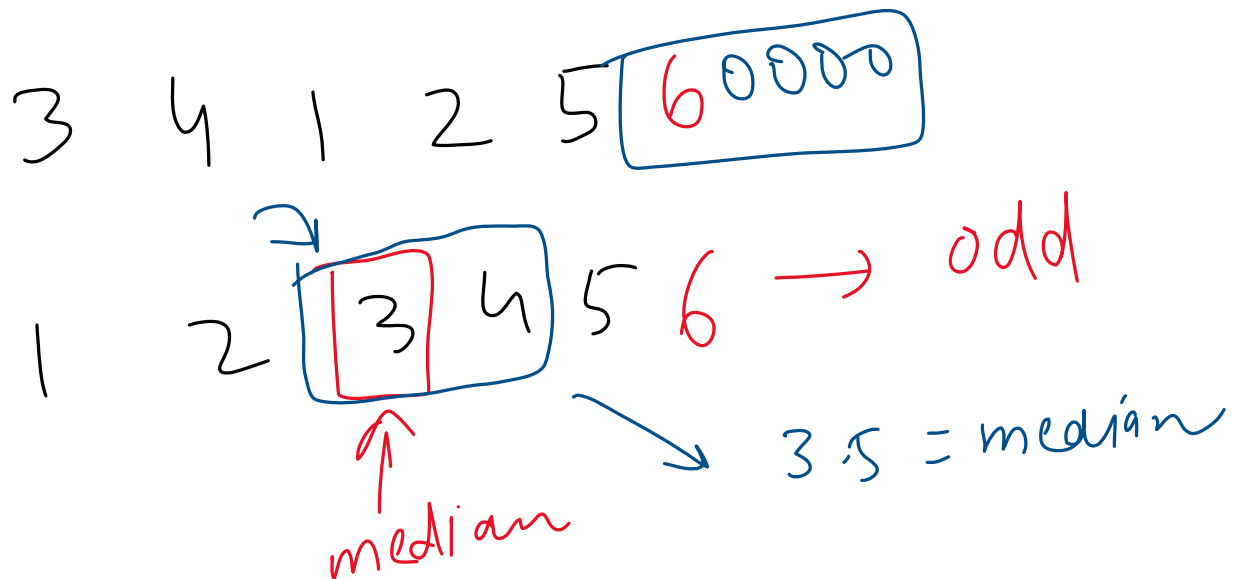
2

5

mean

$$\frac{\sum\limits_{i=1}^{N} x_i}{N}$$

$$\frac{\sum\limits_{i=1}^{n} x_i}{n}$$

Population Mean	Sample Mean
$\mu = \dfrac{\sum\limits_{i=1}^{N} x_i}{N}$	$\overline{X} = \dfrac{\sum\limits_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

2. Median

Median: The median is the middle value in the dataset when the data is arranged in order.

$$3 \quad 4 \quad 1 \quad 2 \quad 5 \quad \boxed{6\,0\,0\,0\,0}$$

$$1 \quad 2 \quad \boxed{3 \quad 4} \quad 5 \quad 6 \quad \rightarrow \quad \text{odd}$$

median

$$3.5 = \text{median}$$

3. Mode

Mode: The mode is the value that appears most frequently in the dataset.

