8 SHOWS STATISTICS, SAMPLING IN FINITE POPULATIONS, SAMPLING ERROR, SAMPLING AND STRATIFIED SAMPLING STRATEGIES.

8.1 SHOWS STATISTICS: Statistics is the study of the collection, organization, analysis, interpretation and presentation of data.[1] It deals with all aspects of data including the planning of data collection in terms of the design of surveys and experiments.[1] When analyzing data, it is possible to use one of two statistics methodologies: descriptive statistics or inferential statistics.

Scope

Statistics is a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data,[3] or as a branch of mathematics.[4] Some consider statistics to be a distinct mathematical science rather than a branch of mathematics.[*vague*][5][6]

Mathematical statistics

Mathematical statistics is the application of mathematics to statistics, which was originally conceived as the science of the state — the collection and analysis of facts about a country: its economy, land, military, population, and so forth. Mathematical techniques which are used for this include mathematical analysis, linear algebra, stochastic analysis, differential equations, andmeasure-theoretic probability theory.[7][8]

Overview

In applying statistics to e.g. a scientific, industrial, or societal problem, it is necessary to begin with a population or process to be studied. Populations can be diverse topics such as "all persons living in a country" or "every atom composing a crystal".

Ideally, statisticians compile data about the entire population (an operation called census). This may be organized by governmental statistical institutes. **Descriptive statistics** can be used to summarize the population data. Numerical descriptors include mean and standard deviation for continuous data types (like income), while frequency and percentage are more useful in terms of describing categorical data (like race).

When a census is not feasible, a chosen subset of the population called a sample is studied. Once a sample that is representative of the population is determined, data

is collected for the sample members in an observational or experimental setting. Again, descriptive statistics can be used to summarize the sample data. However, the drawing of the sample has been subject to an element of randomness, hence the established numerical descriptors from the sample are also due to uncertainty. In order to still draw meaningful conclusions about the entire population, **inferential statistics** is needed. It uses patterns in the sample data to draw inferences about the population represented, accounting for randomness. These inferences may take the form of: answering yes/no questions about the data (hypothesis testing), estimating numerical characteristics of the data (estimation), describing associations within the data (correlation) and modeling relationships within the data (for example, using regression analysis). Inference can extend to forecasting, prediction and estimation of unobserved values either in or associated with the population being studied; it can include extrapolation and interpolation of time series or spatial data, and can also include data mining

8.2 PARAMETER OR STATISTICAL SAMPLE: A **parameter** (from the Ancient Greek παρά, "para", meaning "beside, subsidiary" and μέτρον, "metron", meaning "measure"), in its common meaning, is a characteristic, feature, or measurable factor that can help in defining a particular system. A parameter is an important element to consider in evaluation or comprehension of an event, project, or situation. *Parameter*has more specific interpretations in mathematics, logic, linguistics, environmental science,[1] and other disciplines.

In statistics and econometrics, the probability framework above still holds, but attention shifts to estimating the parameters of a distribution based on observed data, or testing hypothesesabout them. In classical estimation these parameters are considered "fixed but unknown", but in Bayesian estimation they are treated as random variables, and their uncertainty is described as a distribution.[*citation needed*]

A statistic is a numerical characteristic of a sample that can be used as an estimate of the corresponding parameter, the numerical characteristic of the population from which the sample was drawn. For example, the sample mean (usually denoted $\overline{X}$) can be used as an estimate of the *mean* parameter (μ) of the population from which the sample was drawn.

It is possible to make statistical inferences without assuming a particular parametric family of probability distributions. In that case, one speaks of *non-parametric statistics* as opposed to theparametric statistics just described. For example, a test based on Spearman's rank correlation coefficient would be called non-parametric since the statistic is computed from the rank-order of the data

disregarding their actual values (and thus regardless of the distribution they were sampled from), whereas those based on the Pearson product-moment correlation coefficient are parametric tests since it is computed directly from the data values and thus estimates the parameter known as the population correlation.

8.3 ADVANTAGES OF THE CHOICE OF A SAMPLE: In statistics, quality assurance, & survey methodology, **sampling** is concerned with the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population. Each observation measures one or more properties (such as weight, location, color) 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 practice. In business and medical research, sampling is widely used for gathering information about a population.

The sampling process comprises several stages:

- Defining the population of concern
- Specifying a sampling frame, a set of items or events possible to measure
- Specifying a sampling method for selecting items or events from the frame
- Determining the sample size
- Implementing the sampling plan
- Sampling and data collecting
- Data which can be selected

Stratified sampling

Where 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. 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.

There are, however, some potential drawbacks to using stratified sampling. First, identifying strata and implementing such an approach can increase the cost and complexity of sample selection, as well as leading to increased complexity of population estimates. Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata. Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods (although in most cases, the required sample size would be no larger than would be required for simple random sampling.

A stratified sampling approach is most effective when three conditions are met

1. Variability within strata are minimized
2. Variability between strata are maximized
3. The variables upon which the population is stratified are strongly correlated with the desired dependent variable.

Advantages over other sampling methods

1. Focuses on important subpopulations and ignores irrelevant ones.
2. Allows use of different sampling techniques for different subpopulations.
3. Improves the accuracy/efficiency of estimation.
4. Permits greater balancing of statistical power of tests of differences between strata by sampling equal numbers from strata varying widely in size.

