Arthur Jarvis University, Akpabuyo.
STA 221: STATISTICAL INFERENCE II (3UNITS)

Course Outline:

- *Sampling and sampling distribution.*
- *Point and interval estimation.*
- *Principles of hypotheses testing.*
- *Tests of hypotheses concerning population means, proportions and variances of large and small samples, large and small sample cases.*
- *Goodness -fit tests. Analysis of variance*

1. Sampling

Introduction Have you ever tasted a hot soup and decided whether the soup was tasty or not? If yes, then you are a sampler. Sampling is a part of our day-to-day life, which we use either advertently or inadvertently. Another example is a pathologist who takes a few drops of blood and tests for any abnormality in the blood of the whole body. The process of using information obtained from the smaller quantity to make statement about the larger quantity is called sampling. In this lecture, we shall examine why this process is sometimes necessary and the various techniques for doing it. We shall first learn some fundamental concepts, which are related to sampling.

Sampling Definition
Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed, but may include simple random sampling or systematic sampling.

Sampling and Non-Sampling Errors
Any statistical inference based on sample results may not always be correct. This is because sample results are either based on partial or incomplete analysis of the population characteristics. This error is referred to as the sampling error because each sample taken may produce a different estimate of the population characteristic compared to those results that would have been obtained by a complete enumeration of the population.

Non-sampling Errors arise during census as well as sampling surveys because of biases and mistakes such as:
1. Faulty Planning
2. Non-response;
3. Non-random selection of samples;
4. Incompleteness and inaccuracy of returns;
5. Compilation errors.

Sampling Methods
Several sampling methods are available, which are classified into two categories:
1. Random Sampling Methods
2. Non-Random Sampling Methods

Random Sampling Methods
A statistic is a measurable characteristic of a sample, such as a mean or standard deviation. A sampling method is a procedure for selecting sample elements from a population. A random number is a number determined totally by chance, with no predictable relationship to any other number. The following are example of random sampling methods:

- Simple random sampling
- Stratified sampling
- Cluster sampling
- Systematic sampling
- Multi-stage sampling

Simple Random Sampling (SRS)
In this type of sampling, each unit of the population has equal chance of being included in the sample. If the units are drawn one by one in such a way that a unit drawn at a time is replaced back into the population before the subsequent draw, it is known as simple random sampling with replacement (SRSWR). However, if the unit selected once is not included in the population at any subsequent draw, it is called simple random sampling without replacement (SRSWOR). One disadvantage of this method is that all members of the population have to be available for selection. However, this availability may not be possible in most cases.

Stratified Sampling
This method is useful when the population as a whole consists of a number of heterogeneous groups while the units within each group are relatively homogeneous. Thus, population is divided into distinct groups called strata. A simple random sample is drawn from each stratum or group, in proportion to its size. Individual stratum samples are combined into one to obtain an overall sample for analysis.

Cluster Sampling
This method is also referred to as area sampling method. It is useful when the population consists of a very large number of similar groups which are wide-spreading. It is devised to meet the problem of costs or inadequate sampling frame (a complete listing of all members in the population so that each member can be identified by a distinct number). By the use of map references, the entire area to be analyzed is divided into smaller areas and a sample of the desired number of areas is selected by a SRS method. Such groups are termed as clusters. The members of the clusters are called elementary units. From each cluster, we may select a random sample of the desired size.

Systematic Sampling
This method is useful when units in the population are physically arranged in some sequence, and every k^{th} unit is included in the sample after the first has been randomly selected. The value k is called the sampling interval. A systematic sample has the advantage of being quick and easy to use. However, we need to be careful less there be a cyclic variation in the frame, and this is picked up in the sample because it is organized cyclically.

Multi stage Sampling
In multistage sampling, the whole population is divided into a number of primary units called stages, each of which is composed of second stage units. A series of samples are then taken at successive stages. The sample size at each stage is determined by the relative population size at each stage.

Non-Random Sampling Methods
It results in a biased sample, a non-random sample of a population (or non-human factors) in which all individuals, or instances, were not equally likely to have been selected. If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling.

- Quota sampling
- Judgment or Purposive sampling
- Convenience sampling

Quota Sampling
This method is often used in market and social surveys. The selection of respondents lies with the investigator's discretion; although, care must be taken to ensure that each respondent satisfies certain criteria, which are essential for the study. Because quota sampling is non-random, it leads to substantial complications in the statistical analysis of the survey results.

Judgment or Purposive Sampling
In this method, the investigator selects units of the sample that he/she feels are most representative of the population with respect to the population characteristics under study. Great precaution is needed in drawing conclusions based on judgment samples to make inferences about a population.

Convenience Sampling
Perhaps this is the easiest method for collecting data on a particular issue. The investigator simply selects units to be included in the sample at his/her convenience, rather than following a pre specified rule. Precautions are also needed in interpreting the results of convenience sampling that are used to make inferences about a population.

Some Basic Concepts of Sampling
The following are some of the basic concept of sampling

- Census
- Sample

Census
A census involves a complete count (or a complete enumeration) of every individual member of the population of interest, such as persons in a country, households in a town, shops in a city, students in a college, and so on. Apart from the cost and the large amount of resources (such as enumerators, clerical assistance, etc.) that are required, the main problem is the time required to process the data. Thus, the results are not known immediately.

