Descriptive Statistics

Some Basic Definitions

1. Population - A population is the group from which data are to be collected.
2. Sample - A sample is a subset of a population.
3. Variable - A variable is a feature characteristic of any member of a population differing in quality or quantity from one member to another.
4. Quantitative variable - A variable differing in quantity is called quantitative variable, for example, the weight of a person, number of people in a car.
5. Qualitative variable - A variable differing in quality is called a qualitative variable or attribute, for example, color, the degree of damage of a car in an accident.
6. Discrete variable - A discrete variable is one which no value may be assumed between two given values, for example, number of children in a family.
7. Continuous variable - A continuous variable is one which any value may be assumed between two given values, for example, the time for 100-meter run.

Descriptive Statistics

Descriptive statistics is used to summarize data and make sense out of the raw data collected. Since the data usually represents a sample, then the descriptive statistics is a quantitative description of the sample.
The level of measurement of the data affects the type of descriptive statistics.
Nominal and ordinal type data (often termed together as categorical type data) will differ in the analysis from interval and ratio type data (often termed together as continuous type data).

Descriptive statistics for categorical data

Frequency tables are used to tabulate categorical data. A frequency table shows a matrix or table between independent variables at the top row versus a dependent variable on the left column, with the cells indicating the frequency of occurrence of possible combination of levels.

Descriptive statistics for continuous data

The central tendency and variability of the data are the two aspects of descriptive statistics used for continuous type data.

Measures of central tendency "refers to a number (statistic) that best characterizes the group as a whole. It is generally referred to as the average. The three types of averages are:

1. The MEAN (M): is the arithmetic average (sum of all score divided by the number of cases)
2. The MEDIAN (Mdn): is the midpoint of a distribution of data. Half the scores fall above and half below the median.
3. The MODE: is the single score that occurs most often in a distribution of data.

Measures of variability "refers to the spread or dispersion among a set of scores. The different statistics used are the following:

1. The RANGE: is the difference between the highest and lowest score.
2. The STANDARD DEVIATION (sd): It is related to the variability of the data and the way it is clustered around the mean (median and mode are not used here). The larger the standard deviation the wider the data is spread from the mean. The smaller the standard deviation the closer the data are grouped around the mean.
3. The VARIANCE: is the square of the standard deviation.

Graphical representation of data

Several graphical techniques exist for summarizing the data. These graphs can work alone or in conjunction with the statistics described above. Some of the well known types of graphs are the bar graphs, the line graphs, and the pie graphs.

Point Estimates and Interval Estimates

A point estimate is a single value (statistic) used to estimate a population value (parameter).

A confidence interval is a range of values within which the population parameter is expected to occur.

The factors that determine the width of a confidence interval are:
1. The sample size, n.

2. The proportion or variability in the population.
3. The desired level of confidence.

- An interval estimate states the range within which a population parameter probably lies.
- The interval within which a population parameter is expected to occur is called a confidence interval.
- The two confidence intervals that are used extensively are the 95% and the 99%.
- For a 95% confidence interval, about 95% of the similarly constructed intervals will
- contain the parameter being estimated. Also 95% of the sample means for a specified
- sample size will lie within 1.96 standard deviations of the hypothesized population mean.
- For the 99% confidence interval, 99% of the sample means for a specified sample size will lie within 2.575 standard deviations of the hypothesized population mean.

Coefficient of Correlation(r)

Pearson's coefficient of correlation is used to draw an association between two variables. It is denoted by 'r'.
The value of r ranges between +1 to -1.
The coefficient of correlation is used to identify whether there is a positive association, negative association or no association between two variables.
When the value is 0, it indicates that there is no association between two variables. When it is less than 0, it indicates a negative association,
and when the value is more than 0, then it indicates a positive association.