Sampling of Variables

1. **Population:** Group of individuals (items) under study.

2. **Sampling:** Process of taking samples from the population.

3. **Aim of Sampling:** If the population size is large (infinite) or it is not possible to include every member of population under study.

Random Sampling: Each member of population has the equal opportunity to be included in the sample.

Parameters and Statistics:

Statistical constants of the population such as mean (μ), variance (σ^2) etc are known as **parameters.**

Statistical constants computed from the sample observation such as mean ($\overline{x}$), variance (s^2) etc to estimate the population parameters are known as **statistics.**

Test of significance

On the basis of prior experiments, surveys or experience, we propose some hypothetical value to the population characteristics or parameter and test its significance on the basis of sample result whether
(i) sample result deviate from proposed value
(ii) there is any deviation between two independent sample results
to conclude that deviation is significant or might be attributed to chance or fluctuation of sampling and then decide whether the proposed value of the parameter is to be accepted or rejected..

e.g. 1. A manufacturer claim that his product say processor is of 2.8 GB
2. Product X is better than product Y.

Null Hypothesis (H_0) : Hypothesis which is actually tested for acceptance or rejection is termed as Null Hypothesis.
This is known as hypothesis of no difference i.e. We assume that the proposed value of the parameter is true or there is no difference in the values of the parameters of two population.

Alternative Hypothesis (H_1) : Any hypothesis against the null hypothesis.

e.g. To test that population has mean μ_0

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0 \; (\mu > \mu_0 \text{ or } \mu < \mu_0)$ (**Two tail** H_1)

or $H_1 : \mu > \mu_0$ (Assuming $\mu < \mu_0$ is very rare) **one tail** H_1

or $H_1 : \mu < \mu_0$ (Assuming $\mu > \mu_0$ is very rare) **one tail** H_1

Errors in decision making:

Two type:

1. **Type I** : Reject H_0 when it is true
2. **Type II :** Accept H_0 when it is false.

e.g.1. A product is tested for certain standard of quality.
If it meet the standard but we considered it of poor standard then type I error has occurred.
If it doesn't meet the standard but we considered it of good quality then type II error has occurred.

It is not possible to minimize both type of error at the same time.
If we try to minimize one type of error then we tend to commit other type of error.
To minimize both the types error simultaneously, we need to increase sample size which may not feasible.

Type II error can be more severe than type I error.
e.g. 2. A medicine is given to a patient and examined.
If result is positive but considered as no effect therefore medicine is changed then type I error has occurred.
If result is adverse but considered as positive therefore medicine is continued then type II error has occurred.

Therefore to take decision about Null hypothesis (H_0) type II error is minimized even at certain risk of type I error.

Level of Significance (L.S.): It is the quantity of the type I error which we are ready to tolerate in making a decision about H_0
i.e. It is probability of type I error which is tolerable.

If $P\{\text{Reject } H_0 \text{ when it is true}\} = \alpha$

Then $\alpha = \text{Level of significance}$ [probability of making wrong decision]

$1-\alpha = \text{Confidence Limit}$ [probability of making correct decision]

Generally $\alpha = 1\%$ is chosen for high precision and $\alpha = 5\%$ is chosen for moderate precision.

Critical Region :

Let t be the test statistics.

If $t \in W$ leads to the rejection of H_0 then W is called **critical region**

and W^c is called the **region of acceptance**.

If $\alpha = \text{Level of significance}$ then $P\{t \in W\} = \alpha$

Two tail test: A test involving two tail Alternative hypothesis (H_1)

One tail test : A test involving one tail Alternative hypothesis (H_1)

Test of Significance for Large samples $(n \geq 20)$

1. **Test of significance between sample mean and population mean:**

Let there be a sample of size n $(n \geq 20)$ from a population with mean μ and variance σ^2

To test Null hypothesis

H_0: There is no significant difference in the sample mean $(\bar{x})$ and population mean (μ)

The test statistics is

$$z_t = \frac{|\bar{x} - \mu|}{\sqrt{\frac{\sigma^2}{n}}}$$ which follows Standard Normal distribution $N(0,1)$

therefore we apply Normal distribution test.