Population
In statistical sense, population is a group of items, units or subjects, which is under reference of study. It is often referred to as universe by a number of statisticians and scientists. The inhabitants of a region, number of cars in a city, workers in a factory, students in a university, insects in a field, etc., are few examples of populations. Generally, populations or universe is classified into four categories:
Finite population - the number of items or units is fixed, limited and countable, e.g. workers in a factory.
Infinite population- the number of items or units is uncountable, e.g. stars in the sky.
Real population - the items or units in the population are all physically present or visible.
Hypothetical population - the population results from repeated trials, e.g. the tossing of a coin repeatedly results into a hypothetical population of heads and tails, rolling of a die again and again gives rise to a hypothetical population of numbers from 1 to 6, etc.

Sample
A sample is a part or fraction of a population selected on some basis. In principle, a sample should be such that it is a true representative of the population. The process of selecting a sample from the population is called sampling, and the manner or scheme through which the required number of units is selected is called the sampling method. The foremost purpose of sampling is to gather maximum information about the population under consideration at minimum cost, time and resources. Precisely, sampling is inevitable in the following situations:

- When population is infinite
- When the item or unit is destroyed under investigation;
- When the results are required in a short time

- When resources are limited particularly in respect of money and trained persons
- When population is either constantly changing or in a state of movement;
- When the items or units are scattered.

2. Sampling Distribution of a Statistics

Introduction

In the previous study session, population and sampling was discussed. Imagine that you have a large population to study and the description of its characteristics is not possible by census method. Then, in order to make statistical inference, samples of given size are drawn repeatedly from the population and 'statistic' computed for each sample. The computed value of a particular statistic will differ from sample to sample. This implies that, if the same statistic is computed for each of the samples, the value is likely to vary from sample to sample.

Sampling Unit

Elementary units or group of such units which besides being clearly defined, identifiable and observable, are convenient for purpose of sampling are called sampling units. For instance, in a family budget enquiry, usually a family is considered as the sampling unit since it is found to be convenient for sampling and for ascertaining the required information. In a crop survey, a farm or a group of farms owned or operated by a household may be considered as the sampling unit. Population The collection of all units of a specified type in a given region at a particular point or period of time is termed as a population or universe. Thus, we may consider a population of persons, families, farms, cattle in a region or a population of trees or birds in a forest or a population of fish in a tank etc. depending on the nature of data required.

Population Distribution

The population distribution is the distribution of values of its members and has mean denoted by μ, variance σ^2 and standard deviation σ. For example, a population consisting of the numbers 0, 2, 4 and 6 has mean $\mu = 3$ and standard deviation $\sigma = \sqrt{5}$.

Definition [All possible samples of size n]

Let a population consist of N elements.

1. If a random sample of size n is selected from the population with replacement, then there are N^n possible samples of size n that can be drawn from the population.
2. If a random sample of size n is selected from the population without replacement, then there are $({}^{N}C_{n}) = \frac{N!}{n!(N-n)!}$ possible samples of size n that can be drawn from the population.

Example (Do it)

A population consists of the numbers 0, 2, 4 and 6, List all possible samples of size 2 that can be drawn

1. with replacement
2. without replacement.

Sampling Distribution of a sample statistic.

If a particular statistic (e.g. sample mean, sample standard deviation, etc.) is computed for each of the possible samples, the value of the statistic will differ from sample to sample. Thus, it would be theoretically possible to construct a frequency table showing the values assumed by the statistic and their frequency of occurrence. This distribution of values of a statistic is called a sampling distribution. Thus, we see that there would be an overall mean (where it is centered), a standard deviation (representing the spread) and a shape if the histogram is plotted. So, we can talk of the mean of sampling distribution of a statistic (denoted μ m if *m* is the statistic), and standard deviation of sampling distribution of a

statistic (denoted σ_m if m is the statistic). These properties help lay down rules for making statistical inferences about a population on the basis of a single sample drawn from it, that is, without even repeating the sampling process.

Statistic, Estimator and Estimate

Suppose a sample of n units is selected from a population of N units according to some probability scheme and let the sample observations be denoted by $y_1, y_2, \ldots, y_n$. Any function of these values which is free from unknown population parameters is called a statistic. An estimator is a statistic obtained by a specified procedure for estimating a population parameter. The estimator is a random variable and its value differs from sample to sample and the samples are selected with specified probabilities. The particular value, which the estimator takes for a given sample, is known as an estimate.

Standard Error of Statistic

The standard deviation of sampling distribution of a statistic is called the standard error of the statistic. It is clearly different from the population standard deviation (σ). The population standard deviation describes the variation among values of members of the population, whereas the standard deviation of sampling distribution measures the variability among values of the statistic due to sampling error. Standard error is a measure of a reasonable difference between a particular sample statistic and the population parameter. It is used in tests of whether a particular sample could have been drawn from a given parent population. It is also used in working out confidence limits and confidence intervals.

Difference between standard error and standard deviation

The standard deviation, or SD, measures the amount of variability or dispersion for a subject set of data from the mean, while the standard error of the mean, or SEM, measures how far the sample mean of the data is likely to be from the true population mean. The SEM is always smaller than the SD. When dealing with numerical data sets, many people get confused between the standard deviation of the sample and the standard error of the sample mean. We want to stress the difference between these.

Standard deviation (SD)

This describes the spread of values in the sample. The sample standard deviation, s, is a random quantity -- it varies from sample to sample -- but it stays the same on average when the sample size increases.

Standard error of the mean (SE)

This is the standard deviation of the sample mean, , and describes its accuracy as an estimate of the population mean, . When the sample size increases, the estimator is based on more information and becomes more accurate, so its standard error decreases.