Introduction to Statistics

Statistics is a type of mathematical analysis that employs quantified models and representations to analyse a set of experimental data or real-world studies. The main benefit of statistics is that information is presented in an easy-to-understand format. Data processing is the most important aspect of any Data Science plan. When we speak about gaining insights from data, we're basically talking about exploring the chances. In Data Science, these possibilities are referred to as Statistical Analysis.

Importance of Statistics

1) Using various statistical tests, determine the relevance of features.

2) To avoid the risk of duplicate features, find the relationship between features.

3) Putting the features into the proper format.

4) Data normalization and scaling This step also entails determining the distribution of data as well as the nature of data.

5) Taking the data for further processing and making the necessary modifications.

2. Statistics and its types

The Wikipedia definition of Statistics states that "it is a discipline that concerns the collection, organization, analysis, interpretation, and presentation of data."
It means, as part of statistical analysis, we collect, organize, and draw meaningful insights from the data either through visualizations or mathematical explanations.

Statistics is broadly categorized into two types:

1. Descriptive Statistics
2. Inferential Statistics

Descriptive Statistics:

As the name suggests in Descriptive statistics, we describe the data using the Mean, Standard deviation, Charts, or Probability distributions.

Basically, as part of descriptive Statistics, we measure the following:

1. Frequency: no. of times a data point occurs
2. Central tendency: the centrality of the data – mean, median, and mode
3. Dispersion: the spread of the data – range, variance, and standard deviation
4. The measure of position: percentiles and quantile ranks

Inferential Statistics:

In Inferential statistics, we estimate the population parameters. Or we run Hypothesis testing to assess the assumptions made about the population parameters.

In simple terms, we interpret the meaning of the descriptive statistics by inferring them to the population.

For example, we are conducting a survey on the number of two-wheelers in a city. Assume the city has a total population of 5L people. So, we take a sample of 1000 people as it is impossible to run an analysis on entire population data.

From the survey conducted, it is found that 800 people out of 1000 (800 out of 1000 is 80%) are two-wheelers. So, we can infer these results to the population and conclude that 4L people out of the 5L population are two-wheelers.

3. Data Types and Level of Measurement

At a higher level, data is categorized into two types: **Qualitative** and **Quantitative**. Qualitative data is non-numerical. Some of the examples are eye colour, car brand, city, etc.
On the other hand, Quantitative data is numerical, and it is again divided into Continuous and Discrete data.

Continuous data: It can be represented in decimal format. Examples are height, weight, time, distance, etc.

Discrete data: It cannot be represented in decimal format. Examples are the number of laptops, number of students in a class.

Discrete data is again divided into Categorical and Count Data.

Categorical data: represent the type of data that can be divided into groups. Examples are age, sex, etc.

Count data: This data contains non-negative integers. Example: number of children a couple has.

Level of Measurement

In statistics, the level of measurement is a classification that describes the relationship between the values of a variable.

We have four fundamental levels of measurement. They are:

1. Nominal Scale
2. Ordinal Scale
3. Interval Scale
4. Ratio Scale

1. Nominal Scale: This scale contains the least information since the data have names/labels only. It can be used for classification. We cannot perform mathematical operations on nominal data because there is no numerical value to the options (numbers associated with the names can only be used as tags).

Example: Which country do you belong to? India, Japan, Korea.

2. Ordinal Scale: In comparison to the nominal scale, the ordinal scale has more information because along with the labels, it has order/direction.

Example: Income level – High income, medium income, low income.

3. Interval Scale: It is a numerical scale. The Interval scale has more information than the nominal, ordinal scales. Along with the order, we know the difference between the two variables (interval indicates the distance between two entities).

Mean, median, and mode can be used to describe the data.

Example: Temperature, income, et

4. Ratio Scale: The ratio scale has the most information about the data. Unlike the other three scales, the ratio scale can accommodate a true zero point. The ratio scale is simply said to be the combination of Nominal, Ordinal, and Intercal scales.

Example: Current weight, height, etc.

