LESSON SIX: SAMPLING AND SAMPLING DISTRIBUTIONS

6.1 Introduction

- The field of inferential or inductive statistics is concerned with studying facts about populations. Specifically, the interest is in learning about the population parameters. This is accomplished by picking a sample and computing the values of the appropriate statistics.
- A **parameter** is a numerical descriptive measure of a population. Because it is based on the observation in the population, its value is almost always unknown.
- A **Sample statistic** is a numerical descriptive measure of a sample. It is calculated from the observations in the sample.

  NB: The term **statistic** refers to **sample quantity** and the term **parameter** refers to a **population quantity**.
- **Sampling** is the process of selecting a sample from a population.

6.2 Types of sampling Designs

There are two major ways of selecting samples;

a) Probability sampling methods
b) Non - Probability sampling methods

a) Probability sampling methods

i) Simple random sampling

- Assumes that every member of the population has an equal chance of being independently selected. All members of the population are labeled with a number and random numbers should be used to select the sample.
- This is the best method of sampling as independence of sample members is assumed by many statistical tests. Unfortunately all members of the population have to be available for selection and this is rarely the case.

ii) Systematic sampling

- It is useful when the whole sampling frame is not available. The population is listed and every *n*th member is included in the sample after the first has been selected randomly.
- Sampling from a production line may make use of this method.

iii) Stratified random sampling

- Useful when the population consists of a number of distinct subpopulations and there is no difference between the subpopulations than within each of them.
- The population is split into these differing groups – strata. A random sub-sample is then drawn from each, in proportion to the strata size.

iv) Cluster Sampling:

The population is divided into internally heterogeneous subgroups and some are randomly selected for further study. It is used when it is not possible to obtain a sampling frame because the population is either very large or scattered over a large geographical area.

b) Non-probability sampling

It is used when a researcher is not interested in selecting a sample that is representative of the population.

i) Purposive Sampling

It allows the researcher to use cases that have the required information with respect to the objectives of his or her study e.g. educational level, age group, religious sect etc.

ii) Quota Sampling

The researcher purposively selects subjects to fit the quotas identified e.g. Gender: Male or Female; Class Level: Graduate or Undergraduate; Religion: Muslim, Protestant, catholic, Jewish; Social economic class: Upper, middle or lower.

iii) Snow ball sampling

It is used when the population that possesses the characteristics under study is not well known and can be best located through referral networks. Initial subjects are identified who in turn identify others. Commonly used in drug cultures, teenage gang activities, Mungiki sect, insider trading, Mau Mau etc.

iv) Convenience or Accidental Sampling

Involves selecting cases or units of observation as they become available to the researcher e.g. asking a question to the radio listeners, roommates or neighbours.

6.3 Reasons for Sampling

We obtain a sample rather than a complete enumeration (a census) of the population for many reasons. There are six main reasons for sampling in lieu of the census.

i) **Economy**: Directly observing only a portion of the population requires fewer resources than a census.

ii) **The Time factor:** A sample may provide an investigator with needed information quickly

iii) **The very large populations:** Many populations about which inferences must be made are quite large and sample evidence may be the only way to obtain information.

iv) **Partly inaccessible populations:** Some populations contain elementary units so difficult to observe that they are in a sense inaccessible e.g. in determining consumer attitudes not all of the users of a product can be queried.

v) **The Destructive nature of the observation:** Sometimes the very act of observing the desired characteristics of the elementary unit destroys it for the use intended. Classical examples of this occur in quality control

vi) **Accuracy and sampling:** A sample may be more accurate than a census. A sloppily conducted census can provide less reliable information than a carefully obtained sample.

6.4 Bias and Error in sampling

A sample is expected to mirror the population from which it comes from. However, there is no guarantee that any sample will be precisely representative of the population. One of the things that make a sample unrepresentative of its population is the sampling error.

Sampling error: It comprises the difference between the sample and the population that are due solely to the particular elementary units that happen to have been selected.

There are two basic causes for sampling error.

- ✓ **One is Chance**: Bad luck may result in untypical choices. Unusual elementary units do exists, and there is always a possibility that an abnormally large number of them will be chosen. The main protection against this type of error is to use a large enough sample.
- ✓ Another cause of sampling error is **sampling bias.** This is the tendency to favor the selection of elementary units that have particular characteristics. Sampling bias is usually the result of a poor sampling plan.

Non sampling error

- ➢ The other main cause of unrepresentative samples is non sampling error. This type can occur whether a census or a sample is being used.

- A non-sampling error is an error that results solely from the manner in which the observations are made. The simplest example of non sampling error is inaccurate physical measurement due to faulty instruments or poor procedures. Consider the observation of human weights – no 2 answers will be of equal reliability.

6.5 Sampling Distributions

- By sampling distribution of a statistic we mean the theoretical probability distribution of the statistic.

