SAMPLING THEORY

10.1 INTRODUCTION

Before giving the notation of sampling we first define the term population. The group of individuals under study is known as **population** or **universe**. Thus in statistics, population in an average of objects, animate or inanimate, under study, we may speak of the populations of weights, heights of the students in a school, mileages of automobile tyres etc. The population may be finite or infinite.

A population containing finite number of individuals or members is called a **finite population**. For instance, the population of mechanical engineering students in a college. And a population with

infinite number of individuals is called **infinite population**. For instance, the population of pressures at different points in the atmosphere. Here we discuss the procedures for using sample information to draw inferences about uncertain population.

10.2 SAMPLING

A small section selected from the population is called a **sample**. For instance, a housewife tests a small quantity of rice to see that it has been cooked or not. This small quantity is a sample and represents the entire quantity of rice cooked. The number of individuals included in the sample is known as the **size of the sample** and process of selecting a sample from the population is called **sampling**. The population of heads and tails obtained by tossing a coin an infinite no. of times is known as **hypothetical population**. A random sample in one in which each member of population has an equal chance of being including in it.

The statistical containts of the population such as mean (μ), standard deviation (σ) etc. are called **parameters**, similarly constants for the sample drown from the given population is known as **statistics**.

The aim of the theory of sampling is gathering the maximum information about the population with the minimum effort. The object of sampling studies is to obtain the best possible values of the parameters under specific conditions. The logic of the sampling theory is the logic of induction. In induction we pass from a particular (Sample) to general (Population). This types of generalisation here is known as **Statistical inference**. The conclusion in the sampling studies are based not on certainities but on probabilities.

The fundamental assumption underlying most of the theory of sampling is **random sampling** which consists in selecting the individuals from the population in such a way that each individual of the population has the same chance of being selected.

10.3 SAMPLING DISTRIBUTION

Suppose all possible samples of size n which can be drown from a given population at random. Find out the means of each sample, the means of the samples are unequal. The means with their respective frequencies are grouped. The frequency distribution so formed is called as **sampling distribution** of the mean. Similarly, sampling distribution of standard deviation we can have.

10.4 STANDARD ERROR (S.E.)

The standard deviation of the sampling distribution of a statistics is known as its **standard error,** abbreviated as S.E. The standard error is used to assess the difference between expected values and observed values, and the reciprocal of the standard error is known as **precision**. If $n < 30$, a sample is called small, while $n \geq 30$ is called **large sample**. The sampling distribution of large samples is assumed to be normal.

If sample size (n) is increased then variance of ($\bar{x}$), $\left(\frac{\sigma}{\sqrt{n}}\right)$ of $\bar{x}$ is also called **standard error of the mean** and it is denoted by $\sigma_{\bar{x}}$. Sampling from normal population is defined as :

If $$x \sim N\left(\mu, \sigma^2\right) \Rightarrow \bar{x} \sim n\left(x, \frac{\sigma^2}{n}\right)$$

10.6 TESTING A HYPOTHESIS

The testing of hypothesis is a procedure that helps us to ascertain the likelihood of hypothesised population parameter being correct by making use of the sample statistic i.e. It is a process of testing of significance which concerns with the testing of same hypothesis regarding a parameter of the population on the basis of statistic from the sample. The test of hypothesis discloses the fact whether the difference between sample statistic and corresponding hypothesical population parameter is significant or not significant. Thus the test of hypothesis is also known as the **test of significance**.

10.6.1 Null Hypothesis (H_0)

The statistical hypothesis that is set up for a testing a hypothesis is called a null hypothesis. The null hypothesis is set up in testing a statistical hypothesis only to decide whether to accept or reject the null hypothesis. It asserts that there is no difference between the sample statistic and population parameter and whatever difference is there, is attributable to sampling errors. Null hypothesis is usually denoted by H_0.

Prof. R.A. fisher remarked "**Null hypothesis is the hypothesis which is to be tasted for possible rejection under the assumption it is true**".

10.6.2 Alternative Hypothesis (H_1)

Any hypothesis which is not a null hypothesis is called an **Alternative Hypothesis,** and it is denoted by $\mathbf{H_1}$ **or** H_α.

It is set in such a way that the rejection of null hypothesis implies the acceptance of alternative hypothesis.

10.6.3 Error

In sampling theory to draw valid inferences about the population parameter on the basis of the sample results. We decide to reject or to accept the lot after examining a sample from it.

