Design and analysis of sample surveys

School of Science and Informatics

Department of Mathematics, Statistics and Physical Sciences

Taita Taveta University

www.ttu.ac.ke

Dr. Noah Mutai

January 6, 2023

Contents

Terms

- Definition — an explanation of the mathematical meaning of a word.
- Theorem — A statement that has been proven to be true.
- Proposition — A less important but nonetheless interesting true statement.
- Lemma — A true statement used in proving other true statements (that is, a less important theorem that is helpful in the proof of other results).
- Corollary — A true statement that is a simple deduction from a theorem or proposition.
- Proof — The explanation of why a statement is true.
- Conjecture — A statement believed to be true, but for which we have no proof. (a statement that is being proposed to be a true statement).
- Axiom — A basic assumption about a mathematical situation (a statement we assume to be true).

1 Fundamentals

1.1 Introduction

Sample survey, finite population sampling or survey sampling is a method of drawing an inference about the characteristics of a population or universe by **observing under only a part of the population.** Such methods are extensively used by government bodies throughout the world for assessing, among others, different characteristics of national economy as a required for making decisions and for the planning and projection of future economic structure.

Ideally, total information about the population is obtained through census, where every individual in the population is involved in giving out information. However, most of the times due to certain constraints to be discussed later, it is not always possible to carry out a census.

In a sample survey the purpose of the survey statistician is to estimate some functions of the population parameter, $\theta\left(y\right)$, say, by choosing a sample(part of the population) and by observing the values of y only on units selected in the sample. The statistician therefore want to make an inference about the population by observing only a part of it. This is essential and perhaps the only practical method of inference about the characteristics of the population since in many socioeconomic investigations the survey population may be very large, containing say hundreds or thousands of units.

Definition 1.1. Survey population — A finite(survey) population is a collection of known number N of identifiable units labeled $1, 2, 3, ..., i, ..., N$ where i stands for the label as well as the physical unit labeled i. The number N is the size of the population. The parametric functions of general interest for estimation are;

1. Population total, $Y = \sum_{i=1}^{N} Y_i$
2. Population mean: $\bar{Y} = \frac{Y}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i$
3. Population variance: $S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$
4. Population coefficient of variance: $C_Y = \frac{S_Y}{\bar{Y}}$,where S_Y is the population variance and $\bar{Y}$ is the population mean.

Definition 1.2. Sample — is a part of the population/subset of the population selected for study. A sample may be drawn from a population either under with replacement(wr) or under without replacement(wor).

After a sample is selected , data are collected from the sampled units. We shall denote by y_i the value of y on the unit selected at the i^{th} draw $(i = 1, 2, ..., n)$. Thus for example if the sample is $S = \{2, 3, 2\}$,$y_1 = Y_2, y_2 = Y_3, y_3 = Y_2$ .Clearly y_i is a random variable whose possible values lie in the set $\{Y_1, Y_2, ..., Y_N\}$

For a sample s, we shall denote some statistics as follows;

1. Sample total, $y = \sum_{i=1}^{n} y_i$
2. Sample mean, $\bar{y} = \frac{y}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i$
3. Sample variance, $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$
4. Sample coefficient of variation, $c_y = \frac{s_y}{\bar{y}}$,where s_y is the sample variance and $\bar{y}$ is the sample mean.

Definition 1.3. Sampling units: This refers to the individual items whose characteristics are to be measured in the sample survey.

Definition 1.4. Sampling frame: This is the list of all sampling units. It may be a list of units with identification and particulars or a map showing the boundaries of sampling unis e.g. a manufacturing firm may want to determine how popular a newly manufactured product is within the community suggests a possible frame for the survey. The firm may decide to concentrate its surveys in urban residential areas only. In this case, you have a complete list of estates in urban areas. The residents in those chosen estates will be interviewed and inferences are made.

Definition 1.5. Sampled population: It is the set of individuals in the sampling frame. Its actually the subset of the target population. Note: Sampled population is not necessarily the same as target population.

