Unit-4

Definition: The Population is the aggregate or totality of statistical data forming a subject of investigation. This is also called Universe.

Example: The heights of Indians, Technical Institutions in India, Nationalised banks in India, etc..

Definition: The number of observations in the population is defined to be the size of the population. It may be finite or infinite. Size of population is denoted by 'N'.

Definition: A portion of the population which is examined with a view to determine the population characteristics is called **Sample.**

Definition: The number of objects in the sample is called size of the sample. Denoted by 'n'.

Definition: The process of selection of a sample from the population is called **Sampling.**

There are different sampling methods. Like Random sampling, Stratified sampling, Systematic sampling, Purposive sampling, Sequential sampling etc.

Random Sampling: It is the process of drawing a sample from a population in such a way that each member of the population has an equal chance of being included in the sample. The sample obtained by the process of random sampling is called **Random sample.**

If each element of a population may be selected more than once then it is called sampling with replacement whereas if the element cannot be selected more than once, it is called sampling without replacement.

If N is the size of population and n is the size of the sample, then

(i) The number of samples with replacement = N^n

(ii) The number of samples without replacement = N_{C_n}

Classification of Samples: Samples are classified in to two ways.

Large Sample: If the size of the sample (n)≥ 30, then the sample is said to be Large sample.

Small sample: If the size of the sample (n) < 30, then the sample is said to be small sample.

Thus sampling from finite population **with replacement** can be considered theoretically as sampling from **infinite population.**

Whereas in sampling **without replacement** can be considered theoretically as sampling from **finite population.**

Parameter is a statistical measures based on all the observations of a population. Like mean of population (μ), variance of population (σ^2).

Statistic is a statistical measures based only on all the units selected in a sample. Like mean of sample $(\bar{x})$, variance of sample (s^2).

Thus parameters refer to population while statistics refer to sample.

Usually statistic varies from sample to sample. But parameter remains constant.

Statistic is used to estimate the value of unknown parameter obtained from the sample.

Let $X_1, X_2, \ldots X_n$ represents a random sample of size 'n'.

Then mean of sample $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$,

Variance of sample $s^2 = \begin{cases} \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} & \text{[finite population] } i.e., \text{ sample without replacement} \\ \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n} & \text{[infinite population] i.e., sample with replacement} \end{cases}$

Let us consider a finite population of size N and let us draw all possible random samples each of size 'n'. Then we get some 'k' number of samples.

Compute a statistic 't' [where t may be mean, variance, SD etc..]

Let $t_1, t_2, \ldots t_k$ be the values of statistic 't' for the k samples.[values may vary]

Each of these values occur with a definite probability.

Now we can construct a table showing the set of 't' values with their respective probabilities.

This probability distribution of 't' is known as sampling distribution of 't'

This distribution describes how a statistic will vary from one sample to the other of the same size.

If the statistic 't' is mean , then the corresponding distribution of the statistic is known as sampling distribution of the means

Problem 1 : A population consists of five numbers 2,3,6,8 and 11 . Consider all possible samples of size two which can be drawn with replacement from the population . Find (a) the mean of the population

(b) The standard deviation of the population

(c) The mean of the sampling distribution of means and

(d) The standard deviation of the sampling distribution of means.

Solution : Given Population is {2, 3, 6, 8, 11} $\Rightarrow N = 5$

And given size of the sample is 2. i.e., $n = 2$.

(a) Mean of the population is given by $\mu = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6.$

(b) Variance of the population $(\sigma^2) = \sum \frac{(x_i - \mu)^2}{n}$

$$= \frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5}$$

$$= \frac{16+9+0+4+25}{5} = \frac{54}{5} = 10.8$$

Standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{10.8} = 3.2863$

(c) Sampling with replacement (infinite population):

The total number of samples with replacement is $N^n = 5^2 = 25.$

Now the 25 samples are $\begin{Bmatrix} (2,2),\ (2,3),\ (2,6),\ (2,8),\ (2,11), \\ (3,2),\ (3,3),\ (3,6),\ (3,8),\ (3,11), \\ (6,2),\ (6,3),\ (6,6),\ (6,8),\ (6,11), \\ (8,2),\ (8,3),\ (8,6),\ (8,8),\ (8,11), \\ (11,2),\ (11,3),\ (11,6),\ (11,8),\ (11,11) \end{Bmatrix}$

Now compute the athematic mean for each of these 25 samples.

The set of 25 samples means $\bar{x}$ of these 25 samples, gives rise to the distribution of means of the samples known as sampling distribution of means.

