8 Sampling Techniques and Statistical Inference

Pandurang Vasudeo Sukhatme
(July 27, 1911 - January 28, 1997)

Introduction

Pandurang Vasudeo Sukhatme (1911–1997) was an award-winning Indian statistician. He is known for his pioneering work of applying random sampling methods in agricultural statistics and in biometry, in the 1940s. He was also influential in the establishment of the Indian Agricultural Statistics Research Institute. As a part of his work at the Food and Agriculture Organization in Rome, he developed statistical models for assessing the dimensions of hunger and future food supplies for the world. He also developed methods for measuring the size and nature of the protein gap.

In any statistical investigation, the interest lies in the assessment of one or more characteristics relating to the individuals belonging to a group. When all the individuals present in the study are investigated, it is called complete enumeration, but in practice, it is very difficult to investigate all the individuals present in the study. So the technique of sampling is done which states that a part of the individuals are selected for the study and the assessment is made from the selected group of individuals. For example

(i) A housewife tastes a spoonful whatever she cooks to check whether it tastes good or not.
(ii) A few drops of our blood are tested to check about the presence or absence of a disease.
(iii) A grain merchant takes out a handful of grains to get an idea about the quality of the whole consignment.

These are typical examples where decision making is done on the basis of sample information. So sampling is the process of choosing a representative sample from a given population.

Learning Objectives

After studying this chapter students are able to understand

- sampling techniques
- random sampling
- simple random sampling
- stratified random sampling
- systematic sampling
- sampling and non-sampling errors
- sampling distribution
- statistical inference
- estimation
- test of statistical hypothesis

8.1 Sampling

Sampling is the procedure or process of selecting a sample from a population. Sampling is quite often used in our day-to-day practical life.

8.1.1 Basic concepts of sampling

Population

The group of individuals considered under study is called as population. The word population here refers not only to people but to all items that have been chosen for the study. Thus in statistics, population can be number of bikes manufactured in a day or week or month, number of cars manufactured in a day or week or month, number of fans, TVs, chalk pieces, people, students, girls, boys, any manufacturing products, etc…

Finite and infinite population:

When the number of observations/individuals/products is countable in a group, then it is a finite population. Example: weights of students of class XII in a school.

When the number of observations/individuals/products is uncountable in a group, then it is an infinite population. Example: number of grains in a sack, number of germs in the body of a sick patient.

Sample and sample size

A selection of a group of individuals from a population in such a way that it represents the population is called as sample and the number of individuals included in a sample is called the sample size.

Parameter and statistic

Parameter: The statistical constants of the population like mean (μ), variance(σ^2) are referred as population parameters.

Statistic : Any statistical measure computed from sample is known as statistic.

Note

In practice, the parameter values are not known and their estimates based on the sample values are generally used.

Types of sampling

The technique or method of selecting a sample is of fundamental importance in the theory of sampling and usually depends upon the nature of the data and the type of enquiry. The procedures of selecting a sample may be broadly classified as

1. Non-Random sampling or Non-probability sampling.
2. Random Sampling or Probability sampling.

Note

Here we confine ourselves to Random sampling or Probability sampling.

Random sampling or Probability sampling

Random sampling refers to selection of samples from the population in a random manner. A random sample is one where each and every item in the population has an equal chance of being selected.

" Every member of a parent population has had equal chances of being included".- Dr. Yates

"A random sample is a sample selected in such a way that every item in the population has an equal chance of being included".-Harper

The following are different types of probability sampling:

(i) Simple random sampling
(ii) Stratified random sampling
(iii) Systematic sampling

(i) Simple random sampling

In this technique the samples are selected in such a way that each and every unit in the population has an equal and independent chance of being selected as a sample. Simple random sampling may be done, with or without replacement of the samples selected. In a simple random sampling with replacement there is a possibility of selecting the same sample any number of times. So, simple random sampling without replacement is followed.

Thus in simple random sampling from a population of N units, the probability of drawing any unit at the first draw is $\frac{1}{N}$, the probability of drawing any unit in the second draw from among the available $(N-1)$ units is $\frac{1}{(N-1)}$, and so on.. Several methods have been adopted for random selection of the samples from the population. Of those, the following two methods are generally used and which are described below.

(A) Lottery method

This is the most popular and simplest method when the population is finite. In this method, all the items of the population are numbered on separate slips of paper of same size, shape and colour. They are folded and placed in a container and shuffled thoroughly. Then the required numbers of slips are selected for the desired sample size. The selection of items thus depends on chance.

For example, if we want to select 10 students, out of 100 students, then we must write the names/roll number of all the 100 students on slips of the same size and mix them, then we make a blindfold selection of 10 students. This method is called unrestricted random sampling, because units are selected from the population without any restriction. This method is mostly used in lottery draws. If the population or universe is infinite, this method is inapplicable.

(B) Table of Random number

When the population size is large, it is difficult to number all the items on separate slips of paper of same size, shape and colour. The alternative method is that of using the table of random numbers. The most practical, easy and inexpensive method of selecting a random sample can be done through "Random Number Table". The random number table has been so constructed that each of the digits 0,1,2,…,9 will appear approximately with the same frequency and independently of each other.

The various random number tables available are

a. L.H.C. Tippett random number series
b. Fisher and Yates random number series
c. Kendall and Smith random number series
d. Rand Corporation random number series.

Tippett's table of random numbers is most popularly used in practice. Given below the first forty sets from Tippett's table as an illustration of the general appearance of random numbers:

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2670	7483	3408	2762	3563	1089	6913	7991
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

Suppose, if we want to select the required number of samples from a population of size N(<99) then any two digit random number can be selected from (00 to 99) from the above random number table. Similarly if N(< 999) or (< 9999), then any three digit random number or four digit random number can be selected from (000 to 999) or (0000 to 9999).

The procedure of selecting the random samples consists of following steps:

1. Identify the N units in the population with the numbers from 1 to N.
2. Select at random, any page from the 'Random Number Table'.
3. Select the required number of samples from any row or column or diagonal.

One may question, as to how it is ensured that these digits are random. It may be pointed out that the digits in the table were chosen horizontally but the real guarantee of their randomness lies in practical tests. Tippett's numbers have been subjected to numerous tests and used in many investigations and their randomness has been very well established for all practical purposes.

An example to illustrate how Tippett's table of random numbers may be used is given below. Suppose we have to select 20 items out of 6,000. The procedure is to number all the 6,000 items from 1 to 6,000. A page from Tippett's table may be selected and the first twenty numbers ranging from 1 to 6,000 are noted down. If the numbers are above 6000, choose the next number ranging from 1 to 6000. Items bearing those numbers will be selected as samples from the population. Making use of the portion of the random number table given,the requiredrandom samples are shaded. Here, we consider row wise selection of random numbers.

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2670	7483	3408	2762	3563	1089	6913	7991
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

If the population size is 1,000 and suppose we want to select 15 items out of 1,000. All itemsfrom 1 to 1000 should be numbered as 0001 to 1000.Now, we may now select 15 numbers from the random number table.The procedure will be different, as Tippett's random numbers are available only in four digits. Thus, we can select the first three digits from the four digit random sample number.Making use of the portion of the random number table given, the required random samples are shaded in RED colour. Here, we consider row wise selection of random numbers.

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2670	7483	3408	2762	3563	1089	6913	7991
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

If the population size is 100 and suppose we want to select 10 items out of 100. All itemsfrom 1 to 100 should be numbered as 001 to 100. Now, we may now select 10 numbers from the random number table. The procedure will be different, as Tippett's random numbers are available only in four digits. Thus, we can select the first two digits from the four digit random sample number. Making use of the portion of the random number table given, the required random samples are shaded in RED colour.Here, we consider row wise selection of random numbers.

