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.

Population vs Sample

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 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 the college for lectures

Things to be careful about while creating samples:

1. sample size
2. random
3. representative

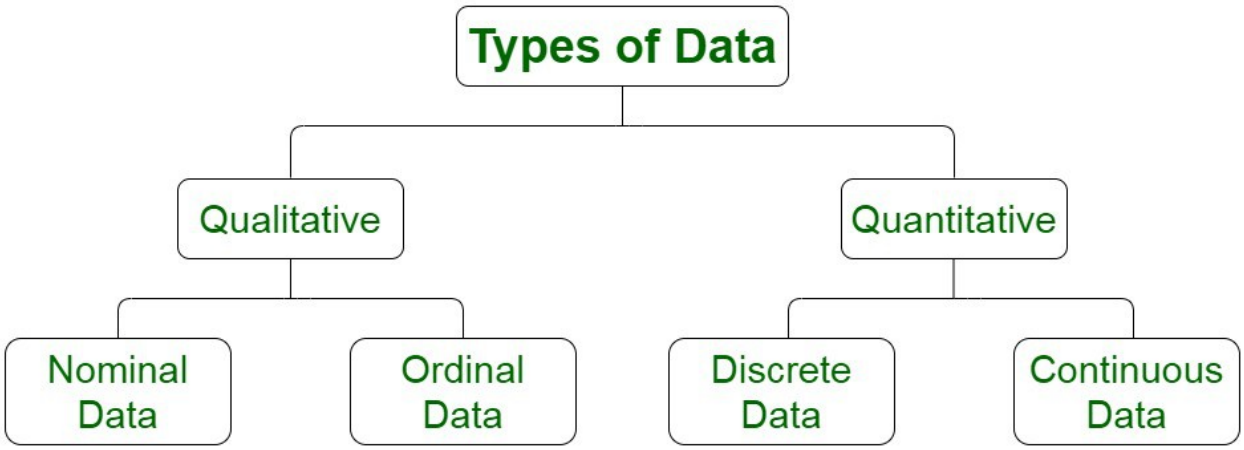
Parameter vs statistics

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.

why ml is closely associated with statistics?

Machine Learning (ML) relies on statistical principles to make predictions and decisions from data. Statistics provides the foundational tools for understanding uncertainty, estimating parameters, and evaluating the performance of ML models.

Types of Data:



Measure of Central Tendencies

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

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

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

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

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 of frequency.

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 describes the spread of variability of a dataset. It provides information about how the data is distributed around the central tendency(mean, median, mode) of the dataset.

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

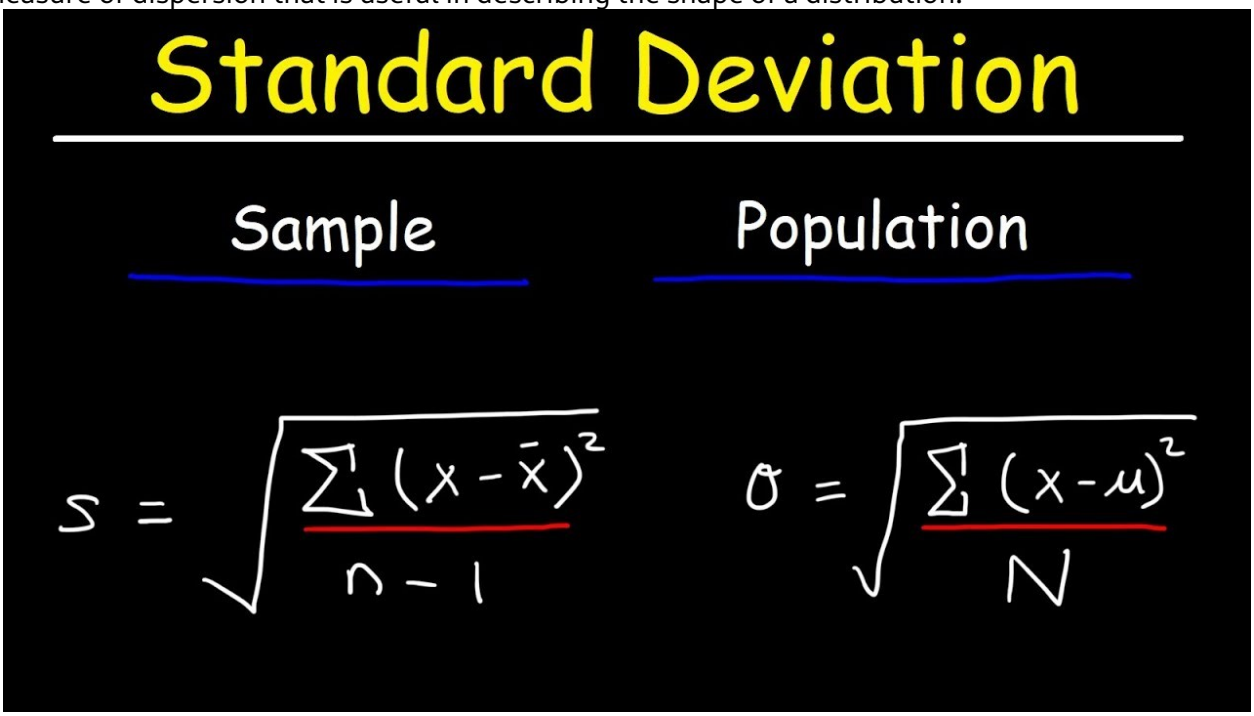
Variance: The variance is the average of the squared differences between each data point and the mean. It measures the average distance of each data point from the mean and is useful in comparing the dispersion of datasets with different means.