The samples means are $\begin{Bmatrix} 2,\ 2.5,\ 4,\ 5,\ 6.5, \\ 2.5,\ 3,\ 4.5,\ \ 5.5,\ 7, \\ 4,\ 4.5,\ 6,\ 7,\ 8.5, \\ 5,\ 5.5,\ 7,\ 8,\ 9.5 \\ 6.5,\ 7,\ 8.5,\ 9.5,\ 11 \end{Bmatrix}$

The mean of 'sampling distribution of means' is the mean of these 25 means.

Mean of sampling distribution, $\mu_{\bar{x}} = \frac{2+2.5+4+5+6.5+\cdots+11}{25} = \frac{150}{25} = 6$

Illustrating that $\mu_{\bar{x}} = \mu$(mean of population)

(d) The variance of sampling distribution of means is , $s^2 = \sum \frac{(\bar{x}_i - \mu_{\bar{x}})^2}{n}$

$$s^2 = \frac{(2-6)^2 + (2.5-6)^2 + (4-6)^2 + \cdots + (11-6)^2}{25} = \frac{135}{25} = 5.4$$

Now standard deviation is $s = \sqrt{5.4} = 2.3238$

Problem 2 : A population consists of four numbers 1,5,6,8. Consider all possible samples of size two which can be drawn without replacement from the population . Find (a) the mean of the population

(b) The standard deviation of the population

(c) The mean of the sampling distribution of means and

(d) The standard deviation of the sampling distribution of means.

(OR)

Let S = {1, 5, 6, 8}, find the probability distribution of the sample mean for random sample of size 2 drawn without replacement.

Solution : Given Population is {1, 5, 6, 8} $\Rightarrow N = 4$

And given size of the sample is 2. i.e., $n = 2$.

(a) Mean of the population is given by $\mu = \frac{1+5+6+8}{4} = \frac{20}{4} = 5.$

(b) Variance of the population $(\sigma^2) = \sum \frac{(x_i - \mu)^2}{N}$

$$= \frac{(1-5)^2+(5-5)^2+(6-5)^2+(8-5)^2}{4}$$

$$= \frac{16+0+1+9}{4} = \frac{26}{4} = 6.5$$

Standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{6.5} = 2.55$

(c) Sampling without replacement (finite population):

The total number of samples without replacement is $N_{C_n} = 4_{C_2} = 6$.

Now the 6 samples are $\{(1,5),\ (1,6),\ (1,8),\ (5,6), (5,8),\ (6,8)\}$

Now compute the athematic mean for each of these 6 samples.

The set of 6 samples means $\bar{x}$ of these 6 samples, gives rise to the distribution of means of the samples known as sampling distribution of means.

The samples means are $\{3,\ 3.5,\ 4.5,\ 5.5,\ 6.5,\ 7\}$

The mean of 'sampling distribution of means' is the mean of these 6 means.

Mean of sampling distribution, $\mu_{\bar{x}} = \frac{3+3.5+\ 4.5+\ 5.5+\ 6.5+\ 7}{6} = \frac{30}{6} = 5$

Illustrating that $\mu_{\bar{x}} = \mu$(mean of population)

(d) The variance of sampling distribution of means is , $s^2 = \sum \frac{(\bar{x}_i - \mu_{\bar{x}})^2}{n-1}$

$$s^2 = \frac{(3-5)^2 + (3.5-5)^2 + (4.5-5)^2 + (5.5-5)^2 + (6.5-5)^2 + (7-5)^2}{6-1}$$
$$= \frac{4 + 2.25 + 0.25 + 0.25 + 2.25 + 4}{5} = \frac{13}{5} = 0.26$$

Now standard deviation is $s = \sqrt{0.26} = 1.612$

Problem 3 : A population consists of four numbers 1,5,6,8. Consider all possible samples of size two which can be drawn with replacement from the population . Find (a) the mean of the population

(b) The standard deviation of the population

(c) The mean of the sampling distribution of means and

(d) The standard deviation of the sampling distribution of means.

(OR)

Let S = {1, 5, 6, 8}, find the probability distribution of the sample mean for random sample of size 2 drawn with replacement.

Solution : Given Population is {1, 5, 6, 8} $\Rightarrow N = 4$

And given size of the sample is 2. i.e., $n = 2$.

(a) Mean of the population is given by $\mu = \frac{1+5+6+8}{4} = \frac{20}{4} = 5$.