Note:

1. Under the two tail $H_1: \mu \neq \mu_0$ and $L.S. = \alpha$, Critical point z_α is given by

$P\{Z \leq z_\alpha\} = 1 - \frac{\alpha}{2}$

2. Under the one tail $H_1: \mu > \mu_0$ or $\mu < \mu_0$ and $L.S. = \alpha$, Critical point z_α is given by

$P\{Z \leq z_\alpha\} = 1 - \alpha$

3. H_0 is accepted iff $z_t \leq z_\alpha$

4. If population variance (σ^2) is not known then it can be estimated by its unbiased estimator i.e. $\sigma^2 = s^2$ (sample variance)

5. If population mean μ is not known, then we can obtain confidence interval (limit) of μ at $C.L. = (1-\alpha)$ i.e. at $L.S. = \alpha$ as

$$z_t \leq z_\alpha \quad \Leftrightarrow \quad \frac{|\bar{x} - \mu|}{\sqrt{\frac{\sigma^2}{n}}} \leq z_\alpha \quad \Leftrightarrow \quad \bar{x} - z_\alpha \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + z_\alpha \sqrt{\frac{\sigma^2}{n}}$$

2. **Test of significance of two sample means:**

Let there be two samples from two populations with following data:

Sample Size	Sample Mean	Population Mean	Population Variance
n_1	$\bar{x}_1$	μ_1	σ_1^2
n_2	$\bar{x}_2$	μ_2	σ_2^2

To test H_0: There is no significant difference between two population means or two samples results is same i.e. $\mu_1 = \mu_2$,

Test statistics is

$$z_t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ which follows Standard Normal distribution $N(0,1)$

Note:

1. If the two samples are taken from same population or two population variance are equal, then $\sigma_1^2 = \sigma_2^2 = \sigma^2$

2. If σ_1^2 & σ_2^2 are not known then they are estimated from the samples by using its unbiased estimator $\sigma_1^2 = s_1^2$ & $\sigma_2^2 = s_2^2$

① A random sample of 50 items gives the mean 6.2 and standard deviation 10.24. Can it be regarded as sample is drawn from a normal population with mean 5.4. Test the claim at 5% significance level.

Solⁿ H_0: Sample is drawn from a population with mean 5.4

i.e. $\mu = 5.4$

$H_1: \mu \neq 5.4$

$\therefore$ Two tail test is applied.

$n = 50, \quad \bar{x} = 6.2, \quad s = 10.24$

$n > 20$

$\therefore$ for large sample size,

$\sigma = s = 10.24$

Test statistics is

$$Z_t = \frac{|\bar{x} - \mu|}{\sqrt{\frac{\sigma^2}{n}}} = \frac{|6.2 - 5.4|}{\sqrt{\frac{(10.24)^2}{50}}}$$

$$= 0.5524$$

At L.S. = 5% i.e. $\alpha = 0.05$, and for two tail test,

critical point Z_α is given by

$P\{Z \leq Z_\alpha\} = 1 - \frac{\alpha}{2} = 0.975$

$\Rightarrow Z_\alpha = 1.96$

$\therefore Z_t < Z_\alpha \Rightarrow H_0$ is accepted

$\Rightarrow$ population mean is 5.4.

② The mean height of 50 male students who showed above average paticipation in college sports was 68.2 inches with a standard deviation of 2.5 inches while 45 male students who showed no interest in such paticipation has a mean height 67.5 inches with a standard deviation of 2.8 inches. Test the hypothesis that students who participated in college sports are taller than other male students.