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2670	7483	3408	2762	3563	1089	6913	7991
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

Merits and demerits of Simple Random Sampling:

Merits

1. Personal bias is completely eliminated.
2. This method is economical as it saves time, money and labour.
3. The method requires minimum knowledge about the population in advance.

Demerits

1. This requires a complete list of the population but such up-to-date lists are not available in many enquiries.
2. If the size of the sample is small, then it will not be a representative of the population.

Example 8.1

Using the Kendall-Babington Smith - Random number table,Draw5 random samples.

23	15	75	48	59	01	83	72	59	93	76	24	97	08	86	95	23	03	67	44
05	54	55	50	43	10	53	74	35	08	90	61	18	37	44	10	96	22	13	43
14	87	16	03	50	32	40	43	62	23	50	05	10	03	22	11	54	36	08	34
38	97	67	49	51	94	05	17	58	53	78	80	59	01	94	32	42	87	16	95
97	31	26	17	18	99	75	53	08	70	94	25	12	58	41	54	88	21	05	13

Solution:

23	15	75	48	59	01	83	72	59	93	76	24	97	08	86	95	23	03	67	44
05	54	55	50	43	10	53	74	35	08	90	61	18	37	44	10	96	22	13	43
14	87	16	03	50	32	40	43	62	23	50	05	10	03	22	11	54	36	08	34
38	97	67	49	51	94	05	17	58	53	78	80	59	01	94	32	42	87	16	95
97	31	26	17	18	99	75	53	08	70	94	25	12	58	41	54	88	21	05	13

There many ways to select 5 random samples from the given Kendall-Babington Smith - Random number table. Assume that at random we select 3rd column 1st value. This location gives the digit to be 75. So the first sample will be 75, and then the next choices can be in the same 3rd column which follows as 55, 16, 67 and 26. Therefore 75,55,16,67 and 26 will be used as random samples. The various shaded numbers can be taken as 5 random sample numbers. Apart from this, one can select any 5 random sample numbers as they like.

Example 8.2

Using the following Tippett's random number table,

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2670	7483	3408	2762	3563	1089	6913	7991
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

Draw a sample of 15 houses from Cauvery Street which has 83 houses in total.

Solution:

There many ways to select 15 random samples from the given Tippet's random number table. Since the population size is 83 (two-digit number). Here the door numbers are assigned from 1 to 83. Assume that at random we first choose 2nd column. So the first sample is 66 and other 14 samples are 74, 52, 39, 15, 34, 11, 14, 13, 27, 61, 79, 72, 35, and 60. If the numbers are above 83, choose the next number ranging from 1 to 83.

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2670	7483	3408	2762	3563	1089	6913	7991
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

Example 8.3

Using the following random number table,

Tippet's random number table							
2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2670	7483	3408	2762	3563	1089	6913	7991
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

Draw a sample of 10 children with theirheight from the population of 8,585 children as classified hereunder.

Height (cm)	105	107	109	111	113	115	117	119	121	123	125
Number of children	2	4	14	41	83	169	394	669	990	1223	1329
Height(cm)	127	129	131	133	135	137	139	141	143	145	
Number of children	1230	1063	646	392	202	79	32	16	5	2	

Solution:

The first thing is to number the population(8585 children). The numbering has already been provided by the frequency table. There are 2 children with height of 105 cm, therefore we assign number 1 and 2 to the children those in the group 105 cm, number 3 to 6 is assigned to those in the group 107 cm and similarly all other children are assigned the numbers. In the last group 145 cms there are two children with assigned number 8584 and 8585.

Height (cm.)	Number of children	Cumulative Frequency
105	2	2
107	4	6
109	14	20
111	41	61
113	83	144
115	169	313
117	394	707
119	669	1376
121	990	2366
123	1223	3589
125	1329	4918
127	1230	6148
129	1063	7211
131	646	7857
133	392	8249
135	202	8451
137	79	8530
139	32	8562
141	16	8578
143	5	8583
145	2	8585
Total	8585	

Now we take 10 samples from the tables, since the population size is in 4 digits we can use the given random number table. Select the10 random numbers from 1 to 8585 in the table, Here, we consider column wise selection of random numbers, starting from first column.

Tippet's random number table

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2670	7483	3408	2762	3563	1089	6913	7991
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

The children with assigned number 2952 is selected and then see the cumulative frequency table where 2952 is present, now select the corresponding row height which is 123 cm, similarly all the selected random numbers are considered for the selection of the child with their corresponding height. The following table shows all the selected 10 children with their heights.

Child with assigned Number	2952	4167	2670	0560	2754
Corresponding Height (cms)	123	125	123	117	123
Child with assigned Number	6641	7483	5246	3992	1545
Corresponding Height (cms)	129	131	127	125	121

Example 8.4

Using the following random number table (Kendall-Babington Smith)

23	15	75	48	59	01	83	72	59	93	76	24	97	08	86	95	23	03	67	44
05	54	55	50	43	10	53	74	35	08	90	61	18	37	44	10	96	22	13	43
14	87	16	03	50	32	40	43	62	23	50	05	10	03	22	11	54	36	08	34
38	97	67	49	51	94	05	17	58	53	78	80	59	01	94	32	42	87	16	95
97	31	26	17	18	99	75	53	08	70	94	25	12	58	41	54	88	21	05	13

Draw a random sample of 10 four- figure numbers starting from 1550 to 8000.

Solution:

Here, we have to select 10 random numbers ranging from 1550 to 8000 but the given random number table has only 2 digit numbers. To solve this, two - 2 digit numbers can be combined together to make a four-figure number. Let us select the 5^{th} and 6^{th} column and combine them to form a random number, then select the random number with given range. This gives 5 random numbers, similarly 8^{th} and 9^{th} is selected and combined to form a random numbers, then select the random number with given range. This gives 5 random numbers, totally 10 four-figure numbers have been selected. The following table shows the 10 random numbers which are combined and selected.

23	15	75	48	59	01	83	72	59	93	76	24	97	08	86	95	23	03	67	44
05	54	55	50	43	10	53	74	35	08	90	61	18	37	44	10	96	22	13	43
14	87	16	03	50	32	40	43	62	23	50	05	10	03	22	11	54	36	08	34
38	97	67	49	51	94	05	17	58	53	78	80	59	01	94	32	42	87	16	95
97	31	26	17	18	99	75	53	08	70	94	25	12	58	41	54	88	21	05	13

Therefore the selected 10 random numbers are

5901	4310	5032	5194	1899
7259	7435	4362	1758	5308

(ii) Stratified Random Sampling

Definition 8.1

In stratified random sampling, first divide the population into sub-populations, which are called strata. Then, the samples are selected from each of the strata through random techniques.The collection of all the samples from all strata gives the stratified random samples.

When the population is heterogeneous or different segments or groups with respect to the variable or characteristic under study, then Stratified Random Sampling method is studied. First, the population is divided into homogeneous number of sub-groups or strata before the sample is drawn. A sample is drawn from each stratum at random. Following steps are involved for selecting a random sample in a stratified random sampling method.

(a) The population is divided into different classes so that each stratum will consist of more or less homogeneous elements. The strata are so designed that they do not overlap each other.

(b) After the population is stratified, a sample of a specified size is drawn at random from each stratum using Lottery Method or Table of Random Number Method.

Stratified random sampling is applied in the field of the different legislative areas as strata in election polling, division of districts (strata) in a state etc…

Example 8.5

From the following data, select 68 random samples from the populationof heterogeneous group with size of 500 through stratified random sampling, considering the following categories as strata.