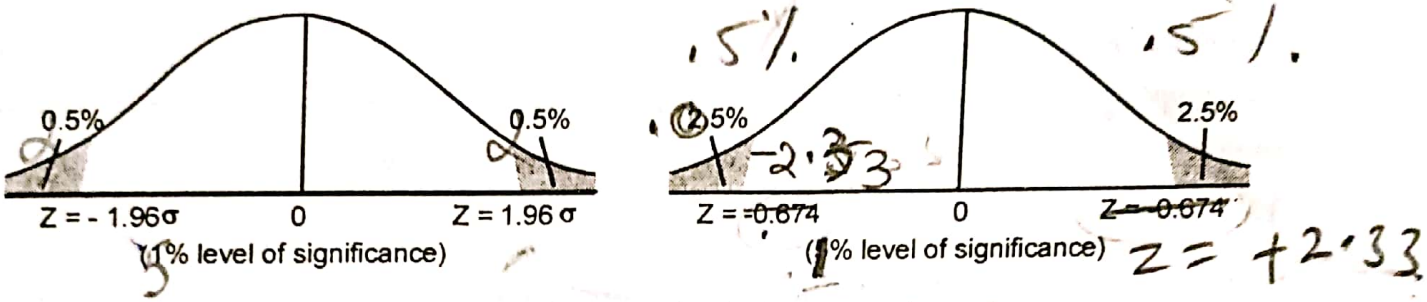
There are two types of errors:

(i) If H_0 is rejected while it should have been accepted, then a **Type 1 Error** is made

(ii) If H_0 is accepted while it should have been rejected, then a **Type II Error** is made.

10.6.4 Level of Significance

The probability level below which we reject the hypothesis is called the **level of significance,** ad the region in which a sample value falling is rejected is called the **critical region**. The commonly used level of significance in practice are 5% (0.05) and 1% (0.01). The shaded portions are the critical regions. Thus the probability of the value of the variate falling in the critical region is the level of significance.



If the variate falls in the critical area, the hypothesis is to be rejected.

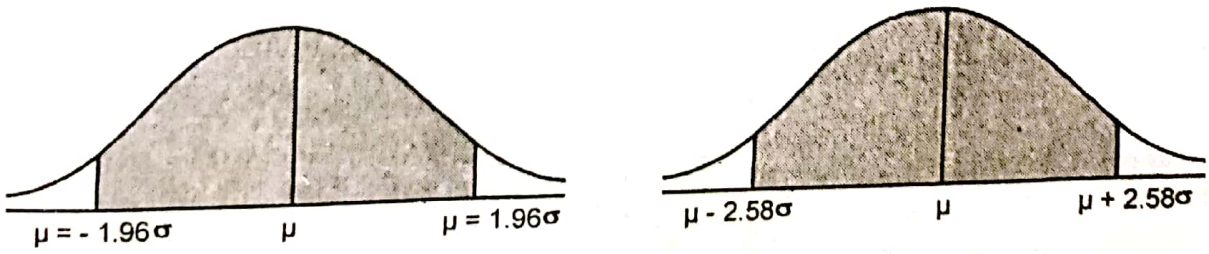
10.6.5 Test of Significance

The procedure which enable us to decide whether to accept or reject the hypothesis is called the test of significance. If the difference between the sample value and population values are so large, it is to be rejected and these difference are so small as to account for fluctuations of sampling.

10.6.6 Confidence Limits (or fiducial limit)

The Americal statistican J. Neyman (1804-1981) developed the modern theory and terminology of confidence limits. Let the sampling distribution of statistic S is normal with mean μ and standard deviation σ. If a sample statistics lies in the interval μ-1.96 σ, μ + 1.96 σ, we call 95% canfidence interval :

Similarly μ-2.58 σ, μ + 2.58 σ is 99% confidence limits as the area between μ - 2.58 σ and μ + 2.58 σ is 99%.



10.7 LARGE SAMPLE - TESTS OF SIGNIFICANCE

In this section, we will discuss the tests of significance when samples are large. Suppose a large no. n of independent bernoullian trials is performed and x successes are obtain. We wish to test

the hypothesis that the probability of success in each trial is p. we also assume that the hypothesis to be correct, and np and npq represent the mean and variance of the sampling distribution therefore for large n,

$$Z = \frac{x - np}{\sqrt{npq}}$$

is distributed as a standard normal variate, Hence from the table ( ), we have $P(|Z| > 3) = 0.0027$

Therefore, in random sampling such a result is extremely unlikely and we conclude that the truth of the hypothesis is itself very improbable and we say that the difference between the observed value of successes x and the expected value of successes np is highly significant.

Than we have following test of signifiacne.

(1) If $|Z| < 1.96$ difference between the observed and expected value of successes is not significant.

(2) If $|Z| > 1.96$, difference is significant at 5% level of significant.

(3) If $|Z| > 2.58$, difference is significant at 1% level of significant.

10.8 COMPARISON OF LARGE SAMPLE

Two large samples of sizes n_1, n_2 be drawn from two populations of proportions of attributes A's as P_1, P_2 respectively.

(i) **Hypothesis :** As regards the attribute A, we combine the two samples to find an estimate of the common value of propertion of A's in the population which is given by

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

If e_1, e_2 be the standard errors in the two samples, then

$$e_1^2 = \frac{pq}{n_1} \text{ and } e_2^2 = \frac{pq}{n_2}$$

and If e be the standard error A the difference between P_1 and P_2 then

$$e^2 = e_1^2 + e_2^2 = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

$$\therefore \quad Z = \frac{P_1 - P_2}{e}$$

(a) If $Z > 3$, the differen between P_1 *and* P_2 are real.

