**CHAPTER-ONE**

**Sampling and Sampling Distributions**

**Objectives:**

After completing this chapter, you should be able to;

* Understand sampling and sampling distributions.
* Identify different types of sampling techniques.
* Compute sampling distributions of the means and proportions.

**1.1 SAM PLING THEORY**

* + 1. **Basic Definitions**

***Definition:***

* Sampling is a technique that is used to select a sample out of a population. It is a process of gathering information from part of the population.

***Common terminologies of sampling***

* ***Population (universe)*** – it is a collection of items or individuals chosen for a study. The characteristic of a population is known as ***parameter***.
* ***Sample*** – it is a subset of a population. It is some representative group of the studied population. The characteristic of a sample is known as ***statistic***.
* ***Census*** – it is a complete enumeration or measurement of every individual or item in the population. It is gathering information from all elements of a population.
* ***Sampling error*** – it is the difference between the population parameter and the observed probability sample statistic.
* ***Non-sampling error*** – it is an error that occurs in the collection, recording and computation of data.
* ***Sampling with replacement*** – a sampling procedure in which sample items are returned to the population; as a result, there is a possibility of their being chosen again in the sample.
* ***Sampling without replacement*** – a sampling procedure in which sample items are not returned to the population; as a result, none of these can be selected in the sample again.
  + 1. **The need for samples**

The purpose of inferential statistics is to find out something about the population based on a sample. In the next section we will discuss about the major reasons for sampling.

Some reasons of sampling are;

1. To contact the whole population would be time consuming.
2. The cost of studying all the items in the population may be prohibitive.
3. The physical impossibility of checking all items in the population.
4. The destructive nature of some tests.
5. The sample results are adequate.
   * 1. **Sampling and non-sampling errors**

Sampling error occurs when the sample is not representative of the population. When random sampling techniques are used to select elements for the sample, sampling errors occurs by chance. Non sampling errors include missing data, recording errors, input processing errors, and analysis errors.

Sampling errors may rise when there is a difference between a sample statistic; say sample mean and the population mean.

**Population Mean:** is the sum of elements in a population divided by the number of elements.

**Sample Mean:** is the sum of the elements in the sample divided by the number of the sample elements.

E.g. suppose that you have a population consists of five households, A, B, C, D, and E. Their monthly income is as follows:

**Households:**  A B C D E Total

**Monthly income** 10,000 7,400 12,000 20,000 5,600 55,000

Assume that you take a sample of three households, A, C, D. Find out sampling and non sampling errors in selecting these samples.

**So/n**

Population mean=x=x/N=55,000/5=11,000, the mean income of the population

Sample mean== =x/n=42,000/3=14,000, the mean income of the sample

Thus the difference between sample mean and population mean is:

 (sample mean)- x(population mean)=14,000-11,000=3,000

Now, suppose that you selected the above mentioned sample comprising A, C, and D households and the income of household C has wrongly been recorded as 15,000. This will result in the sample means;

=10,000+15,000+20,000/3= 15,000

This means that the difference between the sample mean and the population mean is;

- x=15,000-14,000= 1,000

You should recognize that this difference between the sample mean and the population mean is not due to sampling error alone. There is also a non sampling error.

**Sampling error**=3,000

**Non sampling error**=1,000

**Total error**=4,000

* + 1. **Types of samples**

There are two main types of sampling

1. Random(probability) sampling
2. Non-random(non-probability) sampling

In random sampling every unit of the population has the same probability of being selected in the sample. Every element has a known, non zero chance of being selected and included in the sample.

In non-random sampling every unit of the population does not have the same probability of being selected. Here members are not selected by chance.

**Random Sampling Techniques**

1. **Simple random sampling**: the most widely used type of sampling is a simple random sample. A sample selected so that each item or person in the population has the same chance of being included. A table of random numbers is an efficient way to select members of the sample. Every sample of size **n** has the same chance of being chosen. To illustrate the definition, suppose that *Belay, Chaltu, Dawit and Eden* are an office staff population. All four want the same vacation period, but only two can be away at the same time. Consequently, chips lettered B,C, D and E(for *Belay, Chaltu, Dawit and Eden* ) are shaken in the container, and the office manager(blindfolded) draws out two chips. The possible sample of size 2 from the B, C, D, E population are;

**BC BD BE CD CE DE**

Note that B appears in three of the six samples; so

P(B)=3/6=1/2

Similarly, P(C)=P(D)=P(E)=1/2; so i)each element of the population has the same chance, ½, of being chosen to have the desired vacation period. Moreover, ii) each of the six possible samples has the same chance,1/6, of being selected. Consequently, the selection method satisfies the definition of a simple random sample.