Category1: Lower income class -39%

Category2: Middle income class - 38%

Category3: Upper income class- 23%

Solution

Stratum	Homogenous group	Percentage from population	Number of people in each strata	Random Samples
Category1	Lower income class	39	$\frac{39}{100}\times 500 = 195$	$195\times\frac{68}{500} = 26.5\sim 26$
Category2	Middle income class	38	$\frac{38}{100}\times 500 = 190$	$190\times\frac{68}{500} = 26.5\sim 26$
Category3	Upper income class	23	$\frac{23}{100}\times 500 = 115$	$115\times\frac{68}{500} = 15.6\sim 16$
Total		**100**	**500**	**68**

Merits

(a) A random stratified sample is superior to a simple random sample because it ensures representation of all groups and thus it is more representative of the population which is being sampled.

(b) A stratified random sample can be kept small in size without losing its accuracy.

(c) It is easy to administer,if the population under study is sub-divided.

(d) It reduces the time and expenses in dividing the strata into geographical divisions, since the government itself had divided the geographical areas.

Demerits

(a) To divide the population into homogeneous strata (if not divided), it requires more money, time and statistical experience which is a difficult one.

(b) If proper stratification is not done, the sample will have an effect of bias.

(c) There is always a possibility of faulty classification of strata and hence increases variability.

(iii) Systematic Sampling

Definition 8.2

In a systematic sampling, randomly select the first sample from the first k units. Then every k^{th} member, starting with the first selected sample, is included in the sample.

Systematic sampling is a commonly used technique, if the complete and up-to-date list of the sampling units is available. We can arrange the items in numerical, alphabetical, geographical or in any other order. The procedure of selecting the samples starts with selecting the first sample at random, the rest being automatically selected according to some pre-determined pattern. A systematic sample is formed by selecting every item from the population, where k refers to the sample interval. The sampling interval can be determined by dividing the size of the population by the size of the sample to be chosen. That is $k=\frac{N}{n}$,where k is an integer.

k = Sampling interval, N = Size of the population, n = Sample size.

Procedure for selection of samples by systematic sampling method

(i) If we want to select a sample of 10 students from a class of 100 students, the sampling interval is calculated as $k=\frac{N}{n}=\frac{100}{10}=10$.

Thus sampling interval = 10 denotes that for every 10 samples one sample has to be selected.

(ii) The firstsample is selected from the first 10 (sampling interval) samples through random selection procedures.

(iii) If the selected first random sample is 5, then the rest of the samples are automatically selected by incrementing the value of the sampling interval (k = 10) i.e., 5, 15 , 25, 35, 45, 55, 65, 75, 85, 95.

Example:

Suppose we have to select 20 items out of 6,000. The procedure is to number all the 6,000 items from 1 to 6,000. The sampling interval is calculated as $k=\frac{N}{n}=\frac{6000}{20}=300$. Thus sampling interval = 300 denotes that for every 300 samples one sample has to be selected. The first sample is selected from the first 300 (sampling interval) samples through random selection procedures. If the selected first random sample is 50, then the rest of the samples are automatically selected by incrementing the value of the sampling interval (k=300) ie,50, 350, 650, 950, 1250, 1550, 1850, 2150, 2450, 2750, 3050, 3350, 3650, 3950, 4250, 4550, 4850, 5150, 5450, 5750. Items bearing those numbers will be selected as samples from the population.

Merits

1. This is simple and convenient method.
2. This method distributes the sample more evenly over the entire listed population.
3. The time and work is reduced much.

Demerits

1. Systematic samples are not random samples.
2. If N is not a multiple of n, then the sampling interval (k) cannot be an integer, thus sample selection becomes difficult.

8.1.2 Sampling and Non-Sampling Errors:

A sample is a part of the whole population. A sample drawn from the population depends upon chance and as such all the characteristics of the population may not be present in the sample drawn from the same population. The errors involved in the collection, processing and analysis of the data may be broadly classified into two categories namely,

(i) Sampling Errors
(ii) Non-Sampling Errors

(i) Sampling Errors

Errors, which arise in the normal course of investigation or enumeration on account of chance, are called sampling errors. Sampling errors are inherent in the method of sampling. They may arise accidentally without any bias or prejudice. Sampling Errors arise primarily due to the following reasons:

(a) Faulty selection of the sample instead of correct sample by defective sampling technique.
(b) The investigator substitutes a convenient sample if the original sample is not available while investigation.
(c) In area surveys, while dealing with border lines it depends upon the investigator whether to include them in the sample or not. This is known as Faulty demarcation of sampling units.

(ii)Non-Sampling Errors

The errors that arise due to human factors which always vary from one investigator to another in selecting, estimating or using measuring instruments(tape, scale)are called Non-Sampling errors.It may arise in the following ways:

(a) Due to negligence and carelessness of the part of either investigator or respondents.
(b) Due to lack of trained and qualified investigators.
(c) Due to framing of a wrong questionnaire.
(d) Due to apply wrong statistical measure
(e) Due to incomplete investigation and sample survey.

8.1.3 Sampling distribution

Definition 8.3

Sampling distribution of a statistic is the frequency distribution which is formed with various values of a statistic computed from different samples of the same size drawn from the same population.

For instance if we draw a sample of size n from a given finite population of size N, then the total number of possible samples is ${}^{N}C_n = \dfrac{N!}{n!(N-n)!} = k$ (say). For each of these k samples we can compute some statistic, $t = t(x_1, x_2, x_3, \ldots x_n)$, in particular the mean $\bar{x}$, the variance S^2, etc., is given below:

Sample Number	Statistic		
	t	$\bar{x}$	s^2
1	t_1	$\bar{x}_1$	s_1^2
2	t_2	$\bar{x}_2$	s_2^2
3	t_3	$\bar{x}_3$	s_3^2
. . .	. . .	. . .	. . .
k	t_k	$\bar{x}_k$	s_k^2

The set of the values of the statistic so obtained, one for each sample, constitutes the sampling distribution of the statistic.

Standard Error

The standard deviation of the sampling distribution of a statistic is known as its Standard Error abbreviated as S.E. The Standard Errors (S.E.) of some of the well-known statistics, for large samples, are given below, where n is the sample size, σ^2 is the population variance.

Sl. No.	Statistic	Standard Error
1.	Sample mean ($\bar{x}$)	$\sigma/\sqrt{n}$
2.	Observed sample proportion (p)	$\sqrt{PQ/n}$
3.	Sample standard deviation (s)	$\sqrt{\sigma^2/2n}$
4.	Sample variance (s^2)	$\sigma^2\sqrt{2/n}$
5.	Sample quartiles	$1.36263\,\sigma/\sqrt{n}$
6.	Sample median	$1.25331\,\sigma/\sqrt{n}$
7.	Sample correlation coefficient (r)	$(1-\rho^2)/\sqrt{n}$

8.1.4 Computing standard error in simple cases

Example 8.6

A server channel monitored for an hour was found to have an estimated mean of 20 transactions transmitted per minute. The variance is known to be 4. Find the standard error.

Solution:

Given $\sigma^2 = 4$ which implies $\sigma = 2$, $n = 1$ hour = 60 min, $\bar{X} = 20$ /min

Standard Error = $\dfrac{\sigma}{\sqrt{n}} = \dfrac{2}{\sqrt{60}} = 0.2582$

Example 8.7

Find the sample size for the given standard deviation 10 and the standard error with respect of sample mean is 3.