(b) Variance of the population $(\sigma^2) = \sum \frac{(x_i - \mu)^2}{N}$

$$= \frac{(1-5)^2 + (5-5)^2 + (6-5)^2 + (8-5)^2}{4}$$

$$= \frac{16+0+1+9}{4} = \frac{26}{4} = 6.5$$

Standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{6.5} = 2.55$

(c) Sampling with replacement (infinite population):

The total number of samples with replacement is $N^n = 4^2 = 16$.

Now the 16 samples are $\begin{Bmatrix}(1,\ 1),\ (1,5),\ (1,6),\ (1,8),\ (5,1),\ (5,5),\ (5,6),(5,8),\\ (6,\ 1),\ (6,5),\ (6,6),\ (6,8),\ (8,1),\ (8,\ 5),\ (8,6),(8,8)\end{Bmatrix}$

Now compute the athematic mean for each of these 16 samples.

The set of 16 samples means $\bar{x}$ of these 16 samples, gives rise to the distribution of means of the samples known as sampling distribution of means.

The samples means are $\{1,3,\ 3.5,\ 4.5,\ 3,\ 5,\ 5.5,\ 6.5,\ 3.5,\ 5.5,\ 6,7,\ 4.5,\ 6.5,\ 7,8\}$

The mean of 'sampling distribution of means' is the mean of these 16 means.

Mean of sampling distribution, $\mu_{\bar{x}}$

$$\mu_{\bar{x}} = \frac{1 + 3 + 3.5 + 4.5 + 3 + \ 5 + \ 5.5 + \ 6.5 + 3.5 + \ 5.5 + \ 6 + 7 + \ 4.5 + \ 6.5 + \ 7 + 8}{16}$$

$$= \frac{80}{16} = 5$$

Illustrating that $\mu_{\bar{x}} = \mu$(mean of population)

(d) The variance of sampling distribution of means is , $s^2 = \sum \frac{(\bar{x}_i - \mu_{\bar{x}})^2}{n}$ since Infinite population

$s^2 = \frac{(1-5)^2+(3-5)^2+(3.5-5)^2+\cdots+(8-5)^2}{16} = \frac{16+4+2.25+\cdots+9}{16} = \frac{52}{16} = 3.25$

Now standard deviation is $s = \sqrt{3.25} = 1.8027$

The Standard Error (S.E.) of a statistic is the standard deviation of the sampling distribution of the statistic.

It is used for assessing the difference between the expected value and observed value. If gives an idea about the reliability and precision of a sample.

S.E. enables us to determine the confidence limits within which the parameters are expected to lie.

It plays an important role in large sample theory and forms the basis in tests of hypothesis or tests of significance.

Standard Error of a sample mean, S.E.$(\bar{x}) = \frac{\sigma}{\sqrt{n}}$ [in case of infinite population]

[i.e., sample is drawn with replacement]

Standard Error of a sample mean, S.E.$(\bar{x}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ [in case of finite population]

[i.e., sample is drawn without replacement]

The sampling distribution of a statistic $(\bar{x})$ will be approximately NORMAL with mean μ and variance $\sigma_{\bar{x}} = \frac{\sigma^2}{n}$ provided that the sample size is large.

Standardized sample mean, $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

Problem 1. The variance of a population is 2. The size of sample collected from the population is 169. What is the standard error of mean?

Solution: We know that the standard error of the mean, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ for large samples.

Given variance of population, $\sigma^2 = 2 \Rightarrow \sigma = \sqrt{2}$

And size of the sample, $n = 169 \quad (> 30)$

Hence S.E. of mean, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{2}}{\sqrt{169}} = \frac{1.414}{13} = 0.1085$

Problem 2: What is the effect on standard error, if a sample is taken from an infinite population of sample size is increased from 400 to 900 ?

Solution: We know that standard error of the mean , $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ [infinite population]

Let $n = 400$, then S.E.$_1 = \frac{\sigma}{\sqrt{400}} = \frac{\sigma}{20}$

Similarly if $n = 900$, then S.E.$_2 = \frac{\sigma}{\sqrt{900}} = \frac{\sigma}{30} = \frac{2}{3}\left(\frac{\sigma}{20}\right) = \frac{2}{3}$(S.E.$_1$)

$\Rightarrow$ S.E.$_1 = \frac{3}{2}$(S.E.$_2$)

Thus if the sample size is increased from 400 to 900, the S.E. will be divided by $\frac{3}{2}$

Problem 3: The mean height of students in a college is 155 cm and standard deviation is 15. What is the probability that the mean height of 36 students is less than 157 cm ?