1   1   2   3   1   4   5   6   1   2

mode = 1
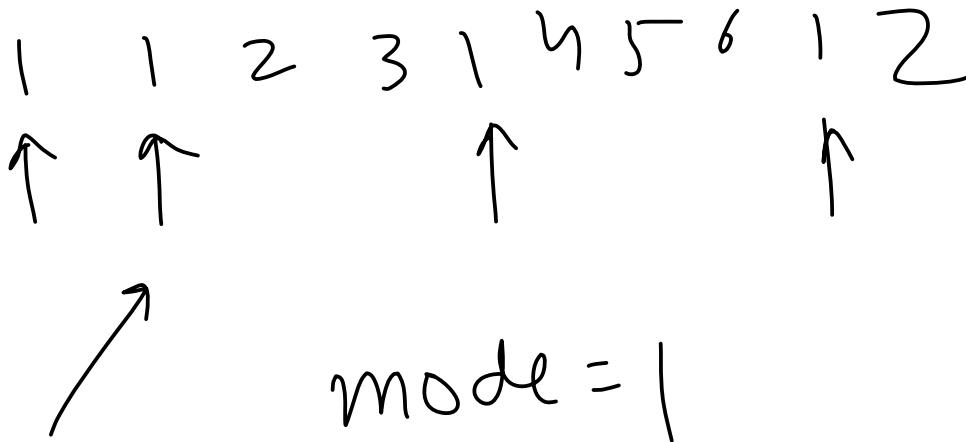
4. Weighted Mean

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

$$\boxed{1} \; \boxed{2} \; \boxed{3} \; = \; \boxed{\dfrac{1+2+3}{3}}$$

$$\rightarrow \boxed{LR} \; \dfrac{0.2}{1} \rightarrow 10l$$

$$\rightarrow \boxed{RF} \; 0.3 \rightarrow 15l$$

$$\rightarrow \boxed{X\$600} \; 0.5 \rightarrow 12l$$

$$\searrow \dfrac{0.2 \times 10l + 0.3 \times 15l + 0.5 \times 12l}{0.2 + 0.3 + 0.5}$$

5. Trimmed Mean

10 March 2023   09:37

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.

20%

$20,000, $22,000, $23,000, $25,000, $28,000, $30,000, $32,000, $35,000, $50,000, $80,000 $36,500.

$25,000, $28,000, $30,000, $32,000, $35,000 $30,000

Measure of Dispersion

A measure of dispersion is a statistical measure that describes 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.

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

$$5 - (-5) = 10$$

$$20$$

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

1. Range

09 March 2023   16:36

Range: The range is the difference between 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.

2. Variance

not find dispersion, directly proportional to dispersion

disadv - prone to outlier

Variance: The variance is the average of the squared 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.

③

$\bar{x} = 3$ $(x_i - \bar{x})$

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

	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

add $-5, 0, 5$

$\frac{10}{5}$ $\sigma = 2$

$$\sigma^2 = \frac{(x_1 - \bar{x})^① + (x_2 - \bar{x})^② + (x_3 - \bar{x})^②}{3}$$

$$\frac{(-5)^2 + (0)^2 + (5)^2}{3} = \frac{50}{3}$$

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

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

$\sigma_1^2 < \sigma_2^2$

$\bar{x}$ -5 0 5 $\bar{x}$

-10 0 10
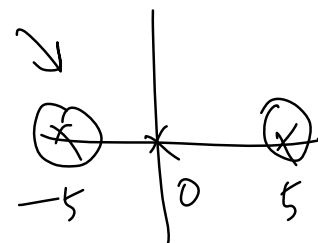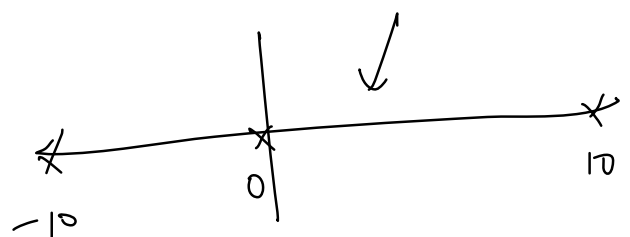
Mean Absolute Deviation

$$MAD = \frac{\sum |x_i - \bar{x}|}{n}$$
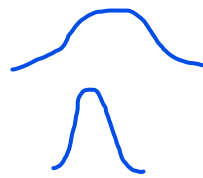
→ inference

not use for prediction or inference

Why std → because it has the same units as the data, whereas variance has squared units

3. Standard Deviation

09 March 2023   16:37

Standard Deviation: The standard deviation is the square root of the variance. It is a widely used measure of dispersion that is useful in describing the shape of a distribution.

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \quad \textbf{Population Standard deviation}$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \quad \textbf{Sample Standard deviation}$$

SD unit → same → data

$\sqrt{2} \quad 1.41$

15
17   → 15
13
14

$\sqrt{\sigma^2} \quad \sqrt{5}$

σ SD

$$\frac{(15-15)^2 + (17-15)^2 + (13-15)^2 + (14-15)}{4}$$

4. Coefficient of Variation

09 March 2023   16:37

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 fields such as biology, chemistry, and engineering.