Disadvantages

1. Requires selection of relevant stratification variables which can be difficult.

2. Is not useful when there are no homogeneous subgroups.
3. Can be expensive to implement.

8.4 SAMPLING ERROR: In statistics, **sampling error** is incurred when the statistical characteristics of a population are estimated from a subset, or sample, of that population. Since the sample does not include all members of the population, statistics on the sample, such as means and quantiles, generally differ from parameters on the entire population. For example, if one measures the height of a thousand individuals from a country of one million, the average height of the thousand is typically not the same as the average height of all one million people in the country. Since sampling is typically done to determine the characteristics of a whole population, the difference between the sample and population values is considered a *sampling error*.[1] Exact measurement of sampling error is generally not feasible since the true population values are unknown; however, sampling error can often be estimated by probabilistic modeling of the sample.

8.5 Sampling Strategies

There are four primary sampling strategies:

- Random sampling
- Stratified random sampling
- Systematic sampling
- Rational sub-grouping

Before determining which strategy will work best, the analyst must determine what type of study is being conducted. There are normally two types of studies: population and process. With a population study, the analyst is interested in estimating or describing some characteristic of the population (inferential statistics).

With a process study, the analyst is interested in predicting a process characteristic or change over time. It is important to make the distinction for proper selection of a sampling strategy. The " I Love Lucy" television show's " Candy Factory" episode can be used to illustrate the difference. For example, a population study, using samples, would seek to determine the average weight of the entire daily run of candies. A process study would seek to know whether the weight was changing over the day.

Random Sampling
Random samples are used in population sampling situations when reviewing historical or batch data. The key to random sampling is that each unit in the population has an equal probability of being selected in the sample. Using random sampling protects against bias being introduced in the sampling process, and hence, it helps in obtaining a representative sample.

In general, random samples are taken by assigning a number to each unit in the population and using a random number table or Minitab to generate the sample list. Absent knowledge about the factors for stratification for a population, a random sample is a useful first step in obtaining samples.

For example, an improvement team in a human resources department wanted an accurate estimate of what proportion of employees had completed a personal development plan and reviewed it with their managers. The team used its database to obtain a list of all associates. Each associate on the list was assigned a number. Statistical software was used to generate a list of numbers to be sampled, and an estimate was made from the sample.

Stratified Random Sampling
Like random samples, stratified random samples are used in population sampling situations when reviewing historical or batch data. Stratified random sampling is used when the population has different groups (strata) and the analyst needs to ensure that those groups are fairly represented in the sample. In stratified random sampling, independent samples are drawn from each group. The size of each sample is proportional to the relative size of the group.

For example, the manager of a lending business wanted to estimate the average cycle time for a loan application process. She knows there are three types (strata) of loans (large, medium and small). Therefore, she wanted the sample to have the same proportion of large, medium and small loans as the population. She first separated the loan population data into three groups and then pulled a random sample from each group.

Systematic Sampling
Systematic sampling is typically used in process sampling situations when data is collected in real time during process operation. Unlike population sampling, a frequency for sampling must be selected. It also can be used for a population study if care is taken that the frequency is not biased.

Systematic sampling involves taking samples according to some systematic rule – e.g., every fourth unit, the first five units every hour, etc. One danger of using

systematic sampling is that the systematic rule may match some underlying structure and bias the sample.

For example, the manager of a billing center is using systematic sampling to monitor processing rates. At random times around each hour, five consecutive bills are selected and the processing time is measured.

Rational Subgrouping

Rational subgrouping is the process of putting measurements into meaningful groups to better understand the important sources of variation. Rational subgrouping is typically used in process sampling situations when data is collected in real time during process operations. It involves grouping measurements produced under similar conditions, sometimes called short-term variation. This type of grouping assists in understanding the sources of variation between subgroups, sometimes called long-term variation.

The goal should be to minimize the chance of special causes in variation in the subgroup and maximize the chance for special causes between subgroups. Subgrouping over time is the most common approach; subgrouping can be done by other suspected sources of variation (e.g., location, customer, supplier, etc.)

For example, an equipment leasing business was trying to improve equipment turnaround time. They selected five samples per day from each of three processing centers. Each processing center was formed into a subgroup.

When using subgrouping, form subgroups with items produced under similar conditions. To ensure items in a subgroup were produced under similar conditions, select items produced close together in time.

8.6 SAMPLING NON-PROBABILITY: Sampling is the use of a subset of the population to represent the whole population. Probability sampling, or random sampling, is a sampling technique in which the probability of getting any particular sample may be calculated. **Nonprobability sampling** does not meet this criterion and should be used with caution. Nonprobability sampling techniques *cannot* be used to infer from the sample to the general population.

The advantage of nonprobability sampling is its lower cost compared to probability sampling. However, one can say much less on the basis of a nonprobability sample than on the basis of a probability sample. Of course, research practice appears to belie this claim, because many analysts draw generalizations (e.g., propose new theory, propose policy) from analyses of nonprobability sampled data. One must ask, however, whether those published works are publishable because tradition

makes them so, or because there really are justifiable grounds for drawing generalizations from studies based on non probability samples.

Some embrace the latter claim, and assert that while probability methods are suitable for large-scale studies concerned with representativeness, non-probability approaches are more suitable for in-depth qualitative research in which the focus is often to understand complex social phenomena (e.g., Marshall 1996; Small 2009). These assertions raise an interesting question—how can one understand a complex social phenomenon by drawing only the most convenient expressions of that phenomenon into consideration? What assumption about homogeneity in the world must one make to justify such assertions? Alas, research indicates only one situation in which a non-probability sample can be appropriate—if one is interested *only* in the specific cases studied (for example, if one is interested in the Battle of Gettysburg), one does not need to draw a probability sample from similar cases.