Solution:

Given $\sigma = 10$, S.E. $\bar{X}$ =3 We know that S.E = $\sigma/\sqrt{n}$

Therefore, $3 = \dfrac{10}{\sqrt{n}} \Rightarrow \sqrt{n} = \dfrac{10}{3}$

Taking Squaring on both sides we get

$$n = \left(\frac{10}{3}\right)^2 = \frac{100}{9} = 11.11 \cong 11,$$

The required sample size is 11.

Example 8.8

A die is thrown 9000 times and a throw of 3 or 4 is observed 3240 times. Find the standard error of the proportion for an unbiased die .

Solution :

If the occurrence of 3 or 4 on the die is called a success, then

Sample size = 9000; Number of Success = 3240

Sample proportion = $p = \dfrac{3240}{9000} = 0.36$

Population proportion (P) = Prob (getting 3 or 4 when a die is thrown)

$$= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3} = 0.3333$$

Thus $P = 0.3333$ and
$Q = 1-P=1- 0.3333 = 0.6667$.

The S.E for sample proportion is given by

$$S.E. = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{(0.3333)(0.6667)}{9000}} = 0.00496$$

Hence the standard error for sample proportion is S.E=0.00496.

Example 8.9

The standard deviation of a sample of size 50 is 6.3. Determine the standard error whose population standard deviation is 6?

Solution:

Sample size $n = 50$

Sample S.D $s = 6.3$

Population S.D $\sigma = 6$

The standard error for sample S.D is given by

$$S.E. = \sqrt{\frac{\sigma^2}{2n}} = \frac{6}{\sqrt{2(50)}} = \frac{6}{\sqrt{100}} = 0.6$$

Thus standard error for
sample S.D = 0.6.

Example 8.10

A sample of 100 students is chosen from a large group of students. The average height of these students is 162 cm and standard deviation (S.D) is 8 cm. Obtain the standard error for the average height of large group of students of 160 cm?

Solution:

Given $n = 100$, $\bar{x}$ =162 cm, s = 8 cm is known in this problem

since σ is unknown, so we consider $\hat{\sigma} = s$ and φ= 160 cm

$$S.E. = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{8}{\sqrt{100}} = 0.8$$

Therefore the standard error for the average height of large group of students of 160 cm is 0.8.

Exercise 8.1

1. What is population?
2. What is sample?
3. What is statistic?
4. Define parameter.
5. What is sampling distribution of a statistic?
6. What is standard error?
7. Explain in detail about simple random sampling with a suitable example.
8. Explain the stratified random sampling with a suitable example.
9. Explain in detail about systematic random sampling with example.
10. Explain in detail about sampling error.
11. Explain in detail about non-sampling error.
12. State any two merits of simple random sampling.
13. State any three merits of stratified random sampling.
14. State any two demerits of systematic random sampling.
15. State any two merits for systematic random sampling.
16. Using the following Tippet's random number table

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2670	7483	3408	2762	3563	1089	6913	7991
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

Draw a sample of 10 three digit numbers which are even numbers.

17. A wholesaler in apples claims that only 4% of the apples supplied by him are defective. A random sample of 600 apples contained 36 defective apples. Calculate the standard error concerning of good apples.
18. A sample of 1000 students whose mean weight is 119 lbs(pounds) from a school

in Tamil Nadu State was taken and their average weight was found to be 120 lbs with a standard deviation of 30 lbs. Calculate standard error of mean.

19. A random sample of 60 observations was drawn from a large population and its standard deviation was found to be 2.5. Calculate the suitable standard error that this sample is taken from a population with standard deviation 3?

20. In a sample of 400 population from a village 230 are found to be eaters of vegetarian items and the rest non-vegetarian items. Compute the standard error assuming that both vegetarian and non-vegetarian foods are equally popular in that village?

Statistical Inference

One of the main objectives of any statistical investigation is to draw inferences about a population from the analysis of samples drawn from that population. Statistical Inference provides us how to estimate a value from the sample and test that value for the population. This is done by the two important classifications in statistical inference,

(i) Estimation;

(ii) Testing of Hypothesis

8.2 Estimation:

It is possible to draw valid conclusion about the population parameters from sampling distribution. Estimation helps in estimating an unknown population parameter such as population mean, standard deviation, etc., on the basis of suitable statistic computed from the samples drawn from population.

Estimation:

Definition 8.4

The method of obtaining the most likely value of the population parameter using statistic is called estimation.

Estimator:

Definition 8.5

Any sample statistic which is used to estimate an unknown population parameter is called an estimator ie., an estimator is a sample statistic used to estimate a population parameter.

Estimate:

Definition 8.6

When we observe a specific numerical value of our estimator, we call that value is an estimate. In other words, an estimate is a specific observed value of a statistic.

Characteristic of a good estimator

A good estimator must possess the following characteristic:

(i) Unbiasedness (ii) Consistency (iii) Efficiency (iv) Sufficiency.

(i) Unbiasedness: An estimator $T_n = T(x_1, x_2, ..., x_n)$ is said to be an unbiased estimator of $\gamma(\theta)$ *if* $E(T_n) = \gamma(\theta)$, for all $\theta \; \varepsilon \Theta$ (parameter space), (i.e) An estimator is said to be unbiased if its expected value is equal to the population parameter. Example: $E(\overline{x}) = \varphi$

(ii) Consistency: An estimator $T_n = T(x_1, x_2, ..., x_n)$ is said to be consistent estimator of $\gamma(\theta)$, if T_n converges to $\gamma(\theta)$ in Probability, i.e., $T_n \xrightarrow{P} \gamma(\theta) \, as \, n \to \infty$, for all $\theta \; \varepsilon \Theta$.

(iii) Efficiency: If T_1 is the most efficient estimator with variance V_1 and T_2 is any other estimator with variance V_2, then the efficiency E of T_2 is defined as $E = \dfrac{V_1}{V_2}$ Obviously, E cannot exceed unity.

(iv) Sufficiency: If $T = t(x_1, x_2, ..., x_n)$ is an estimator of a parameter θ, based on a sample $x_1, x_2, ..., x_n$ of size n from the population with density $f(x,\theta)$ such that the conditional distribution of $x_1, x_2, ..., x_n$ given T, is independent of θ, then T is sufficient estimator for θ.

8.2.1 Point and Interval Estimation:

To estimate an unknown parameter of the population, concept of theory of estimation is used.There are two types of estimation namely,

1. Point estimation
2. Interval estimation

1. Point Estimation

When a single value is used as an estimate, the estimate is called a point estimate of the population parameter. In other words, an estimate of a population parameter given by a single number is called as point estimation.

For example

(i) 55 is the mean mark obtained by a sample of 5 students randomly drawn from a class of 100 students is considered to be the mean marks of the entire class. This single value 55is a point estimate.

(ii) 50 kg is the average weight of a sample of 10 students randomly drawn from a class of 100 students is considered to be the average weight of the entire class. This single value 50 is a point estimate.

Note

The sample mean $(\bar{x})$ is the sample statistic used as an estimate of population mean (μ)

Instead of considering, the estimated value of the population parameter to be a single value, we might consider an interval for estimating the value of the population parameter. This concept is known as interval estimation and is explained below.

2. Interval Estimation

Generally, there are situations where point estimation is not desirable and we are interested in finding limits within which the parameter would be expected to lie is called an interval estimation.

For example,

If T is a good estimator of θ with standard error s then, making use of general property of the standard deviations, the uncertainty in *T*, as an estimator of θ, can be expressed by statements like " We are about 95% certain that the unknown θ, will lie somewhere between T-2s and T+2s", "we are almost sure that θ will in the interval (T-3s and T+3s)" such intervals are called confidence intervals and is explained below.