(b) If $Z < 2$, the difference may be due to fluctuations of sampling.

(c) If $2 < Z < 3$, the difference is significant at 5% level of significant.

(2) Hypothes : If the propertions of A's are not the same in the two populations from which the samples are drawn, then standard error e of the difference $P_1 - P_2$ is

$$e^2 = P_1 + P_2$$

$$= \frac{p_1 q_2}{n_2} + \frac{p_2 q_2}{n_2}, \quad Z = \frac{p_1 - p_2}{e} < 3,$$

the difference is due to fluctuation of samples.

10.9 SAMPLING DISTRIBUTION OF DIFFERENCES OF MEANS

Let $\bar{x}_1$ and $\bar{x}_2$ be the mean of two random sample of size n_1 and n_2 from population means μ_1 and μ_2 and variance σ_1^2 and σ_2^2. Then consider the samples of large.

$$\bar{x}_1 \sim \mu\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \text{ and } \bar{x}_2 \sim \mu\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

Therefore the differences $\bar{x}_1 - \bar{x}_2$ is a normal variate then

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2$$

if $\mu_1 = \mu_2$

$$\mu_1 = \mu_1 = 0$$

Variance, $\text{Var}(\bar{x}_1 - \bar{x}_2) = Var(\bar{x}_1) - Var(\bar{x}_2)$

$$Var(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

and Test of statistic Z is given

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad [\because \bar{x}_1, \bar{x}_2 \text{ are independent}]$$

Case-I: If the samples are drawn from the sample population, i.e. $\sigma_1 = \sigma_2 = \sigma$

$$\because \quad Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Case-II: If σ_1 and σ_2 are not known and $\sigma_1 \neq \sigma_2$ then σ_1 and σ_2 can be approximated standard deviation S_1 and S_2 than

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Case-III: If σ_1 and σ_2 are not known then $\sigma_1 = \sigma_2 = \sigma$ is approximated by $\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$

Hence
$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{n_1 s_1^2 + n_2 s_1^2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1\ n_2}\right)}}$$

$$= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_1}}}$$

10.9.1 Sampling Distribution of Proportions

We consider the two proportions of success are P_1 and P_2 in two large sample of size n_1 and n_2 respectively and let x_1 and x_2 be the number of individual possessing in the two samples.

$$P_1 = \frac{x_1}{n_1} \quad \text{and} \quad P_2 = \frac{x_1}{n_2}$$

Then
$$E(x_1) = n_1 p_1\ ,\ E(x_2) = n_2 p_2$$

$$E(P_1) = P_1$$

Variance $\text{Var}(x_1) = n_1 p_1 q_1$

$$\text{Var}(p_1) = \frac{p_1 q_1}{n_1} \quad [\because q_1 = 1 - p_1]$$

$$E(P_2) = P_2 \text{ and } \text{Var}(P_2) = \frac{p_2 q_2}{n_2}$$

Let $p_1 - p_2$ be a linear combination of two normal variables

then $E(p_1 - p_2) = E(p_1)\, E(p_2) = p_1 - p_2$

and $\text{Var}(p_1 - p_2) = \text{Var}(p_1) + \text{Var}(p_2)$

$$= \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$ [When two samples are independent]

Therefore the test statistic Z is given

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim \mu(0, 1)$$

Case-I: We can check Hypothesis test

if the difference $(p_1 - p_2)$ is a real differece between the two population $P_1 - P_2 = P$

$$Z = \frac{P}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad q = 1 - p$$

Now we test of significance with the aid of the normal curve. If

(i) $|z| < 1.96$ then hypothesis is acceptable at 95% levels.

(ii) at 5% level of significance than $1.96 < z < 2.58$.

(iii) at 1% level of significance then $2.58 < z < 3$

(iv) if different is highly significant i.e. $z > 3$, then hypothesis is not acceptable.

if unbiassed estimate of p

$$E\left\{\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}\right\} = \frac{1}{n_1 + n_2} E(n_1 p_1 + n_2 p_2)$$

Example 1. A Random sample of 400 flower stems has an average length of 10 cm. Can this be regarded as a sample from a large population with mean of 10.2 cm and a standard deviation of 2.25 cm ?

Sol. Here $n = 400$, mean. $\bar{x} = 10$

(i) **Null Hypothesis H_0 :** The sample has been drawn from the normal population with mean $\mu = 10.2$ cm. and standard deviation $\sigma = 2.25\,cm$

Alternative hypothesis, $H_1 : \mu \neq 10.2$

(ii) **Test statistic :**

$$Z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{10 - 10.2}{2.25/\sqrt{400}} = \frac{-0.2 \times 20}{2.25}$$

$$= -\frac{4}{2.25} = -1.777$$

(iii) **Critical value :** The critical value of Z at 5% level of significance is ∓ 1.96

Since the statistic value $|z| = 1.777$ is less than critical value of $z = 1.96$. it falls in the acceptance region. Hence the facts are consistent with the null hypothesis which is accepted with 95% confidence and it is concluded that the sample has been drawn from the normal population with mean of 10.2 cm and a standard deviation of 2.25 cm.

