Introductory Basic Statistics

Statistics is the art and science of collecting and analyzing data. The ultimate goal is to translate data into knowledge. Statistics makes it possible to predict the future data and able to quantify the level of uncertainty associated with the prediction.

Data are the information collected or measured from experiments and surveys. It is a set of values comprised of quantitative or qualitative or both types of information related to a group of objects or individuals.

Variable is a characteristic whose value varies from one person or object to another. Example: age, height, eye color.

Two types of variables:

(a) **Categorical variable** is a non-numerically valued variable, values describe quality or characteristics of individuals.
Categorical variables can be two types:

(i) **Nominal** variable: values are purely qualitative; do not have any order or rank. Example: hair color (blonde, brown, brunette, red, etc.).

(ii) **Ordinal** variable: values can be ordered or ranked. Example: educational experience (elementary school graduate, high school graduate, some college and college graduate), we know college graduates have more educational experience than high school or elementary school graduates.

(b) **Numerical variable** (quantitative variable) is a numerically valued variable, values are measurements and have magnitude.
Numerical variables can be two types.

(i) **Discrete** variable: values are counts, measured in whole number. Example: number of students in biometry class.

(ii) **Continuous** variable: values are real numbers within an interval. Values can be fractions, decimal points are allowed in this case. Example: temperature.

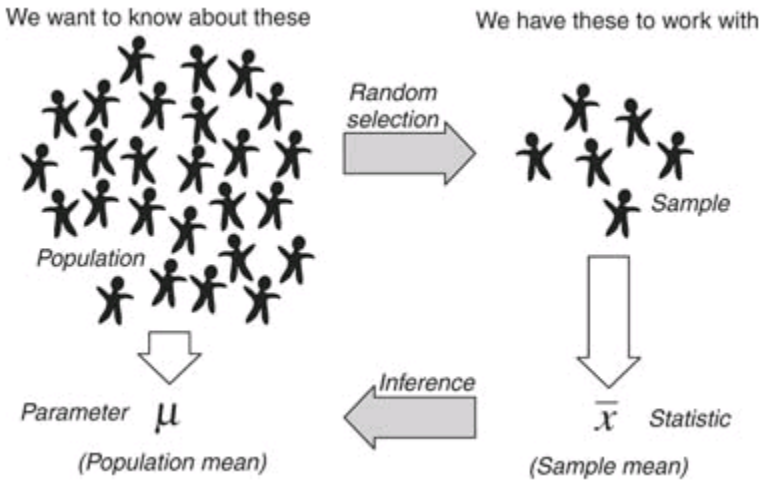
Population is **all** the individuals or items about which investigator desires to make an inference in a statistical study. The study of the entire population is known as **census**.

Sample is a part or subset of the population.

Sometimes the population size may be huge (such as the entire human population in the world), and it is difficult or impossible to get information about the entire population. So instead we choose a sample (such as portions of the entire population) to study.

A **parameter** is a numerical summary of the population. Such as: population mean, population proportion, population median etc. A population parameter is usually

represented by Greek letters. Such as population mean is **μ** (mu), population standard deviation is **σ** (sigma), population variance **σ^2** (sigma squared), etc.



A **statistic** is a numerical summary of a sample taken from the population, such as: sample mean, sample proportion, sample median etc. A sample statistic is usually represented by lower case English alphabets. Such as sample mean is $\bar{x}$, sample standard deviation is s (lower case s), sample variance s^2 , etc.

Sample survey: A study that selects a part or sample of individuals from a population and interviews them to collect data.

Sampling frame: It is the list of subjects or individuals in the population from which sample is selected.

Sampling design: It is the method of selecting subjects from the sampling frame.

Different types of sampling designs:

A simple random sample of *n* subjects from a population of size *N* is one in which each possible sample of same size has equal chance of being selected. In other words, each member of the population has an equal and independent chance of being selected.

Stratified random sample divides the population into separate groups, called strata, and then selects a simple random sample from each stratum proportional to the stratum size.

Cluster random sample divides the population into a large number of clusters, such as city blocks, hospitals, school districts etc. Then select a simple random sample of the clusters. All the subjects in the selected clusters are considered as the sample.

Systematic random sample selects randomly a starting point and then select every *k*th (chosen in such way that all *n* subjects of the sample can be selected) subject until all *n* subjects are selected.

Convenience sample selects members that are easily available to the researcher, not randomly.

Bias in sample surveys

Sampling bias occurs when sample is not representative of the population. This happens when sample is not selected randomly. Another name of this is "under-coverage" of the population.

Non response bias occurs when the subjects in the sample cannot be reached or refused to participate or fail to answer some questions.

Response bias occurs when subject gives an incorrect answer. Subject can do this deliberately (to avoid embarrassment) or unwillingly because of confusing or misleading questions.

