Introduction to Statistics:

Statistics is a branch of mathematics dealing with the collection, organization, analysis, interpretation and presentation of data. In applying statistics to, for example, a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a model process to be studied. Populations can be diverse topics such as "all people living in a country" or "every atom composing a crystal". Statistics deals with all aspects of data including the planning of data collection in terms of the design of surveys and experiments.

1) Descriptive Statistics: is a summary statistic that quantitatively describes or summarizes features of a collection of information, while descriptive statistics in the mass noun sense is the process of using and analyzing those statistics.

Descriptive statistics is distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aims to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent. This generally means that descriptive statistics, unlike inferential statistics, is not developed on the basis of probability theory, and are frequently nonparametric statistics. Even when a data analysis draws its main conclusions using inferential statistics, descriptive statistics are generally also presented.

For example, in papers reporting on human subjects, typically a table is included giving the overall sample size, sample sizes in important subgroups (e.g., for each treatment or exposure group), and demographic or clinical characteristics such as the average age, the proportion of subjects with related comorbidities, etc.

Some measures that are commonly used to describe a data set are measures of central tendency and measures of variability or dispersion. Measures of central tendency include the mean, median and mode, while measures of variability include the standard deviation (or variance), the minimum and maximum values of the variables, kurtosis and skewness.

2. Inferential Statistics: is the process of using data analysis to deduce properties of an underlying probability distribution. Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population.

Statistical inference makes propositions about a population, using data drawn from the population with some form of sampling. Given a hypothesis about a population, for which we wish to draw inferences, statistical inference consists of (first) selecting a statistical model of the process that generates the data and (second) deducing propositions from the model.

Konishi & Kitagawa state, "The majority of the problems in statistical inference can be considered to be problems related to statistical modeling". Relatedly, Sir David Cox has said, "How [the] translation from subject-matter problem to statistical model is done is often the most critical part of an analysis".

The conclusion of a statistical inference is a statistical proposition. Some common forms of statistical proposition are the following:

1) A point estimate, i.e. a particular value that best approximates some parameter of interest;

2) An interval estimate, e.g. a confidence interval (or set estimate), i.e. an interval constructed using a dataset drawn from a population so that, under repeated sampling of such datasets, such intervals would contain the true parameter value with the probability at the stated confidence level;

3) A credible interval, i.e. a set of values containing, for example, 95% of posterior belief;

4) Rejection of a hypothesis;

5) Clustering or classification of data points into groups.

Variable and attribute:

A variable is a symbol, such as $\propto, \beta, \gamma, x, or\ y$, that can assume any of a prescribed set of values, called the domain of the variable. If the variable can assume only one value, it is called a constant. A variable that can theoretically assume any value between two given values is called a continuous variable; otherwise, it is called a discrete variable.

In science and research, an attribute is a characteristic of an object (person, thing, etc.). Attributes are closely related to variables. A variable is a logical set of attributes. Variables can "vary" - for example, be high or low. How high, or how low, is determined by the value of the attribute (and in fact, an attribute could be just the word "low" or "high"). (For example see: Binary option)

While an attribute is often intuitive, the variable is the operationalized way in which the attribute is represented for further data processing. In data processing data are often represented by a combination of items (objects organized in rows), and multiple variables (organized in columns).

Values of each variable statistically "vary" (or are distributed) across the variable's domain. A domain is a set of all possible values that a variable is allowed to have. The values are ordered in a logical way and must be defined for each variable. Domains can be bigger or smaller. The smallest possible domains have those variables that can only have two values, also called binary (or dichotomous) variables. Bigger domains have non-dichotomous variables and the ones with a higher measurement. Semantically, greater precision can be obtained when considering an object's characteristics by distinguishing 'attributes' (characteristics that are attributed to an object) from 'traits' (characteristics that are inherent to the object).

1. Qualitative Data: are properties that are observed and can generally not be measured with a numerical result. They are contrasted to quantitative properties which have numerical characteristics.

Some engineering and scientific properties are qualitative. A test method can result in qualitative data about something. This can be a categorical result or a binary classification (e.g., pass/fail, go/no go, conform/non-conform). It can sometimes be an engineering judgement.

The data that all share a qualitative property form a nominal category. A variable which codes for the presence or absence of such a property is called a binary categorical variable, or equivalently a dummy variable.