Central Tendency in Statistics

1) Mean: The mean (or average) is that the most generally used and well-known measure of central tendency. It will be used with both discrete and continuous data, though it’s most typically used with continuous data (see our styles of Variable guide for data types). The mean is adequate the sum of all the values within the data set divided by the number of values within the data set. So, if we have n values in a data set and they have values x1,x2, …,xn, the sample mean, usually denoted by **“x bar”**, is:

2) Median: The median value of a dataset is the value in the middle of the dataset when it is arranged in ascending or descending order. When the dataset has an even number of values, the median value can be calculated by taking the mean of the middle two values.
The following image gives an example for finding the median for odd and even numbers of samples in the dataset.

3) Mode: The mode is the value that appears the most frequently in your data set. The mode is the highest bar in a bar chart. A multimodal distribution exists when the data contains multiple values that are tied for the most frequently occurring. If no value repeats, the data does not have a mode.

4) Skewness: Skewness is a metric for symmetry, or more specifically, the lack of it. If a distribution, or data collection, looks the same to the left and sight of the centre point, it Is said to be symmetric

Measures of Dispersion

Variability in Statistics

Range: In statistics, the range is the smallest of all dispersion measures. It is the difference between the distribution’s two extreme conclusions. In other words, the range is the difference between the distribution’s maximum and minimum observations.

Range = Xmax – Xmin

Where Xmax represents the largest observation and Xmin represents the smallest observation of the variable values.

Percentiles, Quartiles and Interquartile Range (IQR)

· **Percentiles** — It is a statistician's unit of measurement that indicates the value below which a given percentage of observations in a group of observations fall.

For instance, the value QX represents the 40th percentile of XX (0.40)

· **Quantiles**— Values that divide the number of data points into four more or less equal parts, or quarters. Quantiles are the 0th, 25th, 50th, 75th, and 100th percentile values or the 0th, 25th, 50th, 75th, and 100th percentile values.

Interquartile Range (IQR)— The difference between the third and first quartiles is defined by the interquartile range. The partitioned values that divide the entire series into four equal parts are known as quartiles. So, there are three quartiles. The first quartile, known as the lower quartile, is denoted by Q1, the second quartile by Q2, and the third quartile by Q3, known as the upper quartile. As a result, the interquartile range equals the upper quartile minus the lower quartile.

IQR = Upper Quartile – Lower Quartile

= Q3 − Q1

· **Variance**: The dispersion of a data collection is measured by variance. It is defined technically as the average of squared deviations from the mean.

Standard Deviation: The standard deviation is a measure of data dispersion WITHIN a single sample selected from the study population. The square root of the variance is used to compute it. It simply indicates how distant the individual values in a sample are from the mean. To put it another way, how dispersed is the data from the sample? As a result, it is a sample statistic.

Relationship Between Variables

· **Causality**: The term "causation" refers to a relationship between two events in which one is influenced by the other. There is causality in statistics when the value of one event, or variable, grows or decreases as a result of other events.

Each of the events we just observed may be thought of as a variable, and as the number of hours worked grows, so does the amount of money earned. On the other hand, if you work fewer hours, you will earn less money.

· **Covariance**: Covariance is a measure of the relationship between two random variables in mathematics and statistics. The statistic assesses how much – and how far – the variables change in tandem. To put it another way, it's a measure of the variance between two variables. The metric, on the other hand, does not consider the interdependence of factors. Any positive or negative value can be used for the variance.

The following is how the values are interpreted:

· **Positive covariance:** When two variables move in the same direction, this is called positive covariance.

· **Negative covariance** indicates that two variables are moving in opposite directions.

Correlation: Correlation is a statistical method for determining whether or not two quantitative or categorical variables are related. To put it another way, it’s a measure of how things are connected. Correlation analysis is the study of how variables are connected.

Ø Here are a few examples of data with a high correlation:

1) Your calorie consumption and weight.

2) The amount of time you spend studying and your grade point average

Ø Here are some examples of data with poor (or no) correlation:

1) The expense of vehicle washes and the time it takes to get a Coke at the station.
2) The crime rate and house price

Correlations are useful because they allow you to forecast future behaviour by determining what relationship variables exist. In the social sciences, such as government and healthcare, knowing what the future holds is critical. Budgets and company plans are also based on these facts.

Probability

In a Random Experiment, the probability is a measure of the likelihood that an event will occur. The number of favorable outcomes in an experiment with n outcomes is denoted by x. The following is the formula for calculating the probability of an event.

Probability (Event) = Favourable Outcomes/Total Outcomes = x/n