Example 2. A coin is tossed 410 times and it turns up head 224 times. Find whether the coin may be regarded as unblased one.

Sol. Let x be the number of heads and P be the probability of getting head in a toss.

Set the Hypothesis : coin is unbiased

$$P = \frac{1}{2},\ n = 410 \text{ and } x = 224$$

$$\therefore \quad Z = \frac{x - \mu}{\sqrt{npq}} = \frac{x - np}{\sqrt{npq}}$$

$$Z = \frac{224 - 410 \times \frac{1}{2}}{\sqrt{410 \times \frac{1}{2} \times \frac{1}{2}}} = \frac{19}{10.124} = 1.876$$

$$Z = 1.876 < 1.96$$

Hence hypothesis be correct and critical value of Z at 5% level of significance is ∓ 1.96 and the coin may be regarded as unbiased.

Example 3. A cubical die is thrown 9000 times and a through 4 or a 5 is observed 3240 times. Show that the die cannot be regarded as an unbiased die ?

Sol. Hypothesis of an unbiassed die i.e. $P = \frac{2}{6} = \frac{1}{3}$

$$q = 1 - \frac{1}{3} = \frac{2}{3}$$

mean $$np = 9000 \times \frac{1}{3} = 3000$$

Also standard error (S.E.) of number of success = $\sqrt{npq}$

$$= \sqrt{9000 \times \frac{1}{3} \times \frac{2}{3}} = \sqrt{2000}$$

$$= 44.721$$

Test statistic $$Z = \frac{x - np}{\sqrt{npq}}$$

$$Z = \frac{3240 - 3000}{44.721} = 5.37$$

$|Z| > Z_\alpha$ [$\because z_2 = 1.96$ at 5% level of significance]

Therefore, the difference between x and np is significant i.e. H_0 is rejected that is the die can not be regarded as unbiased i.e. die is certainly biased and $p \neq \frac{1}{3} = H_1$.

Example 4. In a large city X of a random sample of 800 school boys had a slight physical defect in another large city Y, 18% of a random sample 1600 school boys had the same defect. Is the difference between the proportions significant.

Sol. Here Given $p_1 = 30\% = 0.3, p_2 = 18\% = 0.18$

Null Hypothesis H_0 : $p_1 = p_2$

Alternative hypothesis H_1 : $p_1 \neq p_2$

test of statistic $$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where $$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{800 \times 0.3 + 1600 \times 0.18}{800 + 1600}$$

$$P = 0.22,\ q = 1 - 0.22 = 0.78$$

Then $$Z = \frac{0.3 - 0.18}{\sqrt{0.22 \times 0.78\left(\frac{1}{800} + \frac{1}{1600}\right)}} = \frac{0.12}{\sqrt{0.257}}$$

$$Z = \frac{0.12}{0.5073} = 0.236$$

if take $Z_\alpha = 1.96$ at 5% level of significance then

$$|Z| = 0.236 < 1.96$$

$$|Z| < Z_\alpha$$

Therefore the difference between p_1 and p_2 is not significant at 5% level.

Example 5. Before an increase in excise duty on tea, 900 people out of a sample of 1200 were consumers of tea. After the increase in duty. 800 people were consumers of tea in a sample of 1400 persons. Find whether there is significant decrease in the consumption of tea after the increase in duty.

Sol. Let P_1 and P_2 be sampling proportions of the consumers before and after the increase in duty. Then

$$P_1 = \frac{900}{1200} = \frac{3}{4} \text{ and } P_2 = \frac{800}{1400} = \frac{4}{7}$$

Since we know that null hypothesis $H_0 : P_1 = P_2$

Alternative Hypothesis $H_1 : P_1 > P_2$

Then $$Z = \frac{P_1 - P_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{900 + 800}{1200 + 1400} = \frac{1700}{2600}$

$P = 0.653, q = 0.346$

$$Z = \frac{0.75 - 0.57}{\sqrt{0.653 \times .346\left(\frac{1}{1200} + \frac{1}{1400}\right)}} = \frac{0.18}{\sqrt{0.225 \times .0015}}$$

$Z = 9.83$

$|z| > 1.645$ at 5% level of significance

$|z| > 2.33$ at 1% level of significance

Hence the difference of P_1 and P_2 is highly significant of 5% and 1% of level. Hence therefore H_0 is rejected and H_1 is accepted.

Example 6. A sample of 100 students is taken from a large population. The mean height of the students in this sample is 180 cm. Can it be reasonably regarded that in the population, the mean height is 175 cm. and the standard deviation is 10 cm.