Population	Sample
$\sigma^2 = \frac{\Sigma(x_i-\mu)^2}{n}$	$S^2 = \frac{\Sigma(x_i-\bar{x})^2}{n-1}$
μ - Population Average xi - Individual Population Value n - Total Number of Population σ^2 - Variance of Population	x - Sample Average x_i - Individual Population Value n - Total Number of Sample S^2 - Variance of Sample

Mean Absolute Deviation:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - m|$$

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.



Coefficient of Variance: 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 variance(CV) is a statistical measure that expresses the amount of variability in a dataset relative to the mean. It is 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

1. Categorical - Frequency Distribution Table And Cumulative Frequency:

a. **Frequency distribution table** is a table that summarizes the number of times(or frequency) that each value occurs in a dataset. eg: Bar Chart.

b. **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. eg: Pie Chart.

c. **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. eg: Line chart

2. Numerical - Frequency Distribution Table and Histogram:

We plot **Histogram** by forming bins size

Graphs for Bivariate Analysis

1. Categorical-Categorical:

A **contigency table**, also known as a cross-tabulation or crosstab, is a type of table used to 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.

2. Numerical-Numerical:

We plot **scatterplot** in order to study correlation.

3. Categorical-Numerical:

We plot **bar chart** using aggregation function on numerical column.

Quantiles and Percentiles

Quantiles are statistical measures used to divide a set of numerical data into equal-sized groups, with each group containing and equal number of observations.

Quantiles are important measures of variability and can be used to: understand distribution of data, summarize and compare different datasets. They can also be used to identify outliers.

There are several types of quantiles used in statistical analysis, including:

a. **Quartiles:** Divides the data into four equal parts, Q1(25th percentile), Q2(50th percentile), Q3(75th percentile).

b. **Deciles:** Divide the data into ten equal parts, D1(10th percentile), D2(20th percentile), ..., D9(90th percentile).

c. **Percentiles:** Divide the data into 100 equal parts, P1(1st percentile), P2(2nd percentile), ..., P99(99th percentile).

d. **Quintiles:** Divides the data into 5 equal parts.

Things to remember while calculating these measure.

1. Data should be sorted from low to high.
2. You are basiscally finding the location of an observation.
3. They are not actual values in the data.
4. All other tiles can be easily derived from percentiles.

Percentile

A percentile is a statistical measure that represents the percentage of observation in a dataset that fall below a particular value for example, the 75th percentile is the value below which 75% of the observations in the dataset fall.

Formula to calculate the percentile value:

PL = P/100*(N+1)

where:

- PL = the desired percentile value between
- N = the total number of observation in the dataset
- P = the percentile rank

Example:

Find the 75th percentile score from the below data.

82, 88, 93, 94, 96, 98, 99, 84, 91, 78

Step1 - Sort the data

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

PL = 75/100*(10+1) = (3/4)*11 = 33/4 = 8.25

Value = 96 + 0.25*(98 - 96) = 96.5

75th percentile = 96.5

Percentile of a value

Percentile rank = (X + 0.5*Y)/N

X = number of values below the given value

Y = number of values equal to the given value

N = total number of values in the dataset.

Calculate the percentile rank of the value 88:

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

percentile rank = (3 + 0.5*1)/10 = 0.35 = 35%

5 number summary

The five-number summary is a descriptive statistic that provides a summary of a dataset. It consists of five values that divide the dataset into four equal parts, also known as quartiles. The five-number summary includes the following values:

1. Minimum value
2. First quartile(Q1)
3. Median(Q2)
4. Third quartile(Q3)
5. Maximum value

Boxplot

A boxplot, also known as a box-and-whisker plot, is a graphical representation of a dataset that shows the distribution of the data. The box plot displays a summary of the data, including the minimum and maximum values, the first quartile(Q1), the median(Q2), and the third quartile(Q3).

Covariance

Covariance is a statistical measure that describes the degree to which two variables are linearly related. It measures how much two variables change together, such that when one variable increases, does the other variable also increase, or does it decrease?

If the covariance between two variables is positive, it means that the variables tend to move together in the same direction. If the covariance of zero indicates that the variables are not linearly related.

Population Covariance Formula

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

Sample Covariance

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-y)}{N-1}$$

Disadvantages of using Covariance

One limitation of covariance is that it does not tell us about the strength of the relationship between two variables, since the magnitude of covariance is effected by the scale of the variables.

Covariance of a variable with itself is variance

Correlation

Correlation refers to a statistical relationship between two or more variables. Specifically, it measures the degree to which two variables are related and how they tend to change together.

Correlation is often measured using a statistical tool called the correlation coefficient, which ranges from -1 to 1. A correlation coefficient of -1 indicates a perfect negative correlation, a correlation coefficient of 0 indicates no correlation, and a correlation coefficient of 1 indicates a perfect positive correlation. Correlation does not effected by scale.

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)\ Var(Y)}}, \quad -1 \leq \rho \leq 1$$

$$R = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Correlation and Causation

The phrase 'correlation does not imply causation' means that just because two variables are correlation with each other, it does not necessarily mean that one causes the other. In other words a correlation between two variables does not necessarily imply that one variables is the reason for the other variables behaviour.

Suppose there is a positive correlation between the number of firefighters present at a fire and the amount of damage caused by the fire. One might be tempted to conclude that the presence of firefighters causes more damage. However, this correlation could be explained by a third variable - the severity of the fire. More severe fires might require more firefighters to be present, and also cause more damage.

Thus, while correlations can provide valuable insights into how different variables are related, they cannot be used to establish causality. Establishing causality often requires additional evidence such as experiments, randomized controlled trials, or well-designed observational studies.