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

The formula for calculating the coefficient of variation is:

CV = (standard deviation / mean) x 100%
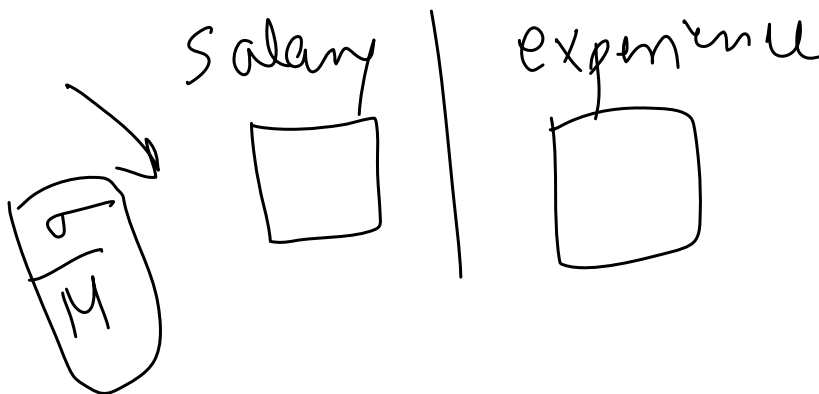
Graphs for Univariate Analysis

Categorial   Numerical
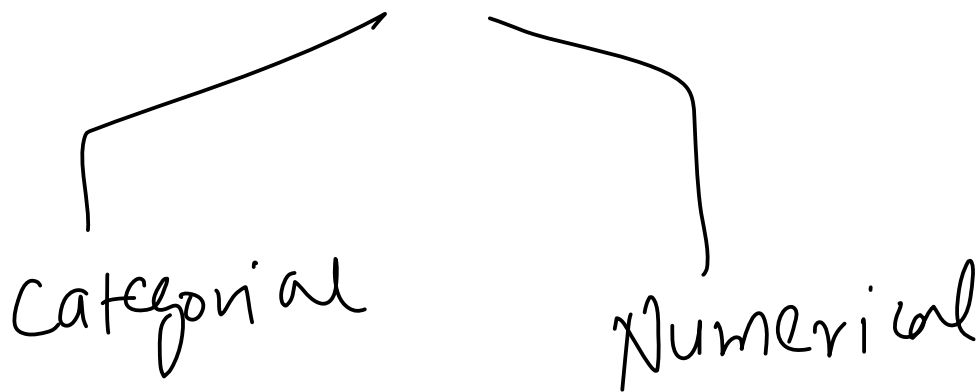
1. Categorical - Frequency Distribution Table & Cumulative Frequency

09 March 2023   16:50

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

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

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 observations in the dataset or sample.

Type 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

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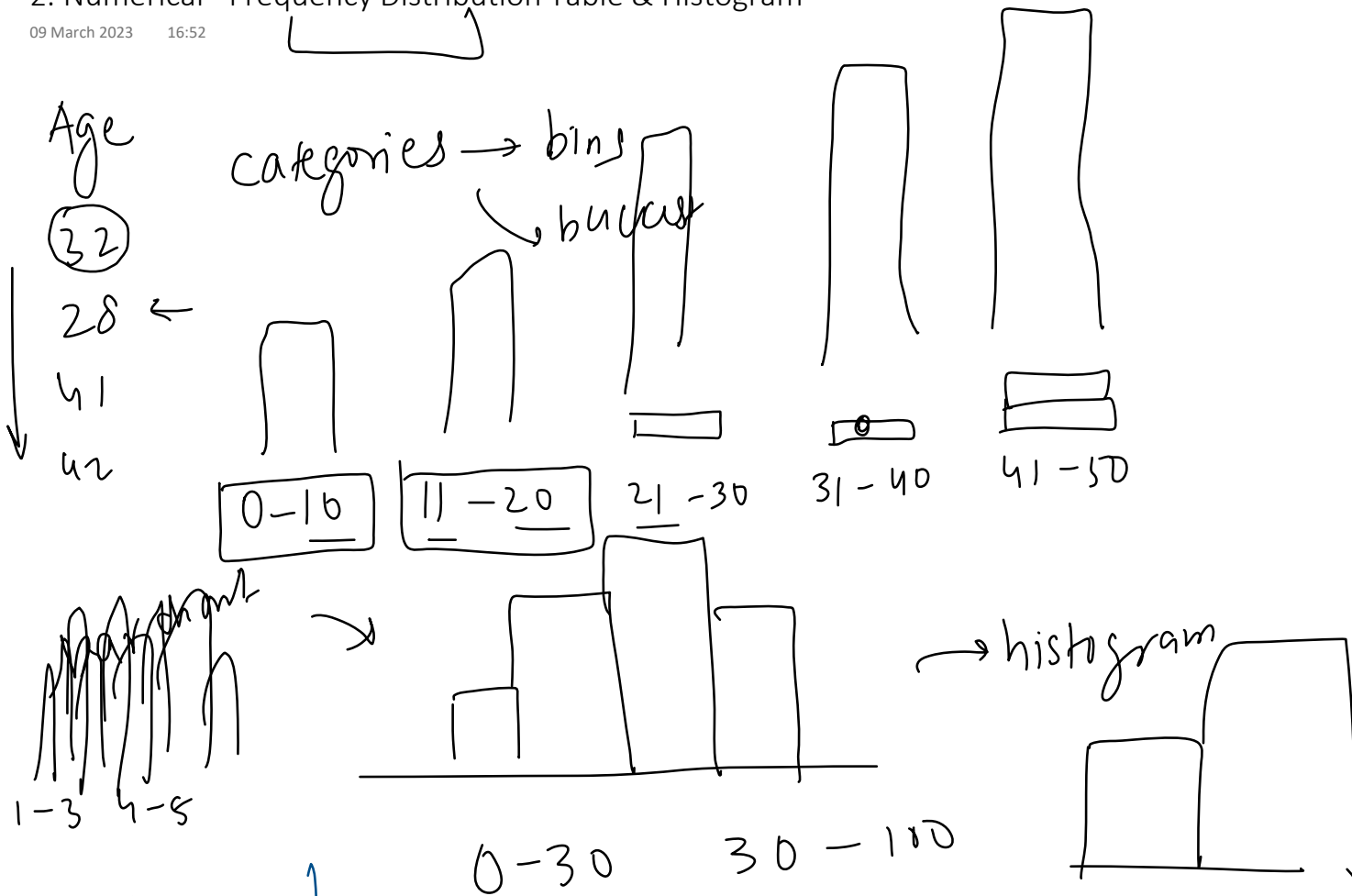
Type of Vacation	Frequency	Relative Frequency	Cumulative Frequency
Beach	60	0.3	60
City	40	0.2	100
Adventure	30	0.15	130
Nature	35	0.175	165
Cruise	20	0.1	185
Other	15	0.075	200