Sol. Here Given $\bar{x} = 180 cm., n = 100$, $\mu = 175$ and $\sigma = 10$

Statistic test $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{180 - 175}{10 / \sqrt{100}} = 5$

$\therefore$ $|Z| > Z_\alpha (1.645)$ at 5% level of significance

The difference is highly then null hypothesis H_0 is rejected here it is not statistically correct to assume that $\mu = 175 cm.$

Example 7. In a sample of 700 men from a certain large city, 500 are found to be smokers. In one of 1000 from another large city, 500 are smokers. Do the data indicate that the cities are significantly w.r.t. prevalence of smoking among men ?

Sol. Here given $P_1 = \frac{500}{700} = \frac{5}{7}$ and $P_2 = \frac{500}{1000} = \frac{1}{2}$

$n_1 = 700, n_2 = 1000$

then $$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where $P = \frac{n_1P_1 + n_2P_2}{n_1 + n_2} = \frac{500 + 500}{700 + 1000} = \frac{1000}{1700}$

$$P = \frac{10}{17} = 0.58 \quad q = 1 - p = 0.41$$

$$Z = \frac{0.71 - 0.50}{\sqrt{0.58 \times 41\left(\frac{1}{700} + \frac{1}{1000}\right)}} = \frac{0.21}{\sqrt{.2378 \times .0024}}$$

$$|Z| = \frac{0.21}{\sqrt{.00057}} = \frac{.021}{.0238} = 8.82$$

$|z| > 1.645$ at 5% level of significance. Thus the difference is highly significant and hence the two cities are significantly different w.r. to prevalance of smoking habit among men.

Example 8. Test the significance of the difference between the means of the samples, drawn from two normal population with the same standard deviation using the following data :

	Size	Mean	S.D.
Sample 1	**110**	**62**	**4**
Sample 2	**205**	**64**	**6**

Sol. Here $n_1 = 110, n_2 = 205$ $\sigma_1 = 4$, and $\sigma_2 = 6$, $\bar{x}_1 = 62$, $\bar{x}_2 = 64$

Null hypothesis

$H_0 : \bar{x}_1 = \bar{x}_2$ or $\mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$ statistic value

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_2} + \frac{\sigma_2^2}{n_1}}}$$

$$Z = \frac{62 - 64}{\sqrt{\frac{16}{205} + \frac{36}{110}}} = \frac{-2}{0.633}$$

Therefore $|Z| > Z_\alpha, Z_\alpha = 1.96$ at 5% level of ignificance

Hence the difference is highly signicicant at 5% level

i.e. H_0 is rejected and H_1 is accepted.

Therefore, the two normal populations, from which the samples are drawn, may not have the same mean, though they may have the same SD.

Example 9. Let $x_1, x_2, ...x_n$ be a random sample from a normal population $N(\mu,1)$ Show that $\theta = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ is an unbiased estimator of $\mu^2 + 1$

Sol. : Here given mean $E(x_i) = \mu$ and $V(x_i) = 1 \quad \forall i = 1, 2, ...n$

$$V(x_i) = E(x_i^2) \pm [E(x_i)]^2$$

$$E(x_i^2) = V(x_i) + \mu^2$$

$$E(x_i^2) = 1 + \mu^2$$

Now $$E(\theta) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(x_i^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(1+\mu^2) = \frac{1}{n}(1+\mu^2)n$$

$$E(\theta) = 1 + \mu^2$$

As $E(\theta) = 1 + \mu^2$ hence 'θ' is an unbiased estimater of $\mu^2 + 1$.

Example 10. The standard deviation of a random sample of 1000 is found to be 2.7 and the standard deviation of another random sample of 500 is 2.8. Assuming the samples to be independent. Find whether the two samples could have came from populations with the same SD.

Sol. Here Given $n_1 = 1000$, $\sigma_1 = 2.7$, $n_2 = 500, \sigma_2 = 2.8$

$$H_0 ; \sigma_1 = \sigma_2$$

$$H_1 ; \sigma_1 \neq \sigma_2$$

Let level of significance be 5% therefore $Z_\alpha = 1.96$

Statistic $$Z = \frac{\sigma_1 - \sigma_2}{\sqrt{\frac{\sigma_1^2}{2n_2} + \frac{\sigma_2^2}{2n_1}}}$$

$$= \frac{2.7 - 2.8}{\sqrt{\frac{(2.7)^2}{1000} + \frac{(2.8)^2}{2000}}}$$

$$= \frac{-0.1}{\sqrt{0.01121}} = -\frac{0.1}{.1058}$$

$$Z = -0.94$$

$$|Z| = 0.94$$

Therefore $|Z| < Z_\alpha$

Hence differente between σ_1 and σ_2 is not significant at 5% level.

i.e. H_0 is accepted, H_1 is rejected.

Example 11. In two large populations, there are 30% and 25% respectively of block haired people. Is this difference likely to be hidden in samples of sizes 1200 and 900 respectively drawn from the two population ?