Confidence interval

After obtaining the value of the statistic 't' (sample) from a given sample, Can we make some reasonable probability statements about the unknown population parameter 'θ'?. This question is very well answered by the technique of Confidence Interval. Let us choose a small value of α which is known as level of significance(1% or 5%) and determine two constants say, c_1 and c_2 such that $P(c_1 < \theta < c_2 | t) = 1 - \alpha$.

The quantities c_1 and c_2, so determined are known as the Confidence Limits and the interval $[c_1, c_2]$ within which the unknown value of the population parameter is expected to lie is known as Confidence Interval. $(1-\alpha)$ is called as confidence coefficient.

Confidence Interval for the population mean for Large Samples (when s is known)

If we take repeated independent random samples of size n from a population with an unknown mean but known standard deviation, then the probability that the true population mean φ will fall in the following interval is $(1-\alpha)$ i.e

$$P = \left(\bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = (1-\alpha)$$

So, the confidence interval for population mean (φ), when standard deviation (σ) is known and is given by $\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

For the computation of confidence intervals and for testing of significance, the critical values Z_α at the different level of significance is given in the following table:

Normal Probability Table

Critical Values Z_α	Level of significance (α)			
	1%	2%	5%	10%
Two-tailed test	$\lvert Z_\alpha \rvert = 2.58$	$\lvert Z_\alpha \rvert = 2.33$	$\lvert Z_\alpha \rvert = 1.96$	$\lvert Z_\alpha \rvert = 1.645$
Right tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 2.055$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -2.055$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

The calculation of confidence interval is illustrated below.

Example 8.11

A machine produces a component of a product with a standard deviation of 1.6 cm in length. A random sample of 64 componentsvwas selected from the output and this sample has a mean length of 90 cm. The customer will reject the part if it is either less than 88 cm or more than 92 cm. Does the 95% confidence interval for the true mean length of all the components produced ensure acceptance by the customer?

Solution:

Here φ is the mean length of the components in the population.

The formula for the confidence interval is

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Here $\sigma = 1.6$, $Z_{\alpha/2} = 1.96$, $\bar{x} = 90$ and n = 64

Then $S.E. = \frac{\sigma}{\sqrt{n}} = \frac{1.6}{\sqrt{64}} = 0.2$

Therefore, $90 - (1.96 \times 0.2) \le \varphi \le 90 + (1.96 \times 0.2)$

i.e. $(89.61 \le \varphi \le 90.39)$

This implies that the probability that the true value of the population mean length of the components will fall in this interval (89.61,90.39) at 95% . Hence we concluded that 95% confidence interval ensures acceptance of the component by the consumer.

Example 8.12

A sample of 100 measurements at breaking strength of cotton thread gave a mean of 7.4 and a standard deviation of 1.2 gms. Find 95% confidence limits for the mean breaking strength of cotton thread.

Solution:

Given, sample size = 100, $\bar{x} = 7.4$, since σ is unknown but s = 1.2 is known.

In this problem, we consider $\breve{\sigma} = s$, $Z_{\alpha/2} = 1.96$

$$S.E. = \frac{\breve{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{1.2}{\sqrt{100}} = 0.12$$

Hence 95% confidence limits for the population mean are

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$7.4 - (1.96 \times 0.12) \le \mu \le 7.4 + (1.96 \times 0.12)$$

$$7.4 - 0.2352 \le \mu \le 7.4 + 0.2352$$

$$7.165 \le \mu \le 7.635$$

This implies that the probability that the true value of the population mean breaking strength of the cotton threads will fall in this interval (7.165,7.635) at 95% .

Example 8.13

The mean life time of a sample of 169 light bulbs manufactured by a company is found to be 1350 hours with a standard deviation of 100 hours. Establish 90% confidence limits within which the mean life time of light bulbs is expected to lie.

Solution:

Given: n = 169, $\bar{x}$ = 1350 hours, σ = 100 hours, since the level of significance is (100-90)% =10% thus α is 0.1, hence the significant value at 10% is $Z_{\alpha/2} = 1.645$

$$S.E. = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{169}} = 7.69$$

Hence 90% confidence limits for the population mean are

$$\bar{x} - Z_{\alpha/2} SE < \mu < \bar{x} + Z_{\alpha/2} SE$$

$$1350 - (1.645 \times 7.69) \leq \mu \leq 1350 + (1.645 \text{ x } 7.69)$$

$$1337.35 \leq \mu \leq 1362.65$$

Hence the mean life time of light bulbs is expected to lie between the interval (1337.35, 1362.65)

8.3 Hypothesis Testing

One of the important areas of statistical analysis is testing of hypothesis. Often, in real life situations we require to take decisions about the population on the basis of sample information. Hypothesis testing is also referred to as "Statistical Decision Making". It employs statistical techniques to arrive at decisions in certain situations where there is an element of uncertainty on the basis of sample, whose size is fixed in advance. So statistics helps us in arriving at the criterion for such decision is known as Testing of hypothesis which was initiated by J. Neyman and E.S. Pearson.

For Example: We may like to decide on the basis of sample data whether a new vaccine is effective in curing cold, whether a new training methodology is better than the existing one, whether the new fertilizer is more productive than the earlier one and so on.

8.3.1 Meaning : Null Hypothesis and Alternative Hypothesis - Level of Significants and Type of Errors

Statistical Hypothesis

Statistical hypothesis is some assumption or statement, which may or may not be true, about a population.

There are two types of statistical hypothesis

(i) Null hypothesis (ii) Alternative hypothesis

Null Hypothesis

Definition 8.7

According to Prof. R.A.Fisher, "Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true", and it is denoted by H_0.

For example: If we want to find the population mean has a specified value μ_0, then the null hypothesis H_0 is set as follows $H_0 : \mu = \mu_0$

Alternative Hypothesis

Any hypothesis which is complementary to the null hypothesis is called as the alternative hypothesis and is usually denoted by H_1.

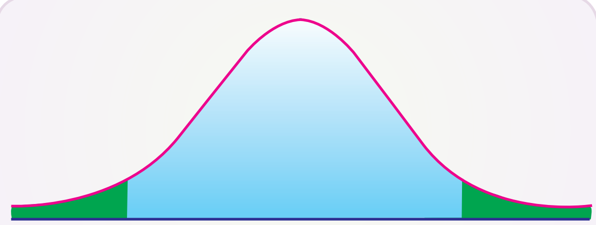
For example: If we want to test the null hypothesis that the population has specified mean μ i.e., $H_0 : \mu = \mu_0$ then the alternative hypothesis could be any one among the following:

(i) $H_1 : \mu \neq \mu_0$ ($\mu >$ or $\mu < \mu_0$)

(ii) $H_1 : \mu > \mu_0$

(iii) $H_1 : \mu < \mu_0$

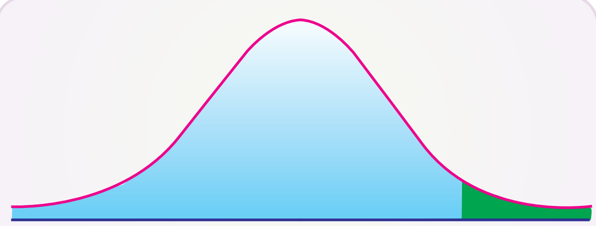
The alternative hypothesis in $H_1 : \mu \neq \mu_0$ is known as two tailed alternative test. Two tailed test is one where the hypothesis about the population parameter is rejected for the value of sample statistic falling into either tails of the sampling distribution. When the hypothesis about the population parameter is rejected only for the value of sample statistic falling into one of the tails of the sampling distribution, then it is known as one-tailed test. Here $H_1: \mu > \mu_0$ and $H_1: \mu < \mu_0$ are known as one tailed alternative.