Solⁿ H_0: There is no difference in the heights of students who participated in college sports and other students.

i.e. $\mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$

∴ one tail test is applied.

saples:

I: $n_1 = 50$ $\quad \bar{x}_1 = 68.2$ $\quad s_1 = 2.5$

II: $n_2 = 45$ $\quad \bar{x}_2 = 67.5$ $\quad s_2 = 2.8$

Since n_1, n_2 are large,

$\sigma_1 = s_1 = 2.5, \quad \sigma_2 = s_2 = 2.8$

The test statistics is

$$Z_t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{|68.2 - 67.5|}{\sqrt{\frac{(2.5)^2}{50} + \frac{(2.8)^2}{45}}}$$

$$= 1.2797$$

At L.S. = 5% i.e. $\alpha = 0.05$ and for one tail test, critical point is given by

$$P\{Z \leq Z_\alpha\} = 1 - \alpha = 1 - 0.05 = 0.95$$

$$\Rightarrow \quad Z_\alpha = 1.65$$

$$\Rightarrow \quad Z_t < Z_\alpha$$

$\Rightarrow$ H_0 is accepted

$\Rightarrow$ There is no difference in the heights.

Test of Significance for small samples $(n < 20)$

Degree of Freedom : No. of values in a set which may be assigned arbitrarily.
i.e. Degree of freedom (d.f.)= No. of variables involve – No. of equations used.

1. **Test of significance between sample mean and population mean:**

Let there be a random sample of size $n\ (n < 20)$ with sample mean $\overline{x}$, sample variance s^2 from a population with mean μ and population variance σ^2 which not known.

Then test statistics is

$$t = \frac{|\overline{x} - \mu|}{\sqrt{\frac{s^2}{n-1}}}$$ which follow student's t distribution with $d.f. = (n-1)$

Note:

1. The critical value (tabulated value) t_α for $\text{level of significance} = \alpha$ and $d.f. = n-1$ is given in the table at probability column $\frac{\alpha}{2}$ **for two tail test** and at probability α **for one tail test.**

2. Null hypothesis H_0 is accepted iff $t \le t_\alpha$

3. If μ is not given, then $(1-\alpha)$ confidence limit for μ is

$$t \le t_\alpha \quad \Leftrightarrow \quad \frac{|\overline{x} - \mu|}{\sqrt{\frac{s^2}{n-1}}} \le t_\alpha \quad \Leftrightarrow \quad \overline{x} - t_\alpha \sqrt{\frac{s^2}{n-1}} \le \mu \le \overline{x} + t_\alpha \sqrt{\frac{s^2}{n-1}}$$

2. Test of significance of means of two samples:

Let there be two samples from two populations with following data:

Sample Size	Sample Mean	Population Mean	Population Variance
n_1	$\bar{x}_1$	μ_1	σ_1^2
n_2	$\bar{x}_2$	μ_2	σ_2^2

where population variances σ_1^2 and σ_2^2 are unknown.

To test H_0: There is no significant difference between two sample results i.e. $\mu_1 = \mu_2$,

Assuming the two population variance are equal i.e $\sigma_1^2 = \sigma_2^2 = \sigma^2$ or two samples are taken from the same population,

Unbiased estimator of σ^2 is given by

$$\tilde{s}^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2\right) = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

and test statistic is given by

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\tilde{s}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ which follow t distribution with $(n_1 + n_2 - 2)$ degree of freedom.

3. Paired t-Test of significance of means of two samples:

If two samples are of the same size i.e. $n_1 = n_2 = n$ and two samples are dependent then we can paired the data and then test statistics is given by

$$t = \frac{|\bar{d}|}{\sqrt{\frac{s^2}{n-1}}}$$ which follow t distribution with $(n-1)$ degree of freedom.

where $d = x_1 - x_2$ $\quad \bar{d} = \text{mean}(d)$ $\quad s^2 = \text{var}(d)$

E.X.

① An efficiency expert claim that by introducing a new type of machinery into a production process, he can decrease substantially the time required for production. Because the expense involved in maintenance of the machines, management feels that unless the production time can be decreased by at least 8 %, they cannot afford to introduce the process. Six resulting experiments show that the time for productions is decreased by 8.4 % with standard deviation 0.32 %. Test the claim that process should be introduce at 1% at L.S.

