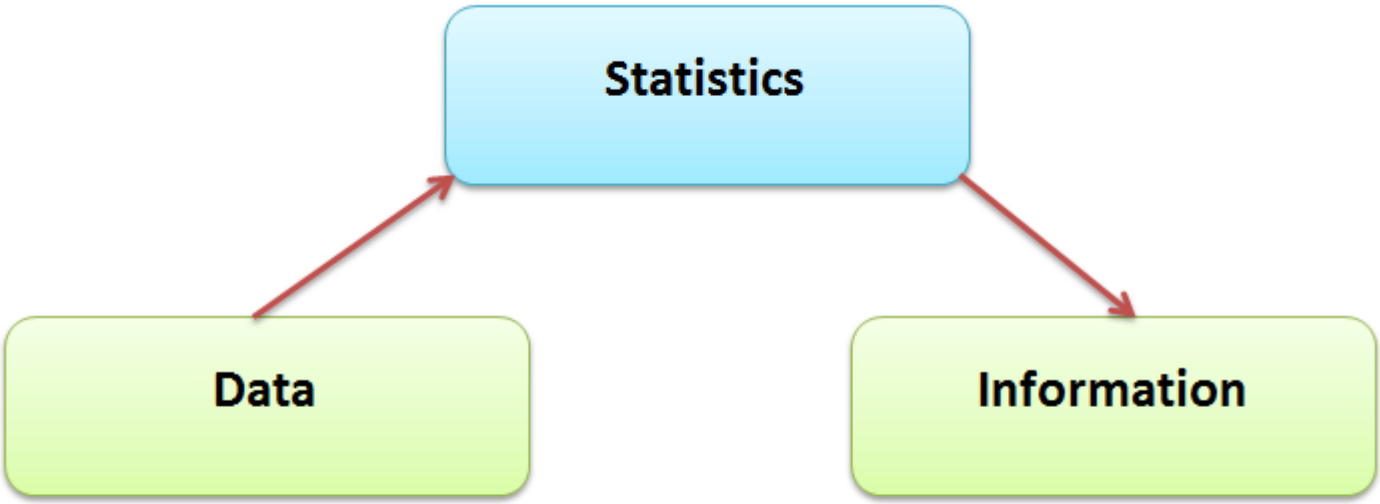
UNIT 108 – UPSC - Descriptive Statistics - Tabular, Graphical and Numerical Methods

Statistics is a branch of math that is used to analyse, interpret, and predict outcomes from data. Descriptive statistics explain the basic concepts used to describe data. This is major field for scholars interested in Data Science, Economics, Psychology, Machine Learning, Sports analytics and just about any other field. Descriptive statistics is basically the analysis of data that helps to describe, show or summarize data in a meaningful way. Descriptive statistics do not allow Statisticians to make conclusions beyond the data they have analysed or reach conclusions regarding any hypotheses they might have made. These are just a way to describe data.

Concept of Descriptive Statistics

These are numbers that are used to summarize and define data. The word "data" denotes to the information that has been collected from an experiment, a survey, a historical record. For example, if statistician evaluate birth certificates, a descriptive statistic might be the percentage of certificates issued in country, or the average age of the mother. Several descriptive statistics are often used at one time to give a full picture of the data. Descriptive statistics are just descriptive. They do not involve generalizing beyond the data in hand.

In general, Descriptive statistics deals with the description or simple analysis of population or sample data. Most simple procedure to analyse data is to classify individual observations into two or more categories according to some attribute that they possess such as people can be classified into employed/unemployed.



A variable is a measurable characteristic which changes from one member of a sample or a population to another, for example, age of a person, GDP of a country.

A continuous variable is a measurable characteristic which potentially can take any value in a continuous range, without any breaks or jumps.

A discrete variable is a measurable characteristic which is restricted to a specific set of values. Discrete data are often represented by bar charts, showing Continuous data are usually represented as histograms, showing frequency density per unit of the variable.

Descriptive statistics has great relevance in maths because simply presenting raw data cannot be understood by people. Descriptive statistics enables people to present the data in a more expressive way, which allows simpler interpretation of the data.

Descriptive statistics makes use of graphical techniques and numerical descriptive measures to summarize and present the data.

Tabular Methods

Tabular method of data presentation is wide spread in all spheres of human life. These methods are used to summarize data from a sample or population into table format. Data is grouped into categories and the number (or frequency) of observations in each category is obtained.
Frequency distribution is a type of tabular method. A frequency distribution is a tabular summary of data showing the frequency of items in each of several non-overlapping classes. The objective is to provide insights about the data that cannot be quickly obtained by looking only at the original data.

Graphical Methods

These methods are applied to visually describe data from a sample or population. The shape of sample data can indicate the shape of the population from which it is taken. Graphs provide visual summaries of data which is more quickly and completely describe essential information than tables of numbers.

Graphs are essential as these provide insight for the analyst into the data under scrutiny, and illustrate important concepts when presenting the results to others. A graphical method is developed which signifies the accuracy of test results. The graphs can be constructed from Producer's scores and Consumer's scores on each of the scales of test score, antigen dose and probability of protection against disease.

There are many types of graphical representation:

1. The Bar Chart: To Construct a Bar Chart, place categories on the horizontal axis, then place frequency (or relative frequency) on the vertical axis. After that construct vertical bars of equal width, one for each category. Its height is proportional to the frequency (or relative frequency) of the category.
2. The Pie Chart: For drawing pie chart, make complete circle that represents the total number of measurements. Partition into slices - one for each category. Then, the size of a slice is proportional to the relative frequency of that category. Determine the angle of each slice by multiplying the relative frequency by 360 degree.

