Understanding Basic Statistics

Statistics is the study of how to collect, organize, analyze, and iterpret numerical information from data.

Statistics is both the science of uncertanity and the technology of extracting information from data.

> Desicion making is an importaint aspect our lives. We make decision based on the information e have, our attitude, and our values. statistical methods help us examine information. Moreover, statistics can be used form making decision when we are faced with uncertainities.

Variables

Variables are characteristics of the individuals to be measured or observed. Individuals are the people or objects on which we need to study.

For example: if we want to do a study about the people who have suffered from corona virus, then the individuals in this study are all peoples, and theirs characteristics like age, height, weight, gender, ..., etc. becomes varaibles

Types of variables

There are two type of variables Qunatitative (numerical) and Qulitative (categorical) variables.

Quantitative Variables: A quantitative variable is a value or numerical measurment on which we can perform methematical operations like add, subtract, divide and multiplication.

Qualitative variable: A qualitative variable describes an individual by placing the individual into a category or group, such as make or female.

Data sets

Population

Population is a dataset which contains the every individual of interest from the selected area or region. a numerical measure that describes an aspect of the population is called **population parameter.**

Sample

Sample is a dataset which only contains a small amout of inidividual of interest from population. a numerical measure that describes an aspect of sample is called sample statistics.

Levels of measurement

we have categorized data as either qualitative or quantitative. Another way to classify data is according to one of the four levels of measurement. These levels indicate the type of arithymetic that is appropriate for the data, such as ordering, taking difference, or taking ratios.

1. **Nominal**: The nominal level ofmeasurement applies to data that consist of names, labels or categories, There are no implied criteria by which the data can be ordered from smallest to largest.
2. **Ordinal**: The ordinal level of measurement applies to data that can be arranged in order, However, differences between data values either cannot be determined or are meaningless.
3. **Interval**: The interval level of measurement applies to data that can be arranged in order. In addition, differences between data values are meaningful.
4. **Ratio**: The ratio level of measurement applies to data that can be arranged in order. In addition, both differences between data values and ratios of data values are meaningful. Data at the ratio level has a true zero.

> The level ofmeasurement tells us which arithmetic processes are appropriatye fro the data. This is important because different statistical process require various kinds of arithmetic. In some instances all we need to do is count the number of data that meet specified criteria. In such acase nominal (and higher) data would

> not be suitable. Many other statistical processes require division, so data need to be at the ratio level.

Types of Statistics

Descriptive Statistics

Descriptive statistics involves method of organizing, picturing, and summarizing information from sample or population.

Inferential Statistics involves methods of using information from a sample to draw conclusion regarding the population.

Random Samples

Simple Random Sample

A **simple random sample** of n measurements from a population is a subset of the populations selected in such a manner that every sample of size n from the population has an equal chance of being selected.

> **Important Features of Simple Random Sample**
>
> For a simple random sample
>
> - Every sample of specified size n from the population has an equal chance of being selected.
> - No researcher bias occurs in the items selected from the sample.
> - A random sample may not always reflects the diversity of the population.

Sampling Techniques

Random Sampling

Use a simple random sample from the entire population.

Stratified sampling

Divide the entire populations into distinct subgroups called strata. The strata are based on a specific characteristic such as age, income, education level, and so on. All members of a stratam share the specific characteristic. Draw random samples from each stratum.

Systematic sampling

Number all members of the populations sequentially. Then, from a starting point selected at random, include every kth member of the population in the sample.

Cluster sampling

Divide the entire population into pre-existing segments or clusters. The clusters are often geographic. Make a random selection of clusters. Include every members of each selected cluseter in the sample.

Multistage sampling

Use a variety of sampling methods to create successively smaller groups at each stage. The final sample consists of cluster.

Convenience sampling

Create a sample by using data from population members that are redily available.

Points to Sampling

- **Sampling Frame**: a sampling frame is a list of individuals from which a sample is actually selected.
- **Undercoverage**: results from omitting population members from the sample frame.

Errors during sampling

- **Sampling Error**: A sampling error is the difference between measurements from a sample and corresponding measurements from the respective population. it is caused by the fact that the sample does ot perfectly represent the population.

- **Nonsampling error:** A nonsampling error is the result of poor sample design, sloppy data collection, faulty measuring instruments, bias inquestionaires, and so on.

Organizing Data

Frequency Tables

When we have a large set of quantitative data, it's useful to organize it into smaller intervals or classes and count how many data values fall into each class. we can do that by using the frequency table.

A **Frequency table** partitions data into classes or intervals of equal width and shows how many data values are in each class. The classes or intervals are constructed so that each data values falls into exactly one class.

> **class limits & width**
>
> - The **lower class limit** is the lowest data value that can fit in a class.
> - The **upper class limit** is the highest data value that can fit in a class.
> - The **class width** is the difference between the lower class limit of one class and the lower class limit of the next class.

Histogram and Relative-Freqency Histograms

Histograms and relative-frequency histograms provide effective visual display of the data orgazinzed into frequency tables. In these graphs, we use bars to represent each class, where the width of the bar is the class width. For histograms, the height of the bar is the class frequency, wheras for relative-frequency histograms, the height of the bar is the relative frequency of that class.

Distribution Shapes

Histograms are valuable and useful tools. If the raw data came from a random sample of population values, the histogram constructed from the sample values should have a distribution shape that is reasonalby similar to that of the population. Servel terms are commonly used to describe histograms adn their associated population distribution.

- **Mound-shaped symmetric**: This term refers to a mound-shpaed histogram in which both side are (more or less) the same when the graph is folded vertically down the middle.
- **Uniform or rectangular**: These terms refer to a histogram in wwhich every class has equal frequency. From on point of view, a uniform distribution is symmetric with the added property that the bars are of the same height.
- **Skewed left or skewed right**: These terms refer to a histogram in which one tail is stretched out longer than the other. The direction of skewness is on the side of the longer tail. So, if the longer tail is on the left, we say the jhistogram is skewed to the left.
- **Bimodal**: This term refers to a histogram in which the two classes with the largest frequencies are separated by at least one class. The top two frequencies of these classes may have slightly different values. This type of situation sometimes indicates that we care sampling from two different populations.