Soln H_0: The new machinery into the production process will be introduced.

i.e. $\mu = 8\%$

H_1: $\mu < 8\%$

∴ one tail test is applied.

sample data:

$n = 6$, $\bar{x} = 8.4\%$, $s = 0.32\%$

n is small;

∴ Test statistics is

$$t = \frac{|\bar{x} - \mu|}{\sqrt{\frac{s^2}{n-1}}} = \frac{|8.4 - 8|}{\sqrt{\frac{(0.32)^2}{5}}}$$

$= 2.795$

which follows t-distribution with $d.f. = n-1 = 5$

At L.S. = 1% i.e. $\alpha = 0.01$,
for one tail test and d.f. = 5,
critical point is
$t_\alpha = 3.36$

$\Rightarrow t < t_\alpha$

$\Rightarrow$ H_0 is accepted

$\Rightarrow$ The new machinery into the production process will be introduced.

(2) A random sample of 10 boys has the I.Q.'s 70, 120, 110, 101, 88, 83, 95, 98, 107 and 100 respectively. Does these data supports the assumption of a population mean I.Q. of 100? Find a reasonable range in which most of the mean I.Q. values of the population lie from which sample is drawn.

$\underline{\text{Sol}^n}$ H_0: Population mean I.Q. is 100 i.e.

$\mu = 100$

H_1: $\mu \neq 100$

$\therefore$ Two tail test is applied.

Sample data:

x: 70, 120, 110, 101, 88, 83, 95, 98, 107, 100

$n = 10$ (small sample size),

$$\bar{x} = \frac{1}{n}\sum x_i = 97.2$$

$$E(x^2) = \frac{1}{n}\sum x_i^2 = 9631.2$$

$$\therefore s^2 = Var(x) = E(x^2) - (E(x))^2$$

$$= 9631.2 - (97.2)^2 = 183.36$$

$\therefore$ Test statistics is

$$t = \frac{|\bar{x} - \mu|}{\sqrt{\frac{s^2}{n-1}}} = \frac{|97.2 - 100|}{\sqrt{\frac{183.36}{9}}}$$

$$= 0.6203$$

At L.S. = 5% i.e. $\alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$,

d.f. = 9

critical point is

$t_\alpha = 2.262$

$\Rightarrow t < t_\alpha \Rightarrow H_0$ is accepted

$\Rightarrow$ Population mean I.Q. is 100.

At L.S. = 5% i.e. at C.L. = 95% (confidence limit),
i.e. $\alpha = 0.05$, d.f = 9, two tail critical point is

$$t_\alpha = 2.262$$

$\therefore$ If μ = population mean, is not known, then based on sample results, 95% confidence limit (fiducial limit) of population mean μ is

$$t \le t_\alpha$$

$$\Rightarrow \frac{|\bar{x} - \mu|}{\sqrt{\frac{s^2}{n-1}}} \le 2.262$$

$$\Rightarrow |97.2 - \mu| \le 2.262 \sqrt{\frac{183.36}{9}}$$

$$\Rightarrow |97.2 - \mu| \le 10.21$$

$$\Rightarrow 97.2 - 10.21 \le \mu \le 97.2 + 10.21$$

$$\Rightarrow 86.99 \le \mu \le 107.41$$

③ The heights of six randomly chosen sailors are (in inches) 65, 63, 68, 69, 71, 72 and those of 10 randomly chosen soldiers are 61, 62, 65, 66, 69, 69, 70, 71, 72, 73. Test whether sailors are on the average taller than soldiers. L.S. = 1%.

Soln Ho: There is no difference in the heights of sailors and soldiers i.e.