Graphical Methods for Quantitative Data include Stem-and-leaf plot and Histogram.

1. Stem-and-leaf plot: Steps for Constructing Stem and Leaf Display are as follows:
    i. Break up each data into two pieces: Stem and leaf. To do this, select one or more leading digits of the data for the stem. The trailing digit or digits become leaves.
    ii. List possible stem values in a (vertical) column, with the smallest stem on top.
    iii. Record the leaf corresponding to each stem beside it in a row.
    iv. Indicate the units for stems and leaves somewhere in the display.
    v. In general, we want the number of stems to be between 5 and 20, if possible.
2. Histogram: A histogram is a graphical representation of a frequency (or relative frequency distribution). A histogram displays the shape of the data. It is useful when it is logical to group data into numerical categories.
3. Quantile plots: These visually portray the quantiles, or percentiles (which equals to the quantiles times 100) of the distribution of sample data. Quantiles of importance such as the median are easily discerned (quantile, or cumulative frequency = 0.5). Main benefits of Quantile plots are as follows:

    i. Arbitrary categories are not required, as with histograms or S-L's.
    ii. All of the data are displayed, unlike a boxplot.
    iii. Every point has a distinct position, without overlap.
4. Boxplots: Boxplot is a very useful and brief graphical display for summarizing the distribution of a data set. Boxplots provide visual summaries of the centre of the data (the median-the centre line of the box), the variation or spread (interquartile range-the box height), the skewness (quartile skew-the relative size of box halves) and presence or

absence of unusual values ("outside" and "far outside" values). Boxplots are even more useful in comparing these attributes among several data sets.

Graphical representation of reports has numerous benefits.

1. Acceptability: Graphical report is acceptable to people who have busy schedule because it easily highlights about the theme of the report. This helps to avoid wastage of time.
2. Comparative Analysis: Information can be compared in terms of graphical representation. Such comparative analysis helps for quick understanding and attention.
3. Less cost: Information, if descriptive, involves huge time to present properly. It involves more money to print the information but graphical presentation can be made in short but catchy view to make the report understandable. It obviously involves less cost.
4. Decision Making: Business executives can view the graphs at a glance and can make decision very quickly which is hardly possible through descriptive report.
5. Logical Ideas: If tables, design and graphs are used to represent information then a logical sequence is created to clear the idea of the audience.
6. Helpful for less educated Audience: Less literate or illiterate people can understand graphical representation easily because it does not involve going through line by line of any descriptive report.
7. Less Effort and Time: To present any table, design, image or graphs require less effort and time. Furthermore, such presentation makes quick understanding of the information.
8. Less Error and Mistakes: Qualitative or informative or descriptive reports involve errors or mistakes. As graphical representations are exhibited through numerical figures, tables or graphs, it usually involves less error and mistake.
9. A complete Idea: Such representation creates clear and complete idea in the mind of audience. Reading hundred pages may not give any scope to make decision. But an instant view or looking at a glance obviously makes an impression in the mind of audience regarding the topic or subject.
10. Use in the Notice Board: Such representation can be hanged in the notice board to quickly raise the attention of employees in any organization.

Graphical representation of reports has some drawbacks also:

1. Expensive: Graphical representations of reports are costly because it involves images, colours and paints. Combination of material with human efforts makes the graphical presentation expensive.
2. More time: Graphical representation involves more time as it requires graphs and figures which are dependent to more time.
3. Errors and Mistakes: Since graphical representations are complex, there is chance of errors and mistake. This causes problems for better understanding to general people.
4. Lack of Privacy: Graphical representation makes full presentation of information which may hamper the objective to keep something secret.
5. Problems to select the appropriate method: Information can be presented through various graphical methods and ways. Which should be the suitable method is very hard to select.

6. Problem of Understanding: All people cannot understand the meaning of graphical representation because it involves various technical matters which are complex to general people.

Numerical Methods

These procedures are used to arithmetically describe data from a sample or population. The numerical measures of a sample can be used to estimate the corresponding numerical measures of the population. The numerical methods can be effectively demonstrated in cases dealing with complex problems for which analytical solutions cannot be obtained or hand calculations cannot be made.

Characteristically, there are two general types of statistic that are used to describe data: Measures of central tendency: These are ways of describing the central position of a frequency distribution for a group of data. Measures of central tendency are numbers that tend to cluster around the "middle" of a set of values. These include the mode, median, and mean.

Mean is the average value, calculated by adding all the observations and dividing by the number of observations. A drawback of the mean is that it is heavily influenced by extreme observations. The median is explained as the middle value when observations are arranged in an ascending or descending order. The median is easy to understand, and it is not greatly affected by extreme observations. It is often used in preference to the mean when extreme observations are present. The mode is described as the most common value of individually recorded observations and as the value of the variable for which the frequency density is greatest for grouped data.

Measures of spread: In this type of statistic, group of data is summarized by describing how spread out the scores are. To describe this spread, a number of statistics are available such as the range, quartiles, absolute deviation, variance and standard deviation. When descriptive statistics is used, it is useful to summarize group of data using a combination of tabulated description, graphical description such as graphs and charts and statistical commentary such as discussion of the results.

To summarize, Descriptive statistics consists of statistical procedures that are used to describe the population that are studying. The data could be collected from either a sample or a population, but the results help statistician organize and describe data. Descriptive statistics can only be used to describe the group that is being studying. That is, the results cannot be generalized to any larger group. There are three methods of descriptive statistics that include tabular, graphical and numerical methods.