Solution: Given the population is students of college. Mean, $\mu = 155\ cm$ and standard deviation of population, $\sigma = 15$.

Size of the sample, $n = 36$

Mean of the sample, $\bar{x} = 157\ cm$

Now we have to find "probability that the mean height of sample is < 157 cm".

i.e., we have to find $P(\bar{x} < 157)$.

Let us consider the sampling distribution of mean has normal distribution.

So we find standardized normal value, $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{157-155}{15/\sqrt{36}} = \frac{6}{15} \times 2 = 0.8$

Hence $P(\bar{x} < 157) = P(z < 0.8) = 0.5 + A(0.8)$

$$= 0.5 + 0.2881 = 0.7881$$

A(0.8) will be find from Standard normal table.

An estimation is a statement made to find an unknown population parameter.

The procedure or rule to determine an unknown population parameter is called an estimator.

i.e., for example, sample mean is an estimator of population mean.

Hence an estimator is a statistic which for all practical purposes, can be used in place of unknown parameter of the population.

The estimations are two types.(i) Point Estimation (ii) Interval Estimation

If an estimate of the population parameter is given by a single value, then the estimate is called a **Point Estimation.**

If an estimate of the population parameter belongs to some interval then that interval is called **Interval Estimation.**

Confidence Interval Estimation of Parameter: In an interval estimation of the population parameter, if we can find two quantities t_1 and t_2 based on sample observations drawn from the population such that the unknown parameter is included in the interval $[t_1, t_2]$ in a specified percentage of cases, then the interval is called a **Confidence Interval** for the parameter.

The computation of confidence interval and confidence limits is based on the sampling distribution of a statistic.

We know that the sampling distribution of the sample mean of a random sample of size (> 30) drawn from a population having mean μ and S.D. σ is approximately a normal distribution with mean $\mu_{\bar{x}}$ and S.D $s = \frac{\sigma}{\sqrt{n}}$.

Therefore $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution with mean 0 and S.D. 1.

From the standard normal curve, we know that

95.45% of the area lies between ordinates $z = \pm 2$

i.e., $P(-2 < z < 2) = 0.9545$

This shows that the inequality $-2 \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq 2$ holds in 95.45% cases and it does not hold in only 4.55% cases.

Thus $-2 \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq 2 \Rightarrow -2\left(\sigma/\sqrt{n}\right) \leq \bar{x} - \mu \leq 2\left(\sigma/\sqrt{n}\right) \Rightarrow \bar{x} - \frac{2\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{2\sigma}{\sqrt{n}}$

Or $\bar{x} - 2(S.E.\ of\ \bar{x}) \leq \mu \leq \bar{x} + 2(S.E.\ of\ \bar{x}) \qquad (S.E.\ of\ \bar{x}) = \frac{\sigma}{\sqrt{n}}$

This interval is called the 95.45% confidence limits for μ. Here 2 is called confidence coefficient.

This 95.45% is called the confidence coefficient or degree of confidence.

It is denoted by $(1-\alpha)100\%$. i.e., $P(-2 < z < 2) = 0.9545 = 1 - \alpha$.

- 90% confidence limits are $\bar{x} \pm 1.64(S.E.\ of\ \bar{x})$

- 95% confidence limits are $\bar{x} \pm 1.96(S.E.\ of\ \bar{x})$
- 99% confidence limits are $\bar{x} \pm 2.58(S.E.\ of\ \bar{x})$
- 99.73% confidence limits are $\bar{x} \pm 3(S.E.\ of\ \bar{x})$

These are for large samples. For small samples the values of z changes according to degree of freedom υ.

Problem 1: In a study of an automobile insurance a random sample of 80 body repair costs had a mean of Rs.472.36 and S.D. of Rs. 62.35. If $\bar{x}$ is used as a point estimate to the true average repair costs, with what confidence we can assert that the maximum error doesn't exceed Rs. 10.

Solution: Given size of random sample, $n = 80$

The mean of random sample, $\bar{x} = 472.36$; Standard deviation, $s = 62.35$

Maximum error of estimate, $E_{max} = 10$

We have $E_{max} = z\alpha_{/2}\left(\frac{\sigma}{\sqrt{n}}\right) \Rightarrow z\alpha_{/2} = E_{max}\frac{\sqrt{n}}{\sigma} = 10\frac{\sqrt{80}}{62.35} = 1.4345$

$\Rightarrow$ The area when $z\alpha_{/2} = 1.43$ from standard normal distribution table $\frac{\alpha}{2}$ is 0.4236.