Two tailed diagram

Fig.8.1

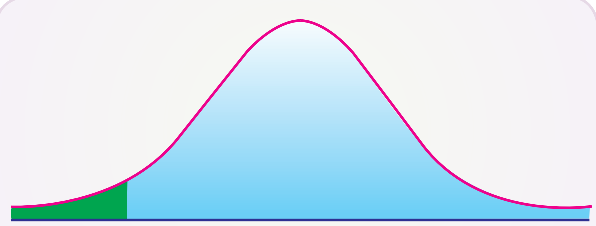
Right tailed test: $H_1: \mu > \mu_0$ is said to be right tailed test where the rejection region or critical region lies entirely on the right tail of the normal curve.



Right tailed diagram

Fig.8.2

Left tailed test: $H_1: \mu < \mu_0$ is said to be left tailed test where the critical region lies entirely on the left tail of the normal curve. (diagram)



Left tailed test

Fig.8.3

Types of Errors in Hypothesis testing

There is every chance that a decision regarding a null hypothesis may be correct or may not be correct. There are two types of errors. They are

Type I error: The error of rejecting H_0 when it is true.

Type II error: The error of accepting when H_0 it is false.

Critical region or Rejection region

A region corresponding to a test statistic in the sample space which tends to rejection of H_0 is called critical region or region of rejection.

Note

The region complementary to the critical region is called the region of acceptance.

Level of significance

The probability of type I error is known as level of significance and it is denoted by α. The level of significance is usually employed in testing of hypothesis are 5% and 1%. The level of significance is always fixed in advance before collecting the sample information.

Critical values or significant values

The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the critical value or significant value. It depend upon

(i) The level of significance

(ii) The alternative hypothesis whether it is two-tailed or single tailed.

For large samples, the standardized variable corresponding to the statistic viz.,

$$Z = \frac{t - E(t)}{\sqrt{Var(t)}} = \frac{t - E(t)}{S.E.(t)} \sim N(0,1) \quad \text{as } n \to \infty \qquad \dots (1)$$

The value of Z given by (1) under the null hypothesis is known as test statistic. The critical values of Z, commonly used at the level of significance for both two tailed and single tailed tests are given in the normal probability table (Refer the normal probably Table).

Since for large n, almost all the distributions namely, Binomial, Poisson, etc., can be approximated very closely by a normal probability curve, we use the normal test of significance for large samples.

Note

For practical purposes, the sample may be regarded as large if $n > 30$.

8.3.2 Testing Procedure : Large sample theory and test of significants for single mean

The following are the steps involved in hypothesis testing problems.

1. Null hypothesis: Set up the null hypothesis H_0
2. Alternative hypothesis: Set up the alternative hypothesis . This will enable us to decide whether we have to use two tailed test or single tailed test (right or left tailed)
3. Level of significance: Choose the appropriate level of significant (α) depending on the reliability of the estimates and permissible risk. This is to be fixed before sample is drawn. i.e., α is fixed in advance.
4. Test statistic : Compute the test statistic using
$$Z = \frac{t - E(t)}{\sqrt{Var(t)}} = \frac{t - E(t)}{S.E.(t)} \sim N(0,1) \text{ as } n \to \infty$$
5. Conclusion: We compare the computed value of Z in step 4 with the significant value or critical value or table value Z_α at the given level of significance .

   (i) If $|Z| < Z_\alpha$ i.e., if the calculated value of is less than critical value we say it is not significant. This may due to fluctuations of sampling and sample data do not provide us sufficient evidence against the null hypothesis which may therefore be accepted.

   (ii) If $|Z| > Z_\alpha$ i.e., if the calculated value of is greater than critical value Z_α then we say it is significant and the null hypothesis is rejected at level of significance α .

Test of significance for single mean

Let $x_i, (i = 1,2,3,\ldots,n)$ is a random sample of size from a normal population with mean μ and variance σ^2 then the sample mean is distributed normally with mean μ and variance $\frac{\sigma^2}{n}$, i.e., $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Thus for large samples, the standard normal variate corresponding to $\bar{x}$ is :

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Under the null hypothesis that the sample has been drawn from a population with mean and variance σ^2, i.e., there is no significant difference between the sample mean $(\bar{x})$ and the population mean (μ), the test statistic (for large samples) is:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Remark:

If the population standard deviation σ is unknown then we use its estimate provided by the sample variance given by $\hat{\sigma}^2 = s_2 \Rightarrow \hat{\sigma} = s$.

Example 8.14

An auto company decided to introduce a new six cylinder car whose mean petrol consumption is claimed to be lower than that of the existing auto engine. It was found that the mean petrol consumption for the 50 cars was 10 km per litre with a standard deviation of 3.5 km per litre. Test at 5% level of significance, whether the claim of the new car petrol consumption is 9.5 km per litre on the average is acceptable.

Solution:

Sample size n =50 Sample mean $\bar{x}$ = 10 km Sample standard deviation s = 3.5 km

Population mean μ =9.5 km

Since population SD is unknown we consider σ = s

The sample is a large sample and so we apply Z-test.

Null Hypothesis: There is no significant difference between the sample average and the company's claim, i.e., $H_0 : \mu = 9.5$

Alternative Hypothesis: There is significant difference between the sample average and the company's claim, i.e., $H_1 : \mu \neq 9.5$ (two tailed test)

The level of significance $\alpha = 5\% = 0.05$

Applying the test statistic

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1);$$

$$Z = \frac{10 - 9.5}{\frac{3.5}{\sqrt{50}}} \sim N(0,1) = \frac{0.5}{0.495} = 1.01$$

Thus the calculated value 1.01 and the significant value or table value $Z_{\alpha/2} = 1.96$

Comparing the calculated and table value ,Here $Z < Z_{\alpha/2}$ i.e., 1.01<1.96.

Inference:Since the calculated value is less than table value i.e., $Z < Z_{\alpha/2}$ at 5% level of sinificance, the null hypothesis H_0 is accepted. Hence we conclude that the company's claim that the new car petrol consumption is 9.5 km per litre is acceptable.

Example 8.15

A manufacturer of ball pens claims that a certain pen he manufactures has a mean writing life of 400 pages with a standard deviation of 20 pages. A purchasing agent selects a sample of 100 pens and puts them for test. The mean writing life for the sample was 390 pages. Should the purchasing agent reject the manufactures claim at 1% level?

Solution:

Sample size n =100, Sample mean $\bar{x}$ = 390 pages, Population mean $\mu = 400$ pages

Population SD σ = 20 pages

The sample is a large sample and so we apply Z -test

Null Hypothesis: There is no significant difference between the sample mean and the population mean of writing life of pen he manufactures, i.e., $H_0 : \mu = 400$

Alternative Hypothesis: There is significant difference between the sample mean and the population mean of writing life of pen he manufactures, i.e., $H_1 : \mu \neq 400$ (two tailed test)

The level of significance α = 1% = 0.01

Applying the test statistic

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1);$$

$$Z = \frac{390 - 400}{\frac{20}{\sqrt{100}}} = \frac{-10}{2} = -5, \therefore |Z| = 5$$

Thus the calculated value $|Z|$ = 5 and the significant value or table value $Z_{\alpha/2}$ = 2.58

Comparing the calculated and table values, we found $Z > Z_{\alpha/2}$ i.e., 5 > 2.58

Inference: Since the calculated value is greater than table value i.e., $Z > Z_{\alpha/2}$ at 1% level of significance, the null hypothesis is rejected and Therefore we concluded that $\mu \neq 400$ and the manufacturer's claim is rejected at 1% level of significance.

