DESCRIPTIVE STASTICS

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 conclusions and make decisions based on the data.

In practice, statistics is used in a wide range of fields, such as business, economic, 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

Descriptive statistics

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

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

Population

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.

Sample

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

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
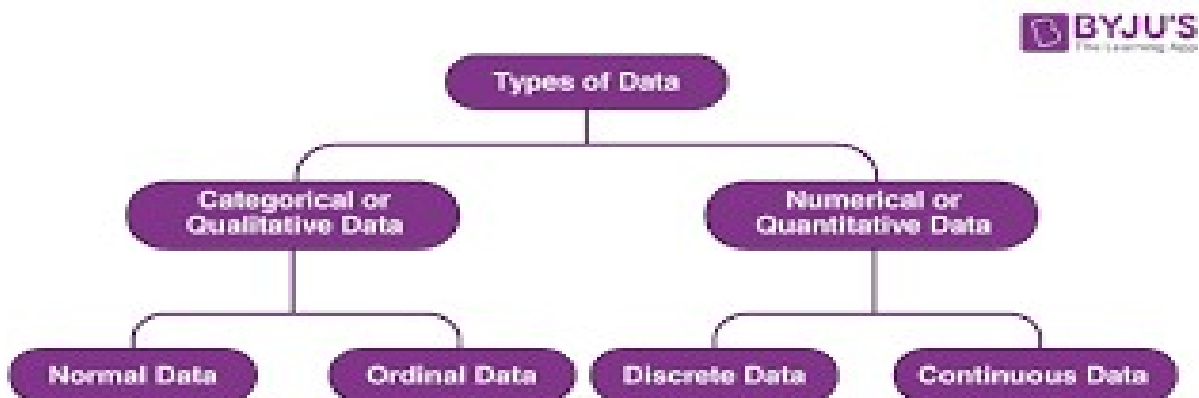
Nominal data-> male/female (albeit somewhat outdated), hair color, nationalities, names of people.

Ordinal data-> measuring economic status using the hierarchy: 'wealthy', 'middle income' or 'poor. '

Discrete data-> The number of product reviews.

Continuous data-> The amount of time required to complete a project.

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.

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

Example 3,4,1,2,5 find the mean   Population mean is 15/5=3

- Population mean $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$ where N is number of item in population
- Sample mean $\bar{X} = \frac{\sum_{i=1}^{N} x_i}{n}$ where n is number of item in population

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

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

4. 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.

Q1Find the mean of the following data set 6 6 6 6 9 9 9 9 9 7 7 7 7 7 7 7

Solution: Use the weighted mean formula weighted mean=$\frac{Sum\ of\ numbers}{Number\ of\ number} = \frac{4*6+5*9+6*7}{15} = 7.4$

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.

Q Find the trimmed mean 70,72,74,68,69,87,82,73,67,70

Ans 67,68,69,70,70,72,73,74,82,87

Q_1=n/4=10/4=2.5=3 Q_3=3n/4=3*10/4=7.5=8 q_1=69 and Q_3=74

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.

1.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: 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.

Population Variance	Sample Variance
$\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
σ^2 = population variance	s^2 = sample variance
x_i = value of ith element	x_i = value of ith element
μ = population mean	$\bar{x}$ = sample mean
N = population size	n = sample size

3.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.

Q1. Compute the standard devation of following sample: 2,3,6,6?

x	x-	
2	2-5=-3	(-3)2
3	3-5=-2	(-2)2
6	6-5=1	12
6	6-5=1	12
8	8-5=-3	32

Step-1: Compute the sample mean   Step-2: Substract the mean from each value

Step-3:Square of each deviation   Step-4: Add the devation to get the sum of square(SS).

Step-5:Divide the SS by(n-1) to get the variance (s^2)

Step-6: Take the square root of variance to get the stand deviatuion.

$\bar{x} = \dfrac{25}{5} = 5$ SS=24 $s2 = \dfrac{24}{5-1} = \dfrac{24}{4} = 6$ $s = \sqrt{6} = 2.449$

4.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%**

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

1. Categorical -Frequency Distribution Table & Cumulative Frequency

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

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

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.