Sol. There $n_1 = 1200,\ P_1 = 30\% = 0.30$

$n_2 = 900,\ P_2 = 25\% = 0.25$

$q_1 = 1 - P_1 = 1 - 0.3 = 0.7$

and $q_2 = 1 - P_2 = 1 - .25 = 0.75$

$H_0 : P_1 = P_2$

$H_1 : P_1 \neq P_2 \qquad Z = \dfrac{P_1 - P_2}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}} \sim N(0,1)$

$$= \frac{0.30 - 0.25}{\sqrt{\dfrac{.30 \times .70}{1200} + \dfrac{.25 \times .75}{900}}} = 2.56$$

$|Z| = 2.56$

$|z| > z_\alpha = 1.96$ at 5% level of significance

$$|z| < z_\alpha = \begin{cases} 2.58 \text{ at } 1\% \text{ LOS} \\ 2.33 \text{ at } 2\% \text{ LOS} \end{cases}$$

Hence the null Hypothesis is rejected at 5% level and we conclude that these samples will reveal the difference in the populations. But H_0 may be accepted at 1% and 2% levels.

		Significance level					
Degrees of freedom	*Two-tailed test:*	10%	5%	2%	1%	0.2%	0.1%
	One-tailed test:	5%	2.5%	1%	0.5%	0.1%	0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
30		1.697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38		1.686	2.024	2.429	2.712	3.319	3.566
40		1.684	2.021	2.423	2.704	3.307	3.551
42		1.682	2.018	2.418	2.698	3.296	3.538
44		1.680	2.015	2.414	2.692	3.286	3.526
46		1.679	2.013	2.410	2.687	3.277	3.515
48		1.677	2.011	2.407	2.682	3.269	3.505
50		1.676	2.009	2.403	2.678	3.261	3.496
60		1.671	2.000	2.390	2.660	3.232	3.460
70		1.667	1.994	2.381	2.648	3.211	3.435
80		1.664	1.990	2.374	2.639	3.195	3.416
90		1.662	1.987	2.368	2.632	3.183	3.402
100		1.660	1.984	2.364	2.626	3.174	3.390
120		1.658	1.980	2.358	2.617	3.160	3.373
150		1.655	1.976	2.351	2.609	3.145	3.357
200		1.653	1.972	2.345	2.601	3.131	3.340
300		1.650	1.968	2.339	2.592	3.118	3.323
400		1.649	1.966	2.336	2.588	3.111	3.315
500		1.648	1.965	2.334	2.586	3.107	3.310
600		1.647	1.964	2.333	2.584	3.104	3.307
∞		1.645	1.960	2.326	2.576	3.090	3.291

(iv) The figure below shows that nature of f-distribution for different combination of degree of freedom.

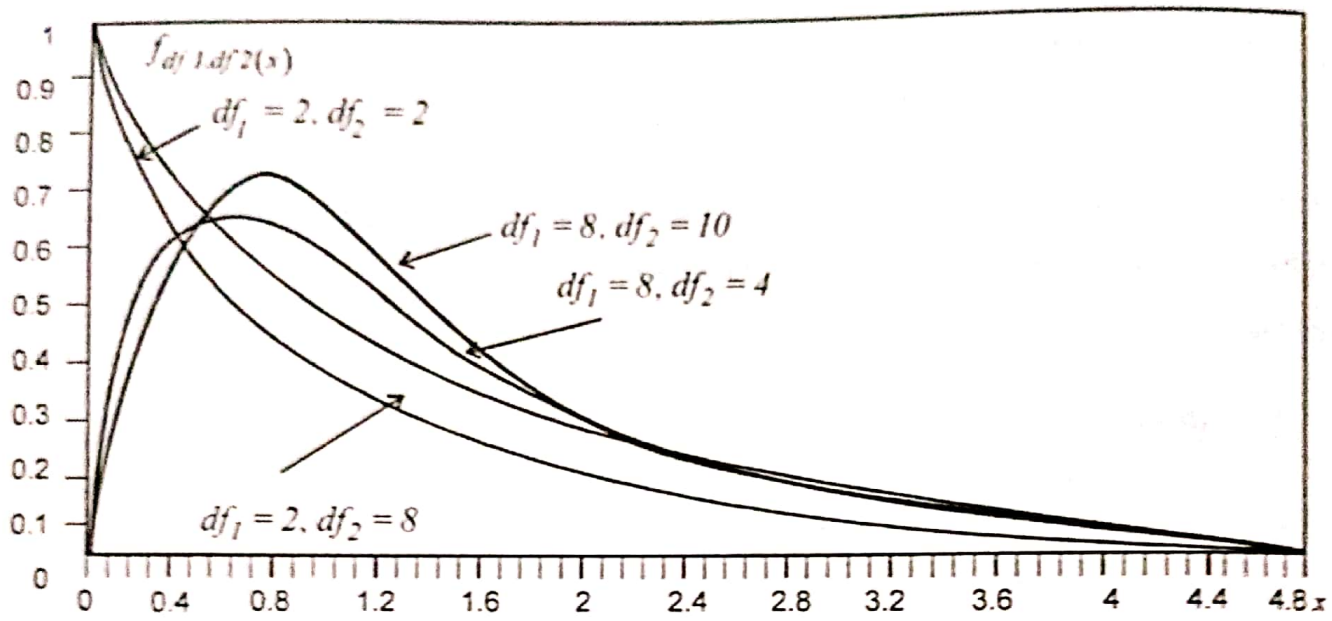


Fig. : F density functions for several values of the degrees of freedom, df_1, df_2

10.15 HYPOTHESIS TESTING

10.15.1 t-Test

10.15.1.1 Testing for a Population Mean (When Sample is Small and Variance is Unknown)

The Z-test statistic is used when the population is normally distributed or when the sample sizes are greater than 30. This is because when the population is normally distributed and σ is known, the sample means will be normally distributed, based on the Central Limit Theorem.