Simple Random Sampling

In statistics, a simple random sample from a population is a sample chosen randomly, so that each possible sample has the same probability of being chosen. One consequence is that each member of the population has the same probability of being chosen as any other. In small populations such sampling is typically done "without replacement", i.e., one deliberately avoids choosing any member of the population more than once. Although simple random sampling can be conducted with replacement instead, this is less common and would normally be described more fully as simple random sampling with replacement. Simple random sampling is a method of selecting n units out of a finite population of size N by giving equal probability to all units, or a sampling procedure in which all possible combinations of n units that may be formed from the finite population of size N units have the same probability of selection. There are N Cn distinct possible samples in the case of sampling without replacement; the chance of selecting each one

of them is 1/N Cn. There are Nn possible samples in the case of sampling with replacement, the chance of selecting each one of them is 1/ Nn . Conceptually, simple random sampling is the simplest of the probability sampling techniques. It requires a complete sampling frame, which may not be available or feasible to construct for large populations. Even if a complete frame is available, more efficient approaches may be possible if other useful information is available about the units in the population.

**Check point**

Three students have taken a class test which is marked out of 10. We want to estimate the mean mark using the sample mean as the estimate of the population mean. We take a sample of size 2 in two cases and suppose the marks of the three students are 1, 2 and 6.

The population mean μ is (1+2+6)/3 = 3

The population variance σ2 =Ʃ (xi- μ)2  / N= 14/3=

i) Sampling without replacement

ii) Sampling with replacement

1. **Systematic random sampling:** In this way of sampling every kth element in the population is selected. First the sampling interval K is calculated as the population size divided by the sample size (N/n). Second a random starting point is selected from 1 up to 9, and then the samples are selected by adding the interval from the starting number.
2. **Stratified random sampling:** In stratified random sampling a population is divided into non overlapping subpopulations, called strata and a sample is randomly selected from each stratum. Stratification is often done by using different variables such as gender, geographic region, religion, ethnicity, educational level, income level etc.
3. **Cluster sampling:** In cluster (area) sampling a population is divided into clusters using naturally occurred boundaries. Then clusters are randomly selected and after clusters have been selected all or part of the elements in each cluster are included in the sample.

**Exercise:** *What is the difference between stratified and cluster sampling?*

**Non-random sampling**

1. **Quota sampling:** It is similar to stratified random sampling. Certain population sub classes such as age, gender, or geographic region are used as strata. Here the researcher uses non-random sampling method to gather data from one stratum until the desired quota of sample is filled.
2. **Judgment sampling:** In this way of sampling elements are selected by the judgment of the researcher. If we believe that some population members have more or better information and more representative of the population than other. And if we want to take small amount of samples we use this method of sampling.
3. **Convenience sampling:** Here elements are selected for the convenience of the researcher. The researcher will choose items that are readily available, nearby, and/or willing to participate.

**1.2 Sampling Distributions**

A sampling distribution is a probability distribution of sample statistics (sample means or sample proportions).

**1.2.1 Sampling distribution of the sample mean**

Sample means vary from sample to sample. A probability distribution of all possible sample means of a given sample size is called Sampling distribution of the sample mean.

The following example illustrates the construction of a sampling distribution of the sample mean.

**Example 1:**

The employees and hourly earnings of a company is given below

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Employee name | A | B | C | D | E | F | G |
| Hourly earning | 7 birr | 7 birr | 8 birr | 8 birr | 7 birr | 8 birr | 9 birr |

1. What is the population mean?
2. What is the sampling distribution of the sample mean for samples of size 2?
3. What is the mean of the sampling distribution?
4. What observations can be made about the population and the sampling distribution?

Solution:

1. The population mean is 7.71.

=∑(7+7+8+8+7+8+9)/7 =7.71

1. First calculate the number of samples of size 2

N C n= N! =7! =21.

n!(N-n)! 2!(7-2)!