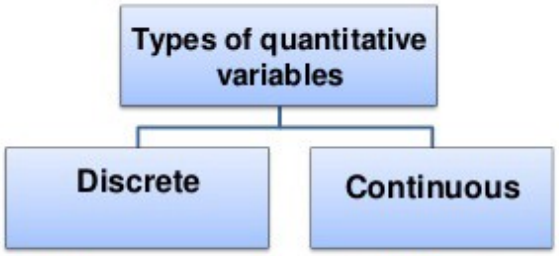
2. Quantitative Data: In natural sciences and social sciences, quantitative research is the systematic empirical investigation of observable phenomena via statistical, mathematical, or computational techniques. The objective of quantitative research is to develop and employ mathematical models, theories, and hypotheses pertaining to phenomena. The process of measurement is central to quantitative research because it provides the fundamental connection between empirical observation and mathematical expression of quantitative relationships.

Quantitative data is any data that is in numerical form such as statistics, percentages, etc. The researcher analyses the data with the help of statistics and hopes the numbers will yield an unbiased result that can be generalized to some larger population. Qualitative research, on the other hand, inquires deeply into specific experiences, with the intention of describing and exploring meaning through text, narrative, or visual-based data, by developing themes exclusive to that set of participants.

1. Continuous variable: a continuous variable is one which can take on infinitely many, uncountable values. For example, a variable over a non-empty range of the real numbers is continuous, if it can take on any value in that range. The reason is that any range of real numbers between a & b with $a \,\&\, b \in R\,; a \neq b$ is infinite and uncountable.

2.Discrete variable: In contrast, a discrete variable over a particular range of real values is one for which, for any value in the range that the variable is permitted to take on, there is a positive minimum distance to the nearest other permissible value. The

number of permitted values is either finite or countably infinite. Common examples are variables that must be integers, non-negative integers, positive integers, or only the integers 0 and 1.Methods of calculus do not readily lend themselves to problems involving discrete variables. Examples of problems involving discrete variables include integer programming.



A discrete variable

is characterized by gaps or interruptions in the values that it can assume.

For example:

- The number of daily admissions to a general hospital,
- The number of decayed, missing or filled teeth per child in an elementary school.

A continuous variable

can assume any value within a specified relevant interval of values assumed by the variable.

For example:

- Height,
- weight,
- skull circumference.

No matter how close together the observed heights of two people, we can find another person whose height falls somewhere in between.

Statistical population:

In statistics, a population is a set of similar items or events which is of interest for some question or experiment. A statistical population can be a group of existing objects (e.g. the set of all stars within the Milky Way galaxy) or a hypothetical and potentially infinite group of objects conceived as a generalization from experience (e.g. the set of all possible hands in a game of poker). A common aim of statistical analysis is to produce information about some chosen population.

In statistical inference, a subset of the population (a statistical sample) is chosen to represent the population in a statistical analysis. The ratio of the size of this statistical

sample to the size of the population is called a sampling fraction. If a sample is chosen properly, characteristics of the entire population that the sample is drawn from can be estimated from corresponding characteristics of the sample.

1.Sample: in statistics and quantitative research methodology, a data sample is a set of data collected and/or selected from a statistical population by a defined procedure. The elements of a sample are known as sample points, sampling units or observations.

Typically, the population is very large, making a census or a complete enumeration of all the values in the population either impractical or impossible. The sample usually represents a subset of manageable size. Samples are collected and statistics are calculated from the samples, so that one can make inferences or extrapolations from the sample to the population.

The data sample may be drawn from a population without replacement (i.e. no element can be selected more than once in the same sample), in which case it is a subset of a population; or with replacement (i.e. an element may appear multiple times in the one sample), in which case it is a multisubset.

2. Sampling: is the selection of a subset (a statistical sample) of individuals from within a statistical population to estimate characteristics of the whole population. Two advantages of sampling are that the cost is lower and data collection is faster than measuring the entire population.