Therefore $\frac{\alpha}{2} = 0.4236 \Rightarrow \alpha = 0.8472 \Rightarrow$ Confidence $= (1-\alpha)100\% = 84.72$

Hence we are 84.72 % confidence that the maximum error is Rs. 10.

Problem 2: What is the maximum error one can expect to make with probability 0.90 when using the mean of a random sample of size 64 to estimate the mean of the population with $\sigma^2 = 2.56$?

Solution: Given $n = 64,\ \sigma^2 = 2.56,$

Probability = 0.90 i.e., confidence limit = 90%

$$1 - \alpha = 0.90 \Rightarrow \alpha = 0.10 \Rightarrow {}^{\alpha}/_{2} = 0.05$$

Now $z\alpha_{/2} = z_{0.05}(degree\ of\ freeedom\ \infty) = 1.645$ from t- table

Hence Maximum Error, $E_{max} = z\alpha_{/2}\left(\frac{\sigma}{\sqrt{n}}\right) = 1.645\left(\frac{\sqrt{2.56}}{\sqrt{64}}\right) = 0.329$

Problem 3: A random sample of size 100 has a standard deviation of 5. What can you say about the maximum error with 95% confidence ?

Solution: Given $n = 100,\ s = 5,\ 1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$

Now $z_{\frac{\alpha}{2}} = z_{0.025}(\ d.f. \infty) = 1.96$

Now Maximum error, $E_{max} = z\alpha_{/2}\left(\frac{\sigma}{\sqrt{n}}\right) = 1.96\left(\frac{5}{\sqrt{100}}\right) = 0.98$

Problem 4: The mean and standard deviation of a population are 11,795 and 14,054 respectively. What can one assert with 95% confidence about the maximum error if $\bar{x} = 11{,}795$ and $n = 50$ and also construct 95% confidence interval for the true mean.

Solution: Given mean of population, $\mu = 11795$

And Standard deviation, $\sigma = 14054$

Mean of sample, $\bar{x} = 11795$

Size of sample, $n = 50$

Confidence coefficient, $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$

From t-distribution table $z_{\alpha/2}(\nu) = z_{0.025}(\nu = \infty) = 1.96$

Now Maximum error, $E = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = 1.96\left(\frac{14054}{\sqrt{50}}\right) = 3899$

Therefore Confidence Interval $(\bar{x} - E, \bar{x} + E)$

$$= (11795 - 3899,\ 11795 + 3899) = (7896,\ 15694)$$

Problem 5: A random sample of size 81 was taken whose variance is 20.25 and mean is 32, construct 98% confidence interval.

Solution: Given $n = 81,\ s^2 = 20.25 \Rightarrow s = \sqrt{20.25} = 4.5,\ \bar{x} = 32$

and $1 - \alpha = 0.98 \Rightarrow\ \alpha = 0.02 \Rightarrow \frac{\alpha}{2} = 0.01$

Now $\underline{z_{\frac{\alpha}{2}}(\nu) = z_{0.01}(\infty) = 2.326}$

Maximum error, $E = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = 2.326\left(\frac{4.5}{\sqrt{81}}\right) = 1.163$

Therefore confidence interval= $(\bar{x} - E, \bar{x} + E) = (32 - 1.163,\ 32 + 1.163)$

$$= (30.837,\ 33.163)$$

Problem 6: Find 95% confidence limits for the mean of a normality distributed population from which the following sample was taken 15, 17, 10, 18, 16, 9, 7, 11, 13,14.

Solution: Given the sample is {15, 17, 10, 18, 16, 9, 7, 11, 13, 14}

Size of sample, $n = 10$

Mean of the sample, $\bar{x} = \frac{15+17+10+18+16+9+7+11+13+14}{10} = \frac{130}{10} = 13$

Variance of sample, $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{4+16+9+25+9+16+36+4+0+1}{9} = \frac{120}{9} = 13.33$

Given confidence coefficient, $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$

Now $z_{\frac{\alpha}{2}}(\nu) = z_{0.025}(\nu = 10 - 1) = 2.262$

Maximum Error, $E = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = 2.262\left(\frac{13.33}{\sqrt{10}}\right) = 9.535$

Hence Confidence interval $= (\bar{x} - E,\ \bar{x} + E) = (13 - 9.535,\ 13 + 9.535)$

$$= (3.465,\ 22.535)$$