Definition 1.6. Sampling scheme: Its the technique by which the elements which constitute the sampled are obtained from the population.

1.2 Types of Sampling

1. Haphazard sampling: No scheme has been used at all it is neither probability nor non-probability sampling.
2. Purposive or judgemental sampling or non probability sampling.
3. Probability/random sampling- statistical theory is used and the kind of inferences made are based on statistical procedures. There is some element of chance associated with selection of items into the sample.

1.3 Properties of random sampling

We are able to define the set of distinct samples, $S_1, S_2, ..., S_N$, which the procedure is capable of selecting if applied to a specific population. This means that we can say precisely what sampling units belong to S_1 to S_2 and so on.

1. Each possible sample S_i has assigned to it a known probability of selection π_i.

2. We select one of the S_i by a process in which each S_i receives its appropriate probability π_i, of being selected.

3. The method for computing the estimate from the sample must be stated and must lead to a unique estimate for any specific sample.

The simplest type of sampling is SRS(simple random sampling). We shall also make use of common terms in statistics like; statistic, estimator, point estimation and interval estimation, regression estimation. Design and analysis of sample survey is about knowing those estimators and design procedures that are good.

1.4 Properties of estimators

- Precision — how much variation there is in the estimation from sample to sample.
- Trueness — on average how close is the estimate to the population characteristics being estimated.
- Accuracy — combination of precision and trueness. Precision of estimates will be measured by variance e.g. if the estimator is X then;

$$Var\left(x\right) = \delta_x^2 = E\left[X - \mu_x\right]^2 \tag{1.1}$$

where $\mu_x = E\left[X\right]$

Trueness of an estimation will be measured by the bias, which is defined as a difference between the expectation of estimate and population parameter of which is an estimate.

$$Bias\left(x\right) = E\left[X\right] - \mu_x \tag{1.2}$$

If $E\left[X\right] - \mu_x = 0$, then it is an unbiased estimator. Accuracy is measured using mean squared error(MSE)

$$MSE\left(x\right) = \sigma_x^2 + \left(Bias\left(x\right)\right)^2 \tag{1.3}$$

Generally estimators which have low variance (high precision) and low bias are preferred.

1.5 Principal steps involved in planning and execution of a sample survey.

The broad steps to conduct any sample surveys are as follows:

1. **Objective of the survey:** The objective of the survey has to be clearly defined and well understood by the person planning to conduct it. It is expected from the statistician to be well versed with the issues to be addressed in consultation with the person who wants to get the survey conducted. In complex surveys, sometimes the objective is forgotten and data is collected on those issues which are far away from the objectives.

2. **Population to be sampled:** Based on the objectives of the survey, decide the population from which the information can be obtained. For example, population of farmers is to be sampled for an agricultural survey whereas the population of patients has to be sampled for determining the medical facilities in a hospital.

3. **Data to be collected:** It is important to decide that which data is relevant for fulfilling the objectives of the survey and to note that no essential data is omitted. Sometimes, too many questions are asked and some of their outcomes are never utilized. This lowers the quality of the responses and in turn results in lower efficiency in the statistical inferences.

4. **Degree of precision required:** The results of any sample survey are always subjected to some uncertainty. Such uncertainty can be reduced by taking larger samples or using superior instruments. This involves more cost and more time. So it is very important to decide about the required degree of precision in the data. This needs to be conveyed to the surveyor also.

5. **Method of measurement:** The choice of measuring instrument and the method to measure the data from the population needs to be specified clearly. For example, the data has to be collected through interview, questionnaire, personal visit, combination of any of these approaches, etc. The forms in which the data is to be recorded so that the data can be transferred to mechanical equipment for easily creating the data summary etc. is also needed to be prepared accordingly.

6. **The frame:** The sampling frame has to be clearly specified. The population is divided into sampling units such that the units cover the whole population and every sampling unit is tagged with identification. The list of all sampling units is called the frame. The frame must cover the whole population and the units must not overlap each other in the sense that every element in the population must belong to one and only one unit. For example, the sampling unit can be an individual member in the family or the whole family.