Example 8.16

(i) A sample of 900 members has a mean 3.4 cm and SD 2.61 cm. Is the sample taken from a large population with mean 3.25 cm. and SD 2.62 cm? (95% confidence limit)

(ii) If the population is normal and its mean is unknown, find the 95% and 98% confidence limits of true mean.

Solution:

(i) Given:

Sample size n =900, Sample mean $\bar{x}$ = 3.4 cm, Sample SD s = 2.61 cm.

Population mean μ= 3.25 cm, Population SD σ = 2.61 cm.

Null Hypothesis $H_0 : \mu = 3.25$ cm (the sample has been drawn from the population mean $\mu = 3.25$ cm and SD $\sigma = 2.61$ cm)

Alternative Hypothesis $H_1 : \mu \neq 3.25$ cm (two tail) i.e., the sample has not been drawn from the population mean $\mu = 3.25$ cm and SD σ = 2.61 cm.

The level of significance $\alpha = 5\% = 0.05$

Teststatistic:

$$Z = \frac{3.4 - 3.25}{\frac{2.61}{\sqrt{900}}} = \frac{0.15}{0.087} = 1.724$$

$\therefore Z = 1.724$

Thus the calculated and the significant value or table value $Z_{\alpha/2} = 1.96$

Comparing the calculated and table values, $Z < Z_{\alpha/2}$ i.e., 1.724 < 1.96

Inference: Since the calculated value is less than table value i.e., $Z < Z_{\alpha/2}$ at 5% level of significance, the null hypothesis is accepted. Hence we conclude that the data doesn't provide us any evidence against the null hypothesis. Therefore, the sample has been drawn from the population mean $\mu = 3.25$ cm and SD, $\sigma = 2.61$ cm.

(ii) Confidence limits

95% confidential limits for the population mean μ are :

$$\bar{x} - Z_{\alpha/2} SE \leq \mu \leq \bar{x} + Z_{\alpha/2} SE$$

$$3.4 - (1.96 \times 0.087) \leq \mu \leq 3.4 + (1.96 \times 0.087)$$

$$3.229 \leq \mu \leq 3.571$$

98% confidential limits for the population mean μ are :

$$\bar{x} - Z_{\alpha/2} SE \leq \mu \leq \bar{x} + Z_{\alpha/2} SE$$

$$3.4 - (2.33 \times 0.087) \leq \mu \leq 3.4 + (2.33 \times 0.087)$$

$$3.197 \leq \mu \leq 3.603$$

Therefore, 95% confidential limits is (3.229,3.571) and 98% confidential limits is (3.197, 3.603).

Example 8.17

The mean weekly sales of soap bars in departmental stores were 146.3 bars per store. After an advertising campaign the mean weekly sales in 400 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful at 95% confidence limit?

Solution:

Sample size	n	= 400 stores
Sample mean	$\bar{x}$	= 153.7 bars
Sample SD	s	= 17.2 bars
Population mean	m	= 146.3 bars

Since population *SD* is unknown we can consider the sample *SD* $s = \sigma$

Null Hypothesis. The advertising campaign is not successful i.e, $H_0 : \mu = 146.3$ (There is no significant difference between the mean weekly sales of soap bars in department stores before and after advertising campaign)

Alternative Hypothesis $H_1 : \mu > 143.3$ (Right tail test). The advertising campaign was successful

Level of significance $\alpha = 0.05$

Test statistic

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

$$Z = \frac{153.7 - 146.3}{\frac{17.2}{\sqrt{400}}}$$

$$= \frac{7.4}{0.86} = 8.605$$

$\therefore Z = 8.605$

Comparing the calculated value $Z = 8.605$ and the significant value or table value $Z_\alpha = 1.645$. we get 8.605 > 1.645.

Inference: Since, the calculated value is much greater than table value i.e., $Z > Z_\alpha$, it is highly significant at 5% level of significance. Hence we reject the null hypothesis H_0 and conclude that the advertising campaign was definitely successful in promoting sales.

Example 8.18

The wages of the factory workers are assumed to be normally distributed with mean and variance 25. A random sample of 50 workers gives the total wages equal to ₹ 2,550. Test the hypothesis $\mu = 52$, against the alternative hypothesis $\mu = 49$ at 1% level of significance.

Solution:

Sample size n = 50 workers

Total wages Σx = 2550

Sample mean $\bar{x} = \frac{total\,wages}{n} - \frac{\Sigma x}{n}$

$= \frac{2550}{50} = 51$ units

Population mean μ = 52

Population variance σ^2 = 25

Population SD σ = 5

Under the null hypothesis $H_0 : \mu$ = 52

Against the alternative hypothesis $H_1 : \mu \neq 52$ (Two tail)

Level of significance μ = 0.01

Test statistic $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

$$Z = \frac{51-52}{\frac{5}{\sqrt{50}}} = \frac{-1}{0.7071} = -1.4142$$

Since alternative hypothesis is of two tailed test we can take $|Z| = 1.4142$

Critical value at 1% level of significance is $Z_{\alpha/2} = 2.58$

Inference: Since the calculated value is less than table value i.e., $Z < Z_{\alpha/2}$ at 1% level of significance, the null hypothesis H_0 is accepted. Therefore, we conclude that there is no significant difference between the sample mean and population mean μ= 52 and SD σ = 5. Therefore μ = 49 is rejected.

Example 8.19

An ambulance service claims that it takes on the average 8.9 minutes to reach its destination in emergency calls. To check on this claim, the agency which licenses ambulance services has then timed on 50 emergency calls, getting a mean of 9.3 minutes with a standard deviation of 1.6 minutes. What can they conclude at 5% level of significance.

Solution:

Sample size $n = 50$

Sample mean $\bar{x}$ = 9.3 minutes

Sample S.D s = 1.6 minutes

Population mean μ = 8.9 minutes

Null hypothesis $H_0 : \mu = 8.9$

Alternative hypothesis $H_1 : \mu = 8.9$ (Two tail)

Level of significance μ = 0.05

Test statistic

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

$$Z = \frac{9.3 - 8.9}{\frac{1.6}{\sqrt{50}}} = \frac{0.4}{0.2263} = 1.7676$$

Calculated value $Z = 1.7676$

Critical value at 5% level of significance is $Z_{\alpha/2} = 1.96$

Inference: Since the calculated value is less than table value i.e., $Z < Z_{\alpha/2}$ at 5% level of significance, the null hypothesis is accepted. Therefore we conclude that an ambulance service claims on the average 8.9 minutes to reach its destination in emergency calls.

Exercise 8.2

1. Mention two branches of statistical inference?
2. What is an estimator?
3. What is an estimate?
4. What is point estimation?
5. What is interval estimation?
6. What is confidence interval?
7. What is null hypothesis? Give an example.
8. Define alternative hypothesis.
9. Define critical region.
10. Define critical value.
11. Define level of significance.
12. What is type I error.
13. What is single tailed test.
14. A sample of 100 items, draw from a universe with mean value 4 and S.D 3, has a mean value 63.5. Is the difference in the mean significant at 0.05 level of significance?
15. A sample of 400 individuals is found to have a mean height of 67.47 inches. Can it be reasonably regarded as a sample from a large population with mean height of 67.39 inches and standard deviation 1.30 inches at 0.05 level of significance?
16. The average score on a nationally administered aptitude test was 76 and the corresponding standard deviation was 8. In order to evaluate a state's education system, the scores of 100 of the state's students were randomly selected. These students had an average score of 72. Test at a significance level of 0.05 if there is a significant difference between the state scores and the national scores.
17. The mean breaking strength of cables supplied by a manufacturer is 1,800 with a standard deviation 100. By a new technique in the manufacturing process it is claimed that the breaking strength of the cables has increased. In order to test this claim a sample of 50 cables is tested. It is found that the mean breaking strength is 1,850. Can you support the claim at 0.01 level of significance.