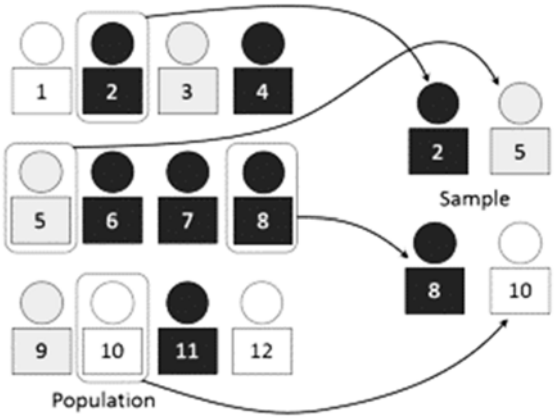
Each observation measures one or more properties (such as weight, location, and colour) of observable bodies distinguished as independent objects or individuals. In survey sampling, weights can be applied to the data to adjust for the sample design, particularly stratified sampling.

Results from probability theory and statistical theory are employed to guide the practice. In business and medical research, sampling is widely used for gathering information about a population. Acceptance sampling is used to determine if a production lot of material meets the governing specifications.

3. Sampling methods:

Within any of the types of frames identified above, a variety of sampling methods can be employed, individually or in combination. Factors commonly influencing the choice between these designs include:

a)Simple random sampling: In a simple random sample (SRS) of a given size, all such subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection: the frame is not subdivided partitioned. Furthermore, any given pair of elements has the same chance of selection as any other such pair (and similarly for triples, and so on).

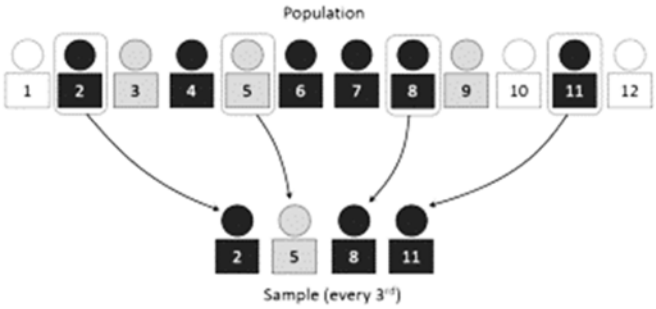


This minimizes bias and simplifies analysis of results. In particular, the variance between individual results within the sample is a good indicator of variance in the overall population, which makes it relatively easy to estimate the accuracy of results.

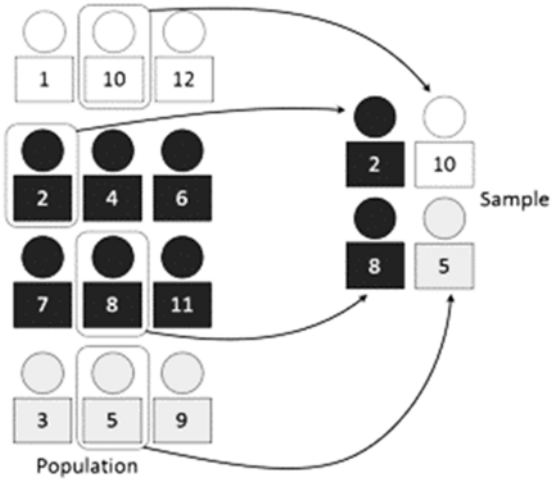
b)Systematic sampling: Systematic sampling (also known as interval sampling) relies on arranging the study population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every kth element from then onwards.

In this case, k =(population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the kth element in the list. A simple example would be to select every

10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10').



c)Stratified sampling: When the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected. The ratio of the size of this random selection (or sample) to the size of the population is called a sampling fraction. There are several potential benefits to stratified sampling.



First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample.

Second, utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). Even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less

efficiency than would simple random sampling, provided that each stratum is proportional to the group's size in the population.

Third, it is sometimes the case that data are more readily available for individual, pre-existing strata within a population than for the overall population; in such cases, using a stratified sampling approach may be more convenient than aggregating data across groups (though this may potentially be at odds with the previously noted importance of utilizing criterion-relevant strata).

Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.

d)Cluster sampling: is commonly implemented as multistage sampling. This is a complex form of cluster sampling in which two or more levels of units are embedded one in the other. The first stage consists of constructing the clusters that will be used to sample from. In the second stage, a sample of primary units is randomly selected from each cluster (rather than using all units contained in all selected clusters). In following stages, in each of those selected clusters, additional samples of units are selected, and so on. All ultimate units (individuals, for instance) selected at the last step of this procedure are then surveyed. This technique, thus, is essentially the process of taking random subsamples of preceding random samples.