Handwritten annotations:

name   Prefer
Nitish   beach

bar chart

beach city
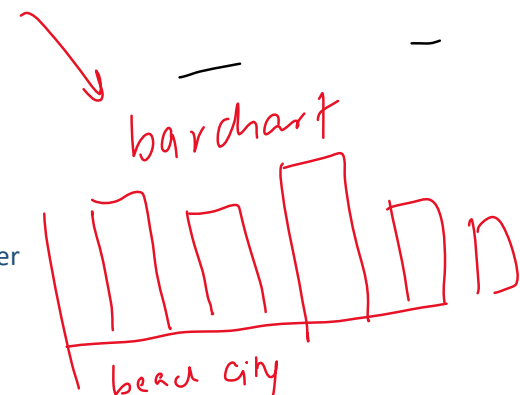
2. Numerical - Frequency Distribution Table & Histogram

Age

(32)

28 ←

41

42

categories → bins
↘ buckets

0 - 10   11 - 20   21 - 30   31 - 40   41 - 50

1 - 3   4 - 5

0 - 30   30 - 100

→ histogram

Shapes of Histogram

SYMMETRIC

BIMODAL

LEFT SKEW

RIGHT SKEW

LEFT SKEW

RIGHT SKEW

UNIFORM

NO PATTERN
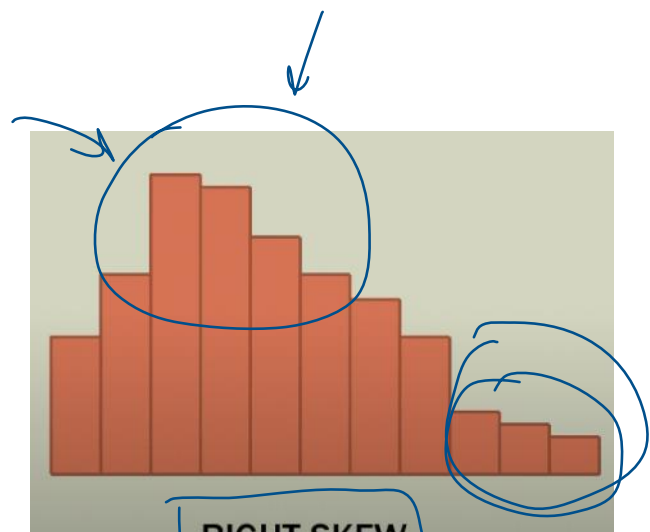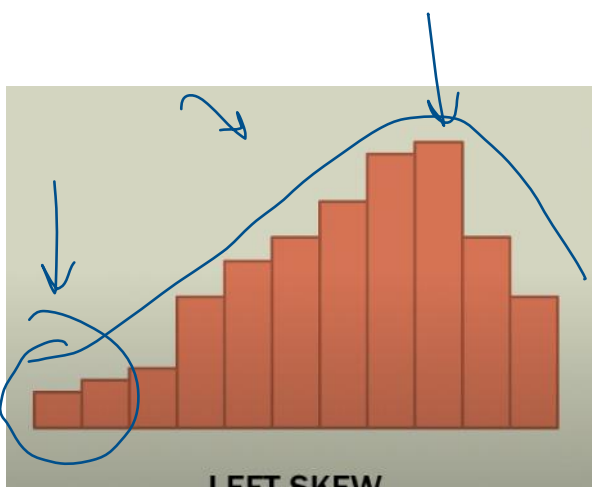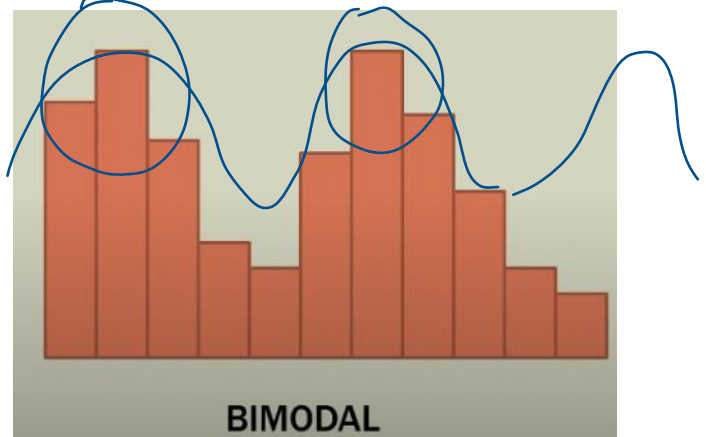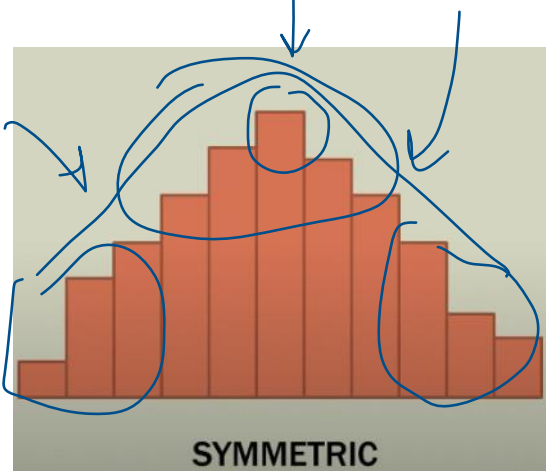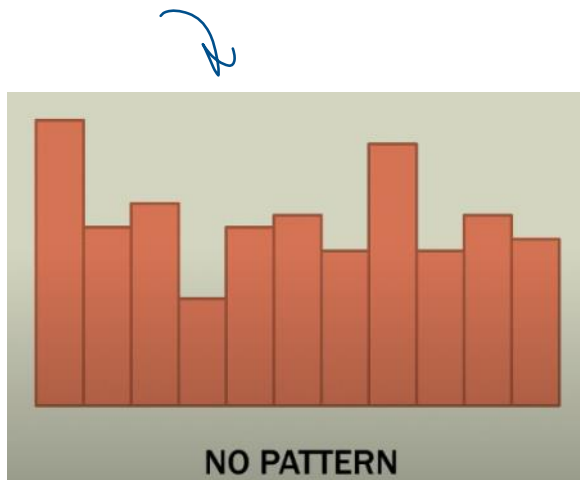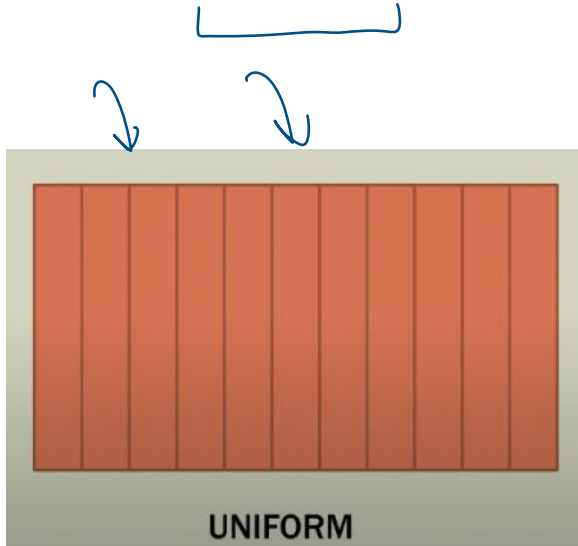
Graphs for Bivariate Analysis