Exercise 8.3

Choose the correct Answer

1. A ………… may be finite or infinite according as the number of observations or items in it is finite or infinite.
   (a) Population (b) census
   (c) parameter (d) none of these
2. A ……… of statistical individuals in a population is called a sample.
   (a) Infinite set (b) finite subset
   (c) finite set (d) entire set
3. A finite subset of statistical individuals in a population is called ………..
   (a) a sample (b) a population
   (c) universe (d) census
4. Any statistical measure computed from sample data is known as …………...
   (a) parameter (b) statistic
   (c) infinite measure (d) uncountable measure
5. A…………..is one where each item in the universe has an equal chance of known opportunity of being selected.
   (a) Parameter (b) random sample
   (c) statistic (d) entire data
6. A random sample is a sample selected in such a way that every item in the population has an equal chance of being included
   (a) Harper (b) Fisher
   (c) Karl Pearson (d) Dr. Yates

7. Which one of the following is probability sampling
   (a) purposive sampling
   (b) judgment sampling
   (c) simple random sampling
   (d) Convenience sampling

8. In simple random sampling from a population of N units, the probability of drawing any unit at the first draw is
   (a) $\frac{n}{N}$ (b) $\frac{1}{N}$ (c) $\frac{N}{n}$ (d) 1

9. In …………….. the heterogeneous groups are divided into homogeneous groups.
   (a) Non-probability sample
   (b) a simple random sample
   (c) a stratified random sample
   (d) systematic random sample

10. Errors in sampling are of
    (a) Two types (b) three types
    (c) four types (d) five types

11. The method of obtaining the most likely value of the population parameter using statistic is called ……………………
    (a) estimation (b) estimator
    (c) biased estimate (d) standard error

12. An estimator is a sample statistic used to estimate a
    (a) population parameter
    (b) biased estimate
    (c) sample size
    (d) census

13. …………..is a relative property, which states that one estimator is efficient relative to another.
    (a) efficiency (b) sufficiency
    (c) unbiased (d) consistency

14. If probability $P[|\hat{\theta}-\theta|<\varepsilon]\to 1$ as $n\to\infty$, for any positive ε then $\hat{\theta}$ is said to …………..estimator of θ.
    (a) efficient (b) sufficient
    (c) unbiased (d) consistent

15. An estimator is said to be ………… if it contains all the information in the data about the parameter it estimates.
    (a) efficient (b) sufficient
    (c) unbiased (d) consistent

16. An estimate of a population parameter given by two numbers between which the parameter would be expected to lie is called an……………..interval estimate of the parameter.
    (a) point estimate
    (b) interval estimation
    (c) standard error
    (d) confidence

17. A ________ is a statement or an assertion about the population parameter.
    (a) hypothesis (b) statistic
    (c) sample (d) census

18. Type I error is
    (a) Accept H_0 when it is true
    (b) Accept H_0 when it is false
    (c) Reject H_0 when it is true
    (d) Reject H_0 when it is false

19. Type II error is
    (a) Accept H_0 when it is wrong
    (b) Accept H_0 when it is true
    (c) Reject H_0 when it is true
    (d) Reject H_0 when it is false

20. The standard error of sample mean is
    (a) $\frac{\sigma}{\sqrt{2n}}$ (b) $\frac{\sigma}{n}$ (c) $\frac{\sigma}{\sqrt{n}}$ (d) $\frac{\sigma^2}{\sqrt{n}}$

Miscellaneous Problems

1. Explain the types of sampling.
2. Write short note on sampling distribution and standard error.
3. Explain the procedures of testing of hypothesis
4. Explain in detail about the test of significance for single mean
5. Determine the standard error of proportion for a random sample of 500 pineapples was

taken from a large consignment and 65 were found to be bad.

6. A sample of 100 students are drawn from a school. The mean weight and variance of the sample are 67.45 kg and 9 kg. respectively. Find (a) 95% and (b) 99% confidence intervals for estimating the mean weight of the students.

7. The mean I.Q of a sample of 1600 children was 99. Is it likely that this was a random sample from a population with mean I.Q 100 and standard deviation 15 ? (Test at 5% level of significance).

Summary

- **Sampling:** It is the procedure or process of selecting a sample from a population.
- **Population:** The group of individuals considered under study is called as population.
- **Sample :**Aselection of a group of individuals from a population.
- **Sample size :**The number of individuals included in a sample.
- **Simple Random Sampling :**The samples are selected in such a way that each and every unit in the population has an equal and independent chance of being selected as a sample.
- **Stratified Random Sampling:** When the population is heterogeneous, the population is divided into homogeneous number of sub-groups or strata. A sample is drawn from each stratum at random.
- **Systematic Sampling:** Select the first sample at random, the rest being automatically selected according to some predetermined pattern.
- **Sampling Distribution:** Sampling distribution of a statistic is the frequency distribution which is formed with various values of a statistic computed from different samples of the same size drawn from the same population.
- **Standard Error:** The standard deviation of the sampling distribution of a statistic is known as its Standard Error.
- **Statistical Inference:** To draw inference about a population of any statistical investigation from the analysis of samples drawn from that population.
- **Estimation :**The method of obtaining the most likely value of the population parameter using statistic is called estimation.
- **Point Estimation:** When a single value is used as an estimate, it is called as point estimation.
- **Interval Estimation:** An interval within which the parameter would be expected to lie is called interval estimation.
- **Test of Statistical Hypothesis:** Statistical technique to arrive at a decision in certain situations where there is an element of uncertainty on the basis of sample
- **Null Hypothesis:** The hypothesis which is tested for possible rejection under the assumption that it is true", denoted by H_0.
- **Alternative Hypothesis:**The hypothesis which is complementary to the null hypothesis is called as the alternative hypothesis, denoted by H_1.
- **Type I error:**The error of rejecting H_0 when it is true.
- Type II **error:**The error of accepting H_0 when it is false.
- **Test of significance for single mean:**

$$Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

GLOSSARY	(கலைச்சொற்கள்)
Alternative hypothesis	மாற்று கருதுகோள்
Confidence interval	நம்பிக்கை இடைவெளி
Estimation	மதிப்பிடுதல்
Interval Estimation	இடைவெளி மதிப்பீடு
Non-Sampling Errors	கூறற்றபிழை
Null hypothesis	இன்மை கருதுகோள்
Parameter	தொகுதிப் பண்பளவை
Point Estimation	புள்ளி மதிப்பீடு
Population	முழுமைத் தொகுதி
Sample	கூறு
Sample size	கூறின் அளவு
Sampling	கூறெடுப்பு
Sampling distribution	கூறு பரவல்
Sampling Errors	கூறு பிழை
Simple Random Sampling	எளிய சமவாய்ப்பு கூறெடுப்பு
Standard Error	திட்டபிழை
Statistic	கூறு பண்பளவை / மாதிரிப் பண்பளவை
Statistical Inference	புள்ளியியல் அனுமானம்
Stratified Random Sampling	படுகை சமவாய்ப்பு கூறெடுப்பு
Systematic Sampling	முறைசார் கூறெடுத்தல்