Statistics

It is mathematical science pretaining to the collection,presentation , analysis and interpretation of data.

1. Statistics analysis: It is the science of collection , exploring and presenting large amounts of the data to identify patterns and trends.it is also called quantitative analysis.
2. Non-statistics analysis: It provides generic information and include text,sound , still images and moving images.it is also called qualitative analysis.

There are two major categories of statistics:

1.Descriptive statistics:It helps organizes data and focuses on the main characteristics of the data. It provides a summary of the data numerically or graphically.

Ex: mean,median,mode,standard deivation

Types of Descriptive Statistics

There are three types of descriptive statistics:

- Measures of Central Tendency
- Measures of Dispersion
- Measures of Frequency Distribution

Measures of Central Tendency

The central tendency is defined as a statistical measure that may be used to describe a complete distribution or dataset with a single value, known as a measure of central tendency. Any of the

central tendency measures accurately describes the whole data distribution. In the following sections, we will look at the central tendency measures, their formulae, applications, and kinds in depth.

- Mean
- Median
- Mode

Mean

Mean is the sum of all the components in a group or collection divided by the number of items in that group or collection. Mean of a data collection is typically represented as x̄ (pronounced “x bar”). The formula for calculating the mean for ungrouped data to express it as the measure is given as follows:

For a series of observations:

$$\bar{x} = \Sigma x / n$$

where,

- x̄ = Mean Value of Provided Dataset
- Σx = Sum of All Terms
- n = Number of Terms

Median

Median of a data set is the value of the middle-most observation obtained after organizing the data in ascending order, which is one of the measures of central tendency. Median formula may be used to compute the median for many types of data, such as grouped and ungrouped data.

Ungrouped Data Median (n is odd): [(n + 1)/2]th term

Ungrouped Data Median (n is even): [(n / 2)th term + ((n / 2) + 1)th term]/2

where,

- n = Number of Terms

Mode

Mode is one of the measures of central tendency, defined as the value that appears the most frequently in the provided data, i.e. the observation with the highest frequency is known as the mode of data.

Mode of Ungrouped Data: Most Repeated Observation in Dataset

Measures of Dispersion

If the variability of data within an experiment must be established, absolute measures of variability should be employed. These metrics often reflect differences in a data collection in terms of the average deviations of the observations.

- Range
- Standard Deviation
- Variance

Range

The range represents the spread of your data from the lowest to the highest value in the distribution. It is the most straightforward measure of variability to compute. To get the range, subtract the data set's lowest and highest values.

Range = Highest Value – Lowest Value

Standard Deviation

Standard deviation (s or SD) represents the average level of variability in your dataset. It represents the average deviation of each score from the mean. The higher the standard deviation, the more varied the dataset is.

To calculate standard deviation, follow these six steps:

Step 1: Make a list of each score and calculate the mean.

Step 2: Calculate deviation from the mean, by subtracting the mean from each score.

Step 3: Square each of these differences.

Step 4: Sum up all squared variances.

Step 5: Divide the total of squared variances by N-1.

Step 6: Find the square root of the number that you discovered.

Variance

Variance is calculated as average of squared departures from the mean. Variance measures the degree of dispersion in a data collection. The more scattered the data, the larger the variance in

relation to the mean. To calculate the variance, square the standard deviation.

Symbol for variance is ***s2***

Mean Deviation

Mean Deviation is used to find the average of the absolute value of the data about the mean, median, or mode. Mean Deviation is some times also known as absolute deviation. The formula mean deviation is given as follows:

Mean Deviation = ∑n1 \|X – μ\|/n

where,

- **μ** is Central Value

Quartile Deviation

Quartile Deviation is the Half of difference between the third and first quartile. The formula for quartile deviation is given as follows:

Quartile Deviation = (Q3 – Q1)/2

where,

- **Q3** is Third Quartile
- **Q1** is First Quartile

Other measures of dispersion include the relative measures also known as the coefficients of dispersion.