Categorical
Categorical

Numerical
Numerical

Numerical
Categorical
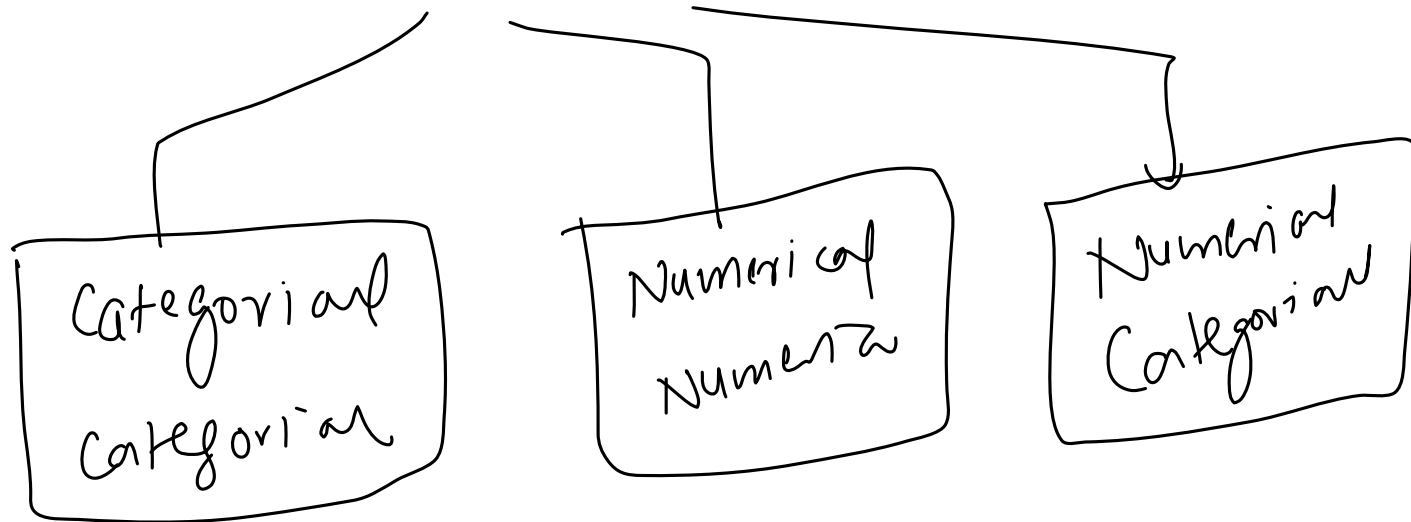
1. Categorical - Categorical

09 March 2023   16:58

Contingency Table/Crosstab

A contingency 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 contingency table displays the frequencies or relative frequencies of the observed values of the two variables, organized into rows and columns.
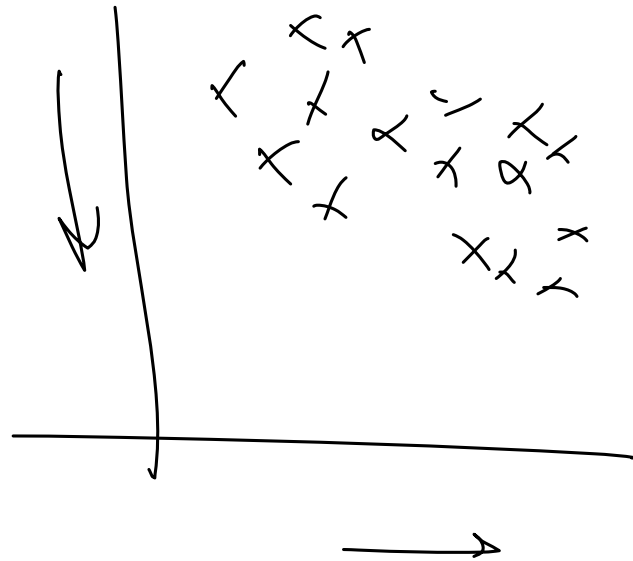
Survived
_____
0
1

P class
_____
1
2
3

→

Pclass

Survived	1	2	3
0	42	31	63
1	71	118	13

$2 \times 3 = 6$

2. Numerical - Numerical

Scatter Plot

3. Categorical - Numerical

Contingenc

	0 – 10	11 – 20	21 – 30
male	32	41	110
female	15	18	120