But what happen, if the sample being analyzed is small $(n \le 30)$, in this situation the Z test statistic would not be appropriate; and the t test comes into a picture. The formula for the t test resembles the one for the Z test but the tables used to compute the values for Z and t are different. Before σ is unknown, it will be replaced by s, the sample standard deviation.

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

The t-distribution is a symmetric distribution and looks very similar to a normal distribution. As $n \to \infty$, the t distribution becomes the normal distribution.

10.15.1.2 The t Confidence Interval on μ

If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2, a $100(1-\alpha)\%$ confidence interval of μ is given by the probability formula.

$$P\left(-t_{\alpha/2,n-1} \le t_{\alpha/2,n-1}\right) = 1-\alpha$$

$$P\left(-t_{\alpha/2,n-1} \le \frac{\bar{X}-\mu}{s/\sqrt{n}} \le t_{\alpha/2,n-1}\right) = 1-\alpha$$

Rearranging the last equation gives

$$P\left(\bar{X} - t_{\alpha/2,n-1} \cdot s/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2,n-1} \cdot s/\sqrt{n}\right) = 1-\alpha$$

where $t_{\alpha/2,n-1}$ is the upper $100.\alpha/2$ percent point of the t distribution with n-1 degrees of freedom.

Example 1. A certain car is rated at 30 mpg. On five trips, each with a tank-full of gas, the miles per gallon were 28.4, 29.2, 30.9, 29.8, and 28.6. Is the rating justified ?

Sol. From the data, $n = 5$, $\bar{X} = 29.4$, *and* $s = 1.0$. The hypothesis being tested are $H_0 : \mu = 30.0$ versus $H_1 : \mu \neq 30.0$. The computed t value is

$$t = \frac{29.4 - 30}{1/\sqrt{5}} = -1.34$$

For $\alpha = 0.05$, from the t-table $t_{\alpha/2,n-1} = 2.776$. Since the computed value of t is between -2.776 and + 2.776, the rating of 30 mpg cannot be rejected.

Confidence Interval :

The $100\,(1-\alpha)\%$ confidence interval for μ is $29.4 \pm 2.776.\frac{1}{\sqrt{5}} = 28.16$ *to* 30.64

The confidence interval includes 30 ; hence, the rating cannot be rejected at the 95% confidence level. However, the confidence interval is relatively wide, suggesting that more data should be collected. Once again, the confidence interval provides more practically useful information than the t-test.

Example 2. A machine used to produce gaskets has been stable and operating under control for many years, but lately the thickness of the gaskets seems to be smaller than they once were. The mean thickness was historically 0.070 inches. A Quality Assurance manager wants to determine if the age of the machine is causing it to produce poorer quality gaskets. He takes a sample of 10 gaskets for testing and finds a mean of 0.074 inches and a standard deviation of 0.008 inches. Test the hypothesis that the machine is working properly with a significance level of 0.05.

Sol. The null hypothesis should state that the population mean is still 0.070 inches-in other words, the machine is still working properly-and the alternate hypothesis should state that the mean is different from 0.070.

$H_0 : \mu = 0.070$ inches
$H_1 : \mu \neq 0.070$ inches

We have equality, therefore we are faced with a two-tailed test and we will have $\alpha/2 = 0.025$ on each side. The degree of freedom $(n-1)$ is equal to 9. The value of t that we will be looking for is $t_{0.025,9} = 2.26$. If the computed value t falls within the interval [-2.26, 2.26] we will not reject the null hypothesis; otherwise, we will.

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{0.074 - .07}{0.008/\sqrt{10}} = 1.012$$

The computed value of t is 1.012, therefore it falls within the interval [-2.262, + 2.262]. We conclude that we cannot reject the null hypothesis.