Measures of Frequency Distribution

Datasets consist of various scores or values. Statisticians employ graphs and tables to summarize the occurrence of each possible value of a variable, often presented in percentages or numerical figures.

Univariate Descriptive Statistics

Univariate descriptive statistics focus on one thing at a time. We look at each thing individually and use different ways to understand it better. Programs like SPSS and Excel can help us with this.

If we only look at the average (mean) of something, like how much people earn, it might not give us the true picture, especially if some people earn a lot more or less than others. Instead, we can also look at other things like the middle value (median) or the one that appears most often (mode). And to understand how spread out the values are, we use things like standard deviation and variance along with the range.

Bivariate Descriptive Statistics

When we have information about more than one thing, we can use bivariate or multivariate descriptive statistics to see if they are related. Bivariate analysis compares two things to see if they change together. Before doing any more complicated tests, it’s important to look at how the two things compare in the middle.

Multivariate analysis is similar to bivariate analysis, but it looks at more than two things at once, which helps us understand relationships even better.

2.Inferential statistics: generalizer the larger dataset and applies probability theory to draw a conclusion.it allows you to infer population parameters based on sample statistics and to model relationships with the data.

1. Hypothesis testing: It is testing is an inferential statistical technique to determine whether there is enough evidence in a data sample to infer that a certan condition holds true for the entire population.

Hypothesis testing is defined in two terms –

- **Null Hypothesis** being the sample statistic to be equal to the population statistic.
- **Alternate Hypothesis** for this example would be that the marks after extra class are significantly different from that before the class.

2. T-tests

· T-tests are very much similar to the z-scores, the only difference being that instead of the Population Standard Deviation, we now use the Sample Standard Deviation. The rest is same as before, calculating probabilities on basis of t-values.

· The Sample Standard Deviation is given as:

3. ANOVA

ANOVA (Analysis of Variance) is used to check if at least one of two or more groups have statistically different means.

To perform an ANOVA, you must have a continuous response variable and at least one categorical factor with two or more levels. ANOVA requires data from approximately normally distributed populations with equal variances between factor levels

4. Chi-square

Chi-square test is used when we have one single categorical variable from the population.

Probability

It is the measure of likelihood that an event will occur. Probability is quantified as a number between 0 and 1. 0 indicates impossibility and 1 indicates certainly.

- **Experiment** – are the uncertain situations, which could have multiple outcomes. Whether it rains on a daily basis is an experiment.

- **Outcome** is the result of a single trial. So, if it rains today, the outcome of today's trial from the experiment is "It rained"
- **Event** is one or more outcome from an experiment. "It rained" is one of the possible event for this experiment.
- **Probability** is a measure of how likely an event is. So, if it is 60% chance that it will rain tomorrow, the probability of Outcome "it rained" for tomorrow is 0.6

Cumulative Probability Distribution

The cumulative probability distribution is also known as a continuous probability distribution. In this distribution, the set of possible outcomes can take on values in a continuous range.

Discrete Probability Distribution

A distribution is called a discrete probability distribution, where the set of outcomes are discrete in nature.

frequency distribution

The **frequency** of a value is the number of times it occurs in a dataset. A **frequency distribution** is the pattern of frequencies of a variable. It's the number of times each possible value of a variable occurs in a dataset.

Types of frequency distributions

There are four types of frequency distributions:

- **Ungrouped frequency distributions:** The number of observations of each **value** of a variable.
    - You can use this type of frequency distribution for categorical variables.
- **Grouped frequency distributions:** The number of observations of each **class interval** of a variable. Class intervals are ordered groupings of a variable's values.
    - You can use this type of frequency distribution for quantitative variables.
- **Relative frequency distributions:** The proportion of observations of each value or class interval of a variable.
    - You can use this type of frequency distribution for **any type of variable** when you're more interested in **comparing frequencies** than the actual number of observations.

- **Cumulative frequency distributions:** The sum of the frequencies less than or equal to each value or class interval of a variable.