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample | Employees | Hourly earnings | Sum | Mean |
| 1 | A,B | 7,7 | 14 | 7 |
| 2 | A,C | 7, 8 | 15 | 7.5 |
| 3 | A,D | 7, 8 | 15 | 7.5 |
| 4 | A,E | 7, 7 | 14 | 7 |
| 5 | A,F | 7, 8 | 15 | 7.5 |
| 6 | A,G | 7, 9 | 16 | 8 |
| 7 | B,C | 7, 8 | 15 | 7.5 |
| 8 | B,D | 7, 8 | 15 | 7.5 |
| 9 | B,E | 7, 7 | 14 | 7 |
| 10 | B,F | 7, 8 | 15 | 7.5 |
| 11 | B,G | 7, 9 | 16 | 8 |
| 12 | C,D | 8, 8 | 16 | 8 |
| 13 | C,E | 8, 7 | 15 | 7.5 |
| 14 | C,F | 8, 8 | 16 | 8 |
| 15 | C,G | 8, 9 | 17 | 8.2 |
| 16 | D,E | 8, 7 | 15 | 7.5 |
| 17 | D,F | 8, 8 | 16 | 8 |
| 18 | D,G | 8, 9 | 17 | 8.5 |
| 19 | E,F | 7, 8 | 15 | 7.5 |
| 20 | E,G | 7, 9 | 16 | 8 |
| 21 | F,G | 8, 9 | 17 | 8.5 |

Sampling distribution of the sample mean for n=2

|  |  |  |
| --- | --- | --- |
| Sample mean | Number of means | probability |
| 7 | 3 | 0.1429 |
| 7.5 | 9 | 0.4285 |
| 8 | 6 | 0.2857 |
| 8.5 | 3 | 0.1429 |
|  |  | 1.0000 |

1. The mean of the sampling distribution of the sample mean is :

∑ (7+7.5+…….+8.5)/21 =162/21 =7.71

1. Observations:

* The mean of the distribution of the sample mean (7.71) is equal to the mean of the population:µ=µx
* the dispersion of the distribution of sample means is narrower than the population distribution
* The sampling distribution of sample means tend to become bell shaped and to approximate the normal probability distribution.

**Example 2:**  A construction company has 310 employees who have an average annual salary of Rs.24; 000.The standard deviation of annual salaries is Rs.5, 000. Suppose that the employees of this company launch a demand that the government should institute a law by which their average salary should be at least Rs. 24500, and, suppose that the government decides to check the validity of this demand by drawing a random sample of 100 employees of this company, and acquiring information regarding their present salaries. What is the probability that, in a random sample of 100 employees, the average salary will exceed Rs.24, 500 (so that the government decides that the demand of the employees of this company is unfounded, and hence does not pay attention to the demand (although, in reality, it was justified))?

**Solution:**  The sample size (n = 100) is large enough to assume that the sampling distribution of‾X(X bar) is approximately *normally* distributed with the following mean and standard deviation:



*Note:* Here we have used finite population correction factor (fpc), because the sample size n = 100 is *greater than 5 percent* of the population size N = 310. Since ‾X is approximately N (24000, 412.20), therefore



is approximately N(0, 1).We are required to evaluate P(‾X > 24,500). At‾x = 24,500, we find that



**24000**

**24500**

**0**

**1.21**

**Z**



**0.3869**

**0.1131**

Using the table of areas under the standard normal curve, we find that the area between z = 0 and z = 1.21 is 0.3869.

Hence,

P(‾X > 24,500)

= P (Z > 1.21) = 0.5 – P (0 < Z < 1.21) = 0.5 – 0.3869 = 0.1131.

Hence, the chances are only 11% that in a random sample of 100 employees from this particular construction company, the average salary will exceed Rs.24, 500. In other words, the chances are 89% that, in such a sample, the average salary will *not* exceed Rs.24, 500. Hence, the chances are considerably *high* that the government *might* pay attention to the employees’ demand.

***Check point:***

1. How many samples of size 4 can be drawn from the population 0, 3, 6, 3, 18? and Compute the sampling distribution of the mean?
2. The mean length of life of a certain cutting tool is 41.5 hours with standard deviation of 2.5 hours. What is the probability that a simple random sample of size 50 drawn from this population will have a mean between 40.5 hours and 42 hours?

