***CHAPTER - SEVEN***

#### 7. Sampling and Sampling Distribution of the Sample mean

##### 7.1. Basic concepts (terms)

* ***Population:*** is the complete set of possible measurements for which inferences are to be made.
* ***Sample:*** is the set of measurements that are actually collected in the course of an investigation.

It should be selected using some pre-defined sampling technique in such a way that they

represent the population very well.

* ***Parameter:*** a numerical value obtained from a population.
* ***Statistic****:* a numerical value obtained from a sample.
* ***Sampling:*** The process or method of sample selection from the population.
* ***Sampling unit:*** the ultimate unit to be sampled or elements of the population to be sampled.
* ***Sampling frame:*** is the list of all elements (sampling units) in the population.
  + There are ***two*** ways of investigation: ***Census*** and ***sample survey***.
* ***Census:*** a complete enumeration of the population. But in most real problems it cannot be realized, hence we take sample.
* There are ***two*** types of ***errors*** in sample survey:

a) ***Sampling error:***

* + - * Is the discrepancy between the population value and sample value.
      * May arise due to in appropriate sampling techniques applied.

b) ***Non sampling errors:*** are errors due to procedure bias such as:

* + - * Due to incorrect responses and measurements
      * Errors at different stages in processing the data.

***7.2 Reasons for Sampling***

* To save money (cost), time and labor.
* For greater speed, scope and accuracy.
* The only option when the population is infinite (***too large*** to cover).
* To avoid destructive tests.

Because of the above consideration, in practice we take sample and make ***conclusion*** about the population values such as population mean and population variance, known as ***parameters*** of the population.

***7.3 Types of sampling techniques***

There are ***two types*** of sampling techniques.

***A. Probability sampling (Random Sampling)***

* Is a method of sampling in which ***all elements*** in the population have an ***equal chance*** to be ***included*** in to the sample.
  + - There are ***four*** method of probability sampling.

1. Simple random sampling 3. Stratified sampling
2. Systematic sampling 4. Cluster sampling
3. ***Simple random sampling:-***

* Is a method of selecting items from a population such that every possible sample of specific size has an ***equal chance*** of being selected? In this case, sampling may be with or without replacement. Or
* All elements in the population have the ***same pre-assigned*** non zero probability to be included in to the sample.
* This can be done either using the ***lottery method*** or ***table of random numbers***.

***Table of Random Numbers***

* Table of random numbers are tables of the digits each digit having an equal chance of selection at any draw. For convenience, the numbers are put in blocks of five. In using these tables to select a simple random sample, ***the steps are***:
* Number the units in the population from (prepare ***frame*** of the population).
* Then proceed in the following way
* If the first digit of N is a number between 5 and 9 inclusively, the following method of selection is adequate. Suppose and we want .
* Select three columns from the table of random numbers, say columns 25 to 27. Go down the three columns selecting the first 10 distinct numbers between 001 & 528. These are Then the units with these roll numbers are our samples.
* **Note**: If sampling is without replacement, reject all the numbers that comes more than once.

1. ***Systematic Sampling****:*

* A complete list of all elements within the population (sampling frame) is required.
* The procedure starts in determining the first element to be included in the sample.
* Then the technique is to take the kth item from the sampling frame.

* Choose any number between 1 and *k*. Suppose it is
* The unit is selected ***at first*** and then until the required sample size is reached.

***Example:*** Let select one of the first 10 elements at random. If 2 is the selected number, then the list of all selected elements are:

1. ***Stratified Random Sampling****:*

* The population will be ***divided*** in to ***non-overlapping*** but exhaustive groups called ***strata***.
* ***Simple random samples*** will be chosen from ***each*** stratum.
* Elements in the ***same strata*** should be more or less ***homogeneous*** while ***different*** in ***different strata.***
* It is applied if the population is ***heterogeneous***.
* Some of the criteria for dividing a population into strata are: Sex (male, female); Age (under 18, 18 to 28, 29 to 39), Marital Status(M,S,D,W)

1. ***Cluster Sampling****:*

* The population is divided in to non-overlapping groups called ***clusters***.
* A simple random sample of ***groups or cluster of elements*** is chosen and ***all the sampling units*** in the selected clusters will be surveyed.
* Clusters are formed in a way that ***elements within a cluster*** are ***heterogeneous***, i.e. observations in each cluster should be more or less ***dissimilar***.
* Cluster sampling is ***useful when it is difficult or costly*** to generate a simple random sample.

***For example,*** to estimate the average annual household income in a large city we use cluster sampling, because to use simple random sampling we need a complete list of households in the city from which to sample. To use stratified random sampling, we would again need the list of households. A less expensive way is to let each block within the city represent a cluster. A sample of clusters could then be randomly selected, and every household within these clusters could be interviewed to find the average annual household income.

***B. Non - Random Sampling or non - probability sampling.***

* It is a sampling technique in which the choice of individuals for a sample depends on the basis of convenience, personal choice or interest.
* The most common types of non-probability sampling are convenience, judgmental and quota sampling.

1. ***Convenience sampling****:*

* In this method, the decision maker selects a sample from the population in a manner that is relatively easy and convenient.

1. ***Judgmental sampling or purposive sampling:***

* In this case, the person taking the sample has direct or indirect control over which items are selected for the sample.

1. ***Quota sampling:***

* In this method, the decision maker requires the sample to contain a certain number of items with a given characteristic. Many political polls are, in part, quota sampling.
* A quota is established (say 30% men) and researchers are free to choose any respondent they wish as long as the quota is met.

***Note:***

• We have possible samples if sampling is ***with replacement***.

