Chapter 1: Introduction to Statistics

Statistics is the branch of Applied Mathematics which consists about the study of collection, organization, analysis, interpretation, and presentation of data. With the help of Statistics, we are able to understand the data and get insights of what is happening.

Speaking about data, there are two types of data:

1. **Qualitative data (Categorical data):** This type of data cannot be counted, measured or easily expressed in numbers. In general, it is the descriptive and conceptual findings collected through questionnaires, interviews, or observation. For example, the color of the cars, observational notes written by a doctor, open-ended surveys, etc. There are two types of Qualitative data:

a) **Nominal:** These are the data which can be labeled or classified into mutually exclusive categories within a variable. These categories cannot be ordered in a meaningful way. For example, the commutes of transportation like car, bike, bus, train, or plane.

b) **Ordinal:** These are the type of data in which the values follow a natural order. In here, we won't be able to determine the differences between the data and the value or we can say there is no meaning. For examples, income levels like less than 50k, 50k – 100k and so on, educational levels like high school, bachelor degree, master degree and so on.

2. **Quantitative data (Numerical data):** This type of data can be counted or measured in numerical values. Hence, the information here can be quantified. There are two types:

a) **Discrete:** This type of data can only take certain numerical values. It involves integers or whole numbers, rather than fractions. For example, the number of students in a classroom is a discrete data which can be counted as whole number. We cannot say as 29.5 students.

b) **Continuous:** this data takes any value and can be given with any fractions. For example, height of a person can be 5.4 feet, the score of a student is 6.7.

Sample v/s Population

1. The population is the entire group that you want to draw conclusions about.
2. The sample is the specific group of individuals that you will collect data form.

Sampling Methods

1. **Probability sampling methods:** In this method, every member of the population has a chance of being selected.

- ***Simple random sampling*:** Every member has equal chances of being selected. Hence, here the members will be selected based on chance randomly. It is kind of lottery system.

- ***Systematic Sampling*****:** It is similar to simple random but it is a bit easier to conduct. The members in the population will be based on the specific intervals. For example, multiple of 2, 3 or 4 or beginning from a number and selecting at a specific interval.
- ***Stratified sampling*****:** This sampling method involves in creating sub-population from the population. We are able to derive precise conclusion from subgroups which are able to on the basis of the data. Each sub-group is called strata.
- ***Cluster sampling*****:** In this method also, we are dividing the population into sub-population but based on the similar characteristics.

2. **Non-probability sampling method:** In this method, individuals are selected based on the non-random criteria. Here, every individual does not have the equal chances of getting selected.

- ***Convenience sampling*****:** It includes the individuals or member who happens to be most accessible to the researcher who is conducting the research.
- ***Quota sampling*****:** It is the sampling method where a researcher involves the individuals that represent the population. They choose based on the specific traits or qualities. For example, which age group prefers to smoke the cigarette?
- ***Purposive sampling*****:** Based on the purpose of the research, the researchers select the sample as per the requirement.
- ***Snowball sampling*****:** Snowball sampling method is used when the participants are not available. Hence, we recruit different participants who in turn recruit more samples.

Types of statistics

Measures of Central Tendency

Measures of central tendency are also known as measures of central location. We derive a single value which describes the data by identifying the central position within that data set. They are also categorized into summary statistics. There are three measures of central tendency:

1. ***Mean*****:** We also say it as average. It can be used on both discrete and continuous data. Based on the formula below, we calculate mean:

$$\text{Mean} = \frac{\text{Sum of all data points}}{\text{Number of all data points}}$$

2. ***Median*****:** The middle value of the data set is the median. First, we have to arrange the data in ascending order and then based on the calculation below, we find median value:

if n is odd, then

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{observation}$$

if n is even, then

$$\text{Median} = \frac{(n/2)^{th} \text{ observation} + (n/2)+1^{th} \text{ observation}}{2}$$

Here n is the number of data points.

3. ***Mode*****:** The most frequent occurring data point is mode. Hence, we can just count the number of occurrences for each repeating data point and then check the highest value of it.

Measures of Variability

Measure of variability is a summary statistic which represents the amount of dispersion in a dataset. From these measures, we can understand how spread out are the values. A low dispersion values indicates that the data points are tend to be clustered at the center. A high dispersion signifies that they tend to fall further away.

1. ***Range*****:** It is the difference between highest value and lowest value.
2. ***Inter-quartile Range*****:** It is the difference between Quartile 3 and Quartile 1.
3. ***Variance*****:** it is the average squared difference of the values of means. On other words, the sum of squared difference between the data points and mean which is divided by the number of observations.
4. ***Standard Deviation*****:** It is typically the difference between the data points and the mean. When the values in a dataset are grouped closer together, you have a smaller standard deviation. On the other hand, when the values are spread out more, the standard deviation is larger because the standard distance is greater.