**1.2.2 Sampling Distribution of the Proportion**

Just as we are interested in relating the sample mean to the population mean, we would also like to relate sample proportions or percentages to population proportions or percentages. This relationship is especially applicable in the analysis of qualitative data. For example, how do we relate the proportion of successes in a sample to the proportion of successes in the population? If we find that the percentage of defective radios from a sample on the assembly line is 2.5%, what can we conclude about the percentage of defective radios in the entire production lot? Similarly, if a sample of students taking the elementary course in statistics suggests that 30% of these students want to become statisticians, then based up on this sample result, what we can say about the percentage of students wanting to become statisticians from all the students taking the elementary statistics course. Similar to sampling distribution of the means, the sampling distribution of proportions provides some answers.

The sampling distribution of the proportions can be defined as a distribution of proportions of all possible random samples of a fixed size **n**.

We use P to mean the parameter population proportion of successes.

**Example 1:** If population has N=5 numbers i.e. 0, 3, 6, 3, and 18

It contains, three even numbers, 0, 6, 18

Two odd numbers, 3 and 3

So, the population proportion of even numbers (“success”) is P=3/5=0.6

Now consider a sample of size n=3, say, the first three population numbers 0, 3, and 6. Two of the numbers in the sample are even, hence the sample proportion of even numbers is =2/3 where = the sample proportion

Thus, is a statistic; it can have different values in different samples:

0, 3, 3 has =1/3

0, 6, 18 has =1

Ten samples of size n=3 can be drawn from the population of N=5. The values for all ten samples of three are; 2/3, 1/3, 2/3, 2/3, 1, 2/3, 1/3, 2/3, 1/3, 2/3

Sampling distribution of the proportion for this example;

|  |  |  |
| --- | --- | --- |
| Sample proportion() | frequency(f) | Probability of |
| 1/3 | 3 | 0.3 |
| 2/3 | 6 | 0.6 |
| 1 | 1 | 0.1 |
| **Total** | 10 | 1.0 |

The symbol represents the mean of the distribution of . Note that =0.6 and the population proportion P is also 0.6. =P

The symbol σis called the standard error of the proportion. It is the standard deviation of all possible sample proportions. Let P be any given population proportion, and q=1-p; then for samples of size n from a population of size N;



When,n> 5% N,

When n<0.05 N

Thus, for our example, P=0.6,(so q=0.4),n=3, and N=5

Standard error of the proportion=0.2

**EXAMPLE 2:** A population consists of six values 1, 3, 6, 8, 9 and 12. Draw all possible samples of size n = 3 *without* replacement from the population and find the proportion of even numbers in each sample. Construct the sampling distribution of sample proportions and verify that:



**SOLUTION:**

The number of possible samples of size n = 3 that could be selected without replacement from a population of size N is. Let  represent the proportion of even numbers in the sample. Then the 20 possible samples and the proportion of even numbers are given as follows:



The sampling distribution of sample proportion is given below:

Sampling Distribution of



# Now

** and **

To verify the given relations, we first calculate the population proportion p. Thus:

**** *Where - X represents the number of even numbers in the population.*

*In other words, *



The sampling distribution of  have the following important properties:

***Property No. 1*:** The mean of the sampling distribution of proportions, denoted by is equal to the population proportion p, that is



***Property No. 2*:** The standard deviation of the sampling distribution of proportions called the *standard error* of and denoted by is given as:



When the sampling is performed *with* replacement

b) When sampling is done *without* replacement from a *finite* population.

(As in the case of the sampling distribution of ‾X(X bar), is known as the finite population correction factor (fpc).)



***Property No. 3:*** *SHAPE OF THE DISTRIBUTION:* The sampling distribution of



is the *binomial* distribution. However, for sufficiently *large* sample sizes, the sampling distribution is approximately normal.

As n → ∞, the sampling distribution of approaches normality:



As a rule of thumb, the sampling distribution of , will be approximately *normal,* whenever both np and nq are equal to or greater than 5. Let us apply this concept to a real-world situation:

**EXAMPLE-3:** Ten percent of the 1-kilogram boxes of sugar in a large warehouse are underweight. Suppose a retailer buys a random sample of 144 of these boxes. What is the probability that at least 5 percent of the sample boxes will be underweight?

**SOLUTION:** Here the statistic is the sample proportion, the sample size (n = 144) is large enough to assume that the sample proportion is approximately normally distributed with mean.

*Mean of the sampling distribution of:*

****

*And Standard Error of:*

****

Therefore, the sampling distribution of is approximately N (0.10, 0.025)



And, hence:

is approximately N(0, 1).

We are required to find the probability that the proportion of underweight boxes in the sample is equal to or greater than 5% i.e., we require