Hence, a summary of the test procedures for both two-and one sided alternative hypothesis follows:

The One-Sample *t*-Test :

Null Hypothesis : $H_0 : \mu = \mu_0$

Test Statistics : $t_0 = \frac{\bar{X} - \mu}{S/\sqrt{n}}$, where $S^2 = \frac{1}{n-1}\sum_{n-1}^{n}(X_i - \bar{X})^2$

Alternative hypothesis	Rejection criteria
$H_1 : \mu \neq \mu_0$	$t_0 > t_{\alpha/2,n-1}$ or $t_0 < \sim -t_{\alpha/2,n-1}$
$H_1 : \mu > \mu_0$	$t_0 > t_{\alpha,\, n-1}$
$H_1 : \mu < \mu_0$	$t_0 < -t_{\alpha,n-1}$

10.15.1.3 Testing Equality of Two Population Mean (When Samples are Small and Variance is Unknown but Equal)

If the population variances σ_1^2 and σ_2^2 are unknown and we assume that they are equal, they can be estimated using the sample variances S_1^2 and S_2^2. The estimate S_p^2 based on the two sample variances is called the **pooled sample variance**. Its formula is given as

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Where $n_1 - 1$ are the degree of freedom of Sample 1 and $n_2 - 1$ are the degree of freedom of Sample 2.

Here the null hypothesis being tested is $H_0 : \mu_1 = \mu_2$ against alternative hypothesis $H_0 : \mu_1 \neq \mu_2$.

The test statistic

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

If the computed value of t exceeds the critical value, H_0 is rejected and the difference is said to be satistically significant.

10.15.1.4 Confidence Interval for Two Sample t Test

The $100\,(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ is

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2, n_1+n_2-2} \quad S_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example 3 : The general manager of jaguar-Semiconductors oversees have two production plants and decided to raise the customer satisfaction index (CSI) to at least 98. To determine if there is a difference in the mean of the CSI in the two plants, random samples are taken over several weeks. For the J-Techo's plant, a sample of 17 weeks has yielded a mean of 96 CSI and a standard deviation of 3, and for the J-Electro plant, a sample of 19 weeks has generated a mean of 98 CSI and a standard deviation of 4. At the 0.05 level, determine if a difference exists in the mean level of CSI for the two plants, assuming that the CSIs are normal and have the same variance.

Sol. Here we have given the following details

$$\alpha = 0.05, n_1 = 17, \bar{x}_1 = 96, \; s_1 = 3 \; n_2 = 19, \bar{x}_1 = 98 \text{ and } s_2 = 4$$

Also the null hypothesis is defined as

$$H_0 : \mu_1 = \mu_2$$

Against the alternatives hypothesis

$$H_1 : \mu_1 \neq \mu_2$$

Estimate the common variance with the pooled sample variance, S_p^2

$$S_p^2 = \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S^2}{n_1 + n_2 - 2}$$

$$S_p^2 = \frac{16(3)^2 + 18(4)^2}{34} = \frac{432}{34} = 12.71$$

The value of the test statistics is

$$t = \frac{(96 - 98)}{\sqrt{12.71\left(\frac{1}{17} + \frac{1}{19}\right)}} = -\frac{2}{1.19} = -1.68$$

Because the alternate hypothesis does not involve "greater than" or "less than" but rather "is different from," we are faced with a two-tailed rejection region with $\alpha/2 = 0.05/2 = 0.025$ at the end of each tail with a degree of freedom of 34. From the *t* table, we obtain $t_{0.025} = 2.03$ and H_0 is not rejected when $-2.03 < t < +2.03$. $t = -1.68$ is well within the interval, we therefore cannot reject the null hypothesis.

Null Hypothesis : $H_0 : \mu_1 = \mu_2$

Test Statistics : $t_0 = \dfrac{\bar{X}_1 - \bar{X}_2}{S_p \Big/ \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$, where $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

Alternative hypothesis	Rejection criteria
$H_1 : \mu_1 \neq \mu_2$	$t_0 > t_{\alpha/2, n_1+n_2-2}$ or $t_0 < -t_{\alpha/2, n_1+n_2-2}$
$H_1 : \mu_1 > \mu_2$	$t_0 > t_{\alpha, n_1+n_2-2}$
$H_1 : \mu_1 < \mu_2$	$t_0 < -t_{\alpha, n_1+n_2-2}$

T-Test

- T Test is a small sample Test.
- It was developed by William Gosset in 1908.
- It is also called students t test (pen name).

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Uses of T-Test / Application

- size of sample is small (n<30)

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Uses of T-Test / Application

- size of sample is small (n<30).
- Degree of freedom is (v=n-1).
- T-Test is used for test of significance of regression coefficient in regression model.

Application of t-test

① To test Significance of mean of random sample [n < 30]

1) Set Null Hypothesis

H_0: $\bar{X} = \mu$

Set Alternative Hypothesis

H_1: $\bar{X} \neq \mu$ [two-tailed]

$\bar{X} > \mu$, $\bar{X} < \mu$ [one-tailed]

2) Test Statistics

$$t = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \quad \text{or} \quad t = \frac{\bar{X} - \mu}{s}\sqrt{n}$$

$\bar{X}$ = Sample mean

μ = Population mean

n = Sample size

s = Sample S.D.

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}$$

d.f. $\nu = n - 1$

Level of Significance [LOS]

$\alpha = 0.05$ or 0.01

Decision

Cal. $|t|$ tab t at certain d.f and at certain LOS α

① Cal. $|t|$ > tab t

H_0 reject

Cal. $|t|$ < tab t

H_0 accept