Volunteer bias occurs when there is a systematic discrepancy between the volunteers and the population.

Two types of statistics:

Descriptive statistics consists of methods for summarizing the data by constructing:

(a) tables,
(b) charts,
(c) graphs,

and also by computing descriptive numbers such as averages, standard deviations, percentages etc. Each of these numbers is known as **sample statistic** or just **statistic**.

Inferential statistics consists of methods for making decisions or predictions about a population based on data obtained from a random sample of that population.

In this case, sample **statistic** is used to **estimate** or predict a corresponding population **parameter**. This particular type of inferential technique is known as **estimation**.

Inferential statistics is also used to validate a claim about a specific value of a parameter by comparing that value with the corresponding sample statistic. This inferential technique is called **test of hypothesis**.

Types of statistical studies

Observational study: Merely observes rather than experiments (or manipulates) with the subjects. This type of study cannot establish cause and effect relationship.

Experimental study: Assigns to each subject a treatment (experimental condition) and then observes the outcome on the response variable. This type of study can establish cause and effect relationship.

Statistical studies have two variables that are of primary interest – a **response variable** and an **explanatory variable**. Response variable is the variable that represents outcome of interest and explanatory variable is the variable that explains the response variable.

Experimental Units are the subjects – the people, animals, cells, etc. or objects to which treatments are applied and measurements are taken.

Treatments are experimental conditions imposed on the experimental units. The treatments correspond to the values of an explanatory variable. Treatment group is the group that receives the treatment (experimental drug).

Control is a type of treatment that provides a comparison group to determine whether a particular treatment is effective or not. So control could be a placebo (inert substance) or no treatment or an existing treatment. Control group is the group receiving control treatment.

Randomization: A process of randomly assigning subjects into different treatment groups. Randomization provides support for cause-effect relationship and prevents bias.

Confounding occurs when an 'outside' variable (which is not a part of the study) influences the response variable. Confounding variable may also be associated with the explanatory variable.

Double-blind study: An experimental study in which neither the subject nor the researcher (data collector) knows the subject's treatment assignment.

Example1: A US study (Muscat, 2000) compared 469 patients with brain cancer to 422 patients who did not have brain cancer in terms of cell-phone usage. Patient's cell-phone use was measured using a questionnaire. In this study, *response variable* is whether or not a subject has cancer and *explanatory variable* is the time amount of cell phone use. *Experimental unit* or subject of this is study patients. This study is an *observational study*, as no treatment was assigned to either of the groups.

From this study, researcher can get some idea about the relationship between brain cancer and cell phone use. But will **NOT** be able to establish that greater amount of cell phone use causes brain cancer. Genetic predisposition to neurological defects could act as a confounding variable and may end up influencing this study result.

Example2: An Australian study (Repacholi, 1997) used 200 transgenic mice, specially bred to be susceptible to cancers of the immune system. Half of the mice were exposed for two half-hour periods a day to the same kind of microwaves with roughly the same power as the kind transmitted from a cell phone. The other half was not exposed. After 18 months, the brain tumor rates of both groups were compared. In this study, *response variable* was whether or not a mouse developed brain tumor and *explanatory variable* was whether or not the subject was exposed to cell phone radiation. *Experimental unit* or subject of this is study is mouse. This study is an *experimental study*; each mouse was assigned one of the two treatments: receiving radiation or not receiving radiation. The group of mice not receiving radiation served as the control group in this experiment and the group receiving radiation is the treatment group.

From this study, researcher will **be able to establish** that either microwaves cause cancers of the immune system or not depending on the results of the experiment.

There are **two** main types of observational studies. These are outlined below.

Case-control study: This is a retrospective study. Subjects with a response outcome of interest, such as cancer serves as cases and other subjects not having that serve as controls. Usually cases and subjects are similar in terms of age, race, gender, education, income, etc. This type of study looks back retrospectively to compare the cases and controls to determine the relationship between an outcome and exposure to a risk factor, such as whether they had been smokers.
Example: Demonstration of the link between tobacco smoking and lung cancer (Richard Doll and Bradford Hill, 1950, *BMJ* 2).

Cohort study: This is a prospective study, planned in advance and carried out over a future period. This type of study design is commonly used in medical research to establish links between risk factors and health outcomes, to identify causes of disease. Cohort study is a type of longitudinal study that follows a group of people who share a common characteristic, over a time period.
Example: The **Framingham Heart Study**, which is a long-term, ongoing cardiovascular cohort study on the residents of the town of Framingham, Massachusetts. This study started in 1948 with 5,209 residents of Framingham, and still continuing with the 3rd generation. This study established the epidemiology of cardiovascular disease, effects of diet and exercise on heart disease, efficacy of aspirin, etc.