$\mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$; one tail test is applied

sample data:

x_1: 65, 63, 68, 69, 71, 72.

x_2: 61, 62, 65, 66, 69, 69, 70, 71, 72, 73

$n_1 = 6, \quad n_2 = 10$

∴ Test statistics is

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $S^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

$\bar{x}_1 = \text{mean}(x_1) = 68$

$s_1^2 = \text{Var}(x_1) = 10$

$\bar{x}_2 = 67.8$

$s_2^2 = Var(x_2) = 15.36$

$\therefore \tilde{s}^2 = \dfrac{6 \times 10 + 10 \times 15.36}{6+10-2} = 15.26$

$\therefore t = \dfrac{|68 - 67.8|}{\sqrt{15.26\left(\frac{1}{6}+\frac{1}{10}\right)}}$

$= 0.099$

At $L.S = 1\%$, $\alpha = 0.01$, $d.f = n_1 + n_2 - 2 = 14$,

$t_\alpha = 2.624$

$t < t_\alpha$

$\Rightarrow$ H_0 is accepted

$\Rightarrow$ There is no difference in the heights of sailors and soldiers.

(4) Eleven Engineering students had given a test in DBMS. They were given a month further coaching and a second test of equal difficulty was held at the end of it. Do the marks given suggest that the students have benefited by the coaching?

test 1: 23 20 19 21 18 20 18 17 23 16 19

test 2: 24 19 22 18 20 22 20 20 23 20 17

Soln: H_0: There is no difference in the performance of two tests.

i.e. $\mu_1 = \mu_2$

$H_1: \mu_1 < \mu_2$

$\therefore$ one tail test is applied.

$n_1 = n_2 = 11$ and data is dependent type

$\Rightarrow$ paired t-test is applied.

test 1 x_1	test 2 x_2	$d = x_1 - x_2$
23	24	-1
20	19	1
19	22	-3
21	18	3
18	20	-2
20	22	-2
18	20	-2
17	20	-3
23	23	0
16	20	-4
19	17	-2

$\bar{d} = \text{mean}(d)$

$= -1.364$

$s^2 = \text{var}(d)$

$= 3.686$

Test statistics is

$$t = \frac{|\bar{d}|}{\sqrt{\frac{s^2}{n-1}}} = \frac{|-1.364|}{\sqrt{\frac{3.686}{10}}}$$

$$= 2.247$$

At L.S. = 5% i.e. $\alpha = 0.05$, d.f. = 10,

$t_\alpha = 1.812$

$\Rightarrow t > t_\alpha$

$\Rightarrow$ Ho is rejected

$\Rightarrow$ Students are benefited from the coaching.

(5) The means of two random samples of sizes 9 and 7 are 196.42 and 198.82 respectively. The sum of the squares of the deviation from the means are 26.94 and 18.73 respectively. Can the samples be considered to have been drawn from the same population?

Soln Ho: The two samples are considered to be drawn from the same population.

i.e. $\mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$\therefore$ Two tail test is applied.

sample data;

I : $n_1 = 9$, $\bar{x}_1 = 196.42$,

$$\Sigma (x_1 - \bar{x}_1)^2 = 26.94$$

II : $n_2 = 7$, $\bar{x}_2 = 198.82$,

$$\Sigma (x_2 - \bar{x}_2)^2 = 18.73$$

The test statistics is

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\tilde{s}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\tilde{s}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{(n_1 + n_2 - 2)}$$

$$s_1^2 = \frac{1}{n_1} \Sigma (x_1 - \bar{x}_1)^2 \Rightarrow n_1 s_1^2 = 26.94$$

$$s_2^2 = \frac{1}{n_2} \Sigma (x_2 - \bar{x}_2)^2 \Rightarrow n_2 s_2^2 = 18.73$$

$$\therefore \tilde{s}^2 = \frac{26.94 + 18.73}{9 + 7 - 2} = 3.262$$

$$\therefore t = \frac{|196.42 - 198.82|}{\sqrt{3.262\left(\frac{1}{9} + \frac{1}{7}\right)}} = 2.637$$

At L.S = 5% i.e. $\alpha = 0.05 \Rightarrow \alpha/2 = 0.025$,

d.f $= n_1 + n_2 - 2 = 14$, $t_\alpha = 2.145 \Rightarrow$ Ho is rejected.