7. **Selection of sample:** The size of the sample needs to be specified for the given sampling plan. This helps in determining and comparing the relative cost and time of different sampling plans. The method and plan adopted for drawing a representative sample should also be detailed.

8. **The Pre-test:** It is advised to try the questionnaire and field methods on a small scale. This may reveal some troubles and problems beforehand which the surveyor may face in the field in large scale surveys.

9. **Organization of the field work:** How to conduct the survey, how to handle business administrative issues, providing proper training to surveyors, procedures, plans for handling the non-response and missing observations etc. are some of the issues which need to be addressed for organizing the survey work in the fields. The procedure for early checking of the quality of return should be prescribed. It should be clarified how to handle the situation when the respondent is not available.

10. **Summary and analysis of data:** It is to be noted that based on the objectives of the data, the suitable statistical tool is decided which can answer the relevant questions. In order to use the statistical tool, a valid data set is required and this dictates the choice of responses to be obtained for the questions in the questionnaire, e.g., the data has to be qualitative, quantitative, nominal, ordinal etc. After getting the completed questionnaire back, it needs to be edited to amend the recording errors and delete the erroneous data. The tabulating procedures, methods of estimation and tolerable amount of error in the estimation needs to be decided before the start of survey. Different methods of estimation may be available to get the answer of the same query from the same data set. So the data needs to be collected which is compatible with the chosen estimation procedure.

11. **Information gained for future surveys:** The completed surveys work as guide for improved sample surveys in future. Beside this they also supply various types of prior information required to use various statistical tools, e.g., mean, variance, nature of variability, cost involved etc. Any completed sample survey acts as a potential guide for the surveys to be conducted in the future. It is generally seen that the things always do not go in the same way in any complex survey as planned earlier. Such precautions and alerts help in avoiding the mistakes in the execution of future surveys.

12. **Pilot Survey** In planning a survey efficiently, some prior information about the population under consideration and the operational and cost aspects of of data collection will be needed. When such information is not available

1.6 Advantages of Sampling

Sample surveys have potential advantages over complete enumeration(census). They include;

1. **Reduced cost** — If data are secured from only a small fraction of the aggregate, expenditures may be expected to be smaller than if a complete census is attempted

2. **Greater speed** — For the same reason, the data can be collected and summarized more quickly with a sample than with a complete count. This may be a vital consideration when the information is urgently needed.

3. **Greater scope** — In certain types of inquiry, highly trained personnel or specialized equipment, limited in availability, must be used to obtain the data. A complete census may then be impracticable: the choice lies between obtaining the information by sampling or not at. Thus surveys which rely on sampling have more scope and flexibility as to the types of information that can be obtained.

4. **Greater accuracy** — Because personnel of higher quality can he employed and can be given intensive training, a sample may actually produce more accurate results than the kind of complete enumeration that it is feasible to take.

5. **Risk** — When a survey involves risky tests such as testing a new drug, sampling should be used.

1.7 Exercises

1. Discuss the statement: "The need to collect statistical information arises in almost every conceivable sphere of human activity."

2. Describe briefly each of the following terms:

   (a) Primary data

   (b) Secondary data

   (c) Mail inquiry

   (d) Questionnaire/schedule

   (e) Population

   (f) Census

   (g) Element

   (h) Sample

   (i) Sampling unit

(j) Sampling frame

3. Differentiate between target and sampled population. What problem arises if two populations are not same?

4. What is the primary advantage of probability sampling over the non probability sampling? Cite three situations where non probability sampling is to be preferred

5. Assume a sample survey shall be carried out to find out about how satisfied students are with their faculty.

   (a) How would you define the population?

   (b) Would you consider a census of all students or rather a sample survey? (Why?)

   (c) How would you operationalise? being satisfied with their faculty?

   (d) What is a sampling frame and how could one be obtained in the example?

   (e) How could a random sample be obtained?

   (f) How do you consider the idea of obtaining a sample from alumni?