• We have possible samples if sampling is ***with out replacement***.

***Sampling Distribution***

* Given a variable X, if we arrange its values in ***ascending order*** and assign a probability to each of the values or if we present in the form of ***relative frequency*** distribution the result is called ***sampling distribution of X.***

***Sampling Distribution of the sample mean***

* Sampling distribution of the sample mean is a theoretical probability distribution that shows the functional relationship between the possible values of a given sample mean based on samples of size and the probability associated with each value, for all possible samples of size drawn from that particular population.

***Steps for constructing(finding) the sampling distribution of the sample mean***

1. From a finite population of size randomly draw all possible samples of size

2. Calculate the mean for each sample.

3. Summarize the mean obtained in step 2 in terms of frequency distribution.

***Example:*** Suppose we have a population of size , consisting of the age of three children: 1, 2 and 6. If we take 2 samples from the population ***with replacement,*** then find

a The population mean and variance. b. The sampling distribution of the sample mean.

c. The mean of the sample means. d. The variance of the sample means.

***Solution: a.*** *and*

***Step 1:*** Draw all possible samples: ***Step 2:*** Calculate the mean for each sample:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1st 2nd | 1 | 2 | 6 |  | 1st 2nd | 1 | 2 | 6 |
| 1 | (1,1) | (1,2) | (1,6) | 1 | 1 | 1.5 | 3.5 |
| 2 | (2,1) | (2,2) | (2,6) | 2 | 1.5 | 2 | 4 |
| 6 | (6,1) | (6,2) | (6,6) | 6 | 3.5 | 4 | 6 |

***Step 3:*** Summarize the mean obtained in step 2 in terms of frequency distribution.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 1.5 | 2 | 3.5 | 4 | 6 |
|  | 1 | 2 | 1 | 2 | 2 | 1 |

* Therefore, the sampling distribution the sample mean is:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 1.5 | 2 | 3.5 | 4 | 6 |
|  | 1/9 | 2/9 | 1/9 | 2/9 | 2/9 | 1/9 |

c). The mean of sample means is:

Which is the same as the population mean. Hence,

d). The variance of the sample means is:

.

Which is the same as the population variance, divided by :

***Exercise:*** Re-do the above example when sampling is without replacement?

***Remark:*** 1. In general if sampling is ***with replacement***:

2. If sampling is ***without replacement*** :

3. In any case the sample mean is ***unbiased*** estimator of the population mean. i.e.

**Example 2**: Suppose we have a population of size, consisting of the age of five children: 6, 8, 10, 12, and 14



Take samples of size 2 with replacement and construct sampling distribution of the sample mean.

* When sampling is from a normally distributed population, the distribution of will possess the following property.

1. The distribution of the sample mean will be normal.

2. The mean of sample meanis equal to the population mean. i.e.

3. The variance of sample meanis equal to the population variance divided by the sample size, i.e.

***Central limit theorem***

* If a random sample of size is selected from population of size having mean µ and variance , then as the distribution of ***sample mean*** will be normal with mean µ and variance .

***Example****:* If the uric acid values in normal adult males are approximately normally distributed with mean 5.7 mgs and standard deviation 1mg find the probability that a sample of size 9 will yield a mean.

i. greater than 6. ii. between 5 and 6. iii. less than 5.2.

***Solution:*** Let be the amount of uric acids in normal adult males.

### Example

A factory making soft-drinks has an automatic process that fills its bottles. The volume of the soft drink in each bottle is supposed to be 330*ml*, but the machine fills the bottles with a random amount of soft drink that has a mean of 330*ml* and a standard deviation of 5*ml*. Suppose we take a sample of 100 bottles of drink. What is the probability that the mean volume of drink in the sample is more than 331*ml*?

### Solution

As we have taken a large (*n*=100) sample we can use the Central Limit Theorem. This says that the mean amount of soft drink in the bottles of the sample can be approximated by a normal random variable with a mean of 330 and a variance of 52/100 = ¼ (i.e. a standard deviation of ½). That is, μ=330, σ2/n=1/4 and =1/2. If we let Y be the volume of the soft drink in each bottle, it is required to find **P**(*Y* > 331) where *Y*~N (330, ¼):

**P**(*Y* > 331) = 

**=** = P (Z>2.00) = 0.5 – 0.4772 = 0.228

### Example 2

Suppose that 150 customers enter a supermarket on a given day. Each customer spends a random amount. All they knew about the distribution of these expenditures that its mean is 7.50 birr and its standard deviation is 3.40 birr.

What is the probability that a person, on average, spent more than 8.00 birr during the day?

Solution

We have n = 150 which is large enough to use the Central Limit Theorem.

Mean = μ =7.50 and standard deviation = σ= 3.40 ⇒ σ2 = 3.42 = 11.56

Let X be the amount of an individual’s expenditure during the day. X~ *N* (7.50, 11.56)

It is required to find *P(X>8.00)*

 =0.5 – P (0<Z<1.80) = 0.5 – 0.4641 = 0.0359

This means there is only 0.0359 probabilities that a person will spent larger than 8.00 birr.

Exercises

1. Suppose that the population distribution of the gripping strengths of industrial workers is known to have a mean of 110 and standard deviation of 10. For a random sample of 75 workers, what is the probability that the sample mean gripping strength will be
2. Between 109 and 112?
3. Greater than112?
4. The amount of sulphur in a daily emission from a factory has a normal distribution with mean of 134 pounds and a standard deviation of 22 pounds. For a day selected randomly, find the probability that the mean amount of sulphur emission will be less than 130 pounds.