6.6.1 Sampling Distribution of the Mean

- If samples of size *n* are drawn with replacement from a population with mean μ and variance δ^2 ,the mean and variance of the sampling distribution of $\bar{x}$ are given by $\mu_{\bar{x}} = \mu$ **and** $\delta_{\bar{x}}^2 = \frac{\delta^2}{n}$**.**
- When random samples of size *n* are drawn without replacement from a finite population of size *N* that has a mean μ and a variance δ^2 , the mean and the variance of the sampling distribution of $\bar{x}$ are given by

$$\mu_{\bar{x}} = \mu \qquad \textbf{and} \qquad \delta_{\bar{x}}^2 = \frac{\delta^2}{n} \cdot \frac{N-n}{N-1}$$

  - If the population size is large compared to the sample size, $\delta_{\bar{x}}^2 = \frac{\delta^2}{n}$, approximately
- The <u>standard deviation</u> of the sampling distribution of $\bar{x}$ is commonly known as the <u>standard error of the mean.</u> It is $\frac{\delta}{\sqrt{n}}$ when sampling with replacement. For a sample drawn without replacement from a finite population of size *N*, the standard error of the mean is $\frac{\delta}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$
- In the latter case it is approximately $\frac{\delta}{\sqrt{n}}$ if the population is very large compared to the sample size. In our discussion, we shall assume that the population is large enough that $\frac{\delta^2}{n}$ can be taken as the value of $\delta_{\bar{x}}^2$ even when sampling without replacement.
- The **standard error** of the **mean** then depends on two quantities, δ^2 and *n*. It will be large if δ^2 is large, i.e. if the scatter in the parent population is large. On the other hand, the standard error

will be small if the **sample size *n* is large**. Since with a larger sample we can get more information about the population mean μ and consequently less scatter of the sample mean about μ.

- The variance of the parent population is usually not under the experimenter's control. Therefore one sure way of reducing the standard error of the mean is by picking a large sample – the larger the better.

- So for we have concerned ourselves with two parameters of the sampling distribution of $\bar{x} \cdot \left(\mu_{\bar{x}} \text{ and } \delta_{\bar{x}}^2\right)$. We now turn our attention to the distribution itself
- The probability distribution of $\bar{x}$ will very much depend on the distribution of the sampled population.
- **Note that if *n* the sample size, is large, the distribution of $\bar{x}$ is close to a normal distribution** of course with mean μ and variance $\frac{\delta^2}{n}$. The statement of this result is contained in the central limit theorem.

Central Limit Theorem

The distribution of the sample mean $\bar{x}$ of a random sample drawn from practically any population with mean μ and variance δ^2 can be approximated by means of a normal distribution with mean μ and variance $\frac{\delta^2}{n}$, provided the sample size is large.

- The central limit theorem tells us that the shape of the distribution is approximately normal. We already know that if the population has mean μ and variance δ^2, then $\mu_{\bar{x}} = \mu$ and $\delta_{\bar{x}}^2 = \frac{\delta^2}{n}$.
- Converting to the *z* scale, we can give an alternate version of the central limit theorem.

When the sample size is large, the distribution of $\frac{\bar{x} - \mu}{\delta / \sqrt{n}}$ is close to that of a standard normal variable *z*.

(Recall that to convert to the *z* scale the rule is: subtract the mean and divide by the standard deviation of the r.v in question)

- Since the central limit theorem applies if the sample size is large, a natural question is, how large is large enough?

  This will depend on the **nature** of the **sampled population**
- **If the parent population is normally distributed, then the distribution of $\bar{x}$ is normal for any sample size,**
- If the parent population has a symmetric distribution, the approximation to the normal distribution will be reached for a moderately small sample size, as low as 10.
- In most instances, the tendency towards normality is so strong that the approximation is fairly satisfactory with a sample size of about 30.

Example 1

The records of the Dept of health, education and welfare show that the mean expenditure incurred by a student during 2010 was $5000 and the standard deviation of the expenditure was $800. Find the approximate probability that the mean expenditure of 64 students picked at random was

a) More than $4820
b) Between $4800 and $5120

Example 2

The length of life (in hours) of a certain type of electric bulb is a random variable with a mean life of 500 hours and a standard deviation of 35 hours.

What is the approximate probability that a random sample of 49 bulbs will have a mean life between 488 and 505 hours?

6.6.2 Sampling Distribution of the Proportion

- If n items are picked independently from a population where the probability of success is p (not very close to 0 or 1) and if n is large, then the distribution of the sample proportion x/n is approximately normal with **mean p and variance** pq/n where $p+q=1$.

- Converting to the *z* scale, it follows that $\frac{\frac{x}{n} - p}{\sqrt{pq/n}}$ has a distribution that is very close to the standard normal distribution provided *n* is large. This leads to the conclusion that $\frac{x - np}{\sqrt{npq}}$ is distributed approximately as a standard normal variable.

Example 1

Suppose 10% of the tubes produced by a Machine are defective. If a sample of 100 tubes is inspected at random

a) Find the expected proportion of defectives in the sample
b) Find the variance of the proportion of defective in the sample
c) Find the approximate distribution of the sample proportion
d) Find the probability that the proportion of defective will exceed 0.16

Example 2

If 60% of the population feels that the president is doing a satisfactory job, find the approximate probability that in a sample of 900 people interviewed at random, the proportion who share this view will

a) Exceed 0.65
b) Be less than 0.56