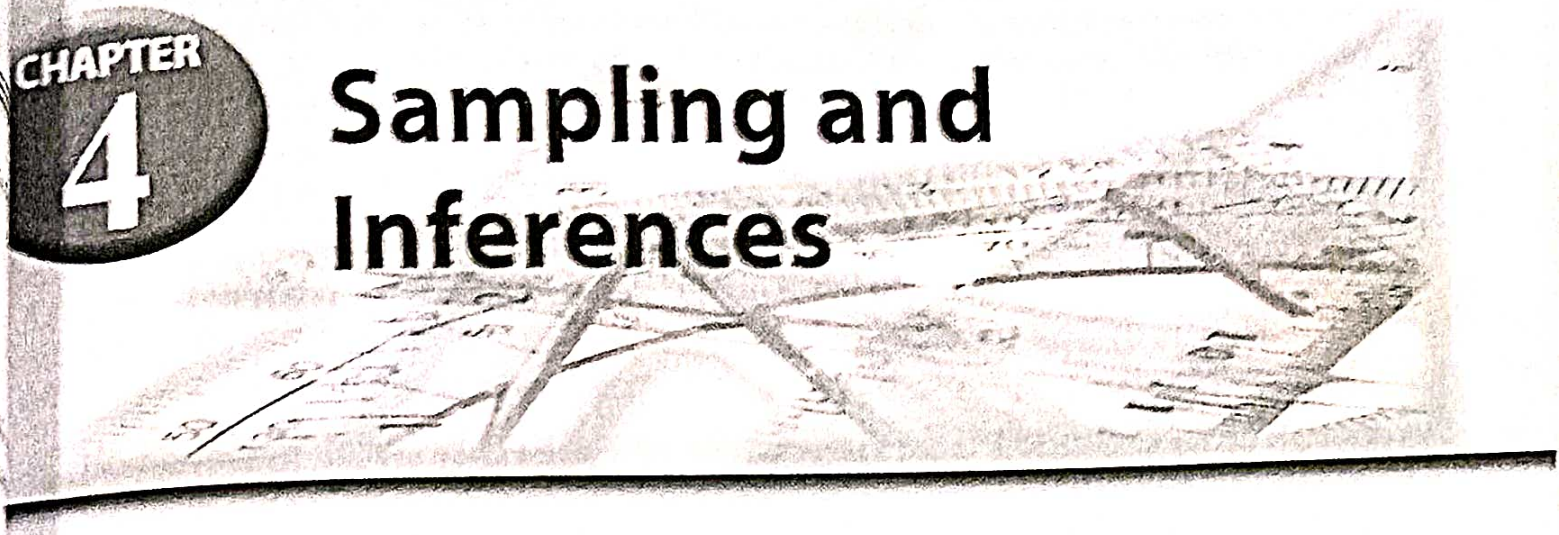
CHAPTER 4

Sampling and Inferences



4.1 INTRODUCTION

In our daily life, most of our decisions depend very much upon the inspection or examination of only a few objects out of the total lot. For example, in a shop, we assess the quality of rice, wheat or any other commodity by taking a handful of it from the bag and then decide whether to purchase it or not. Such examples are related to the concept of sampling. In this chapter, we will learn about random sampling and large sample tests. Then, we will also discuss sampling distribution and small sample tests.

4.2 POPULATION AND SAMPLING

We know that population is the set of all the observations under study, and a sample is a subset of the population drawn for analysis. The process of selecting such samples from the population is called **sampling**. The samples must be selected at random to exclude the possibility of any biasedness. This is called **random sampling**. Thus, in random sampling, each member of the population has equal chance of being included in the sample. A special type of random sampling, is **simple sampling**. In this sampling, each event has the same probability of success and the chance of success of different events is independent whether previous trials have been made or not.

The main objective of sampling is to draw as much inferences as possible of the population by examining only the sample. At the same time, sampling helps in minimising the effort, cost and time. The logic of sampling theory is similar to that of the logic of induction where we pass from particular (sample) to general (population). Such generalisation of inferences from the sample to the population is called **statistical inference**.

The statistical measures or constants of the population such as mean (μ) and variance (σ^2) are called the **parameters** of the population, whereas the statistical measures computed from the samples such as mean ($\bar{x}$) and variance (s^2) are called **statistics**. Population parameters are denoted using Greek letters or capital letters, and sample statistics are denoted using Roman letters. Most often, the population parameters are not

known and their estimates given by the corresponding samples (sample statistics) are used for the analysis of population. However, sample statistics based on different samples can vary from one sample to another. Sampling determines the reliability of these estimates.

4.3 SAMPLING DISTRIBUTION

Let all possible samples of size n be drawn from a population at random. Then, we compute some statistic, say mean ($\bar{x}$), for each of the samples. The means of different samples will not be the same.

If these different means are grouped according to their frequencies, the frequency distribution thus obtained is called the **sampling distribution of mean**. Similarly, we can obtain the **sampling distribution of variance**, etc.

A sample having size $n \geq 30$ is called a **large sample**, otherwise a **small sample**. If the sample is large then the sampling distribution of a statistic approaches a normal distribution. We observe that population may or may not be normal.

If we draw a sample from the population and make some measurement on it and put it back in the population before drawing another sample so that the parent population remains unchanged, then this is called **sampling with replacement**. Henceforth, we will only use the concept of sampling with replacement.

4.3.1 Standard Error

The standard deviation of the sampling distribution of a statistic is called the **standard error** (SE) of the statistic. Thus, the standard error of means is the standard deviation of the sampling distribution of means. Standard error is used to evaluate the difference between the sample statistic and the corresponding population parameter and between two sample statistics. The reciprocal of standard error is called **precision**.

For large samples, the standard errors of some well-known statistics are given in Table 4.1.

Table 4.1 Standard errors of some well-known statistics

Statistic	Standard error
Difference between sample mean ($\bar{x}$) and population mean (μ)	$\sigma/\sqrt{n}$
Difference between sample standard deviation (s) and population standard deviation (σ)	$\sigma/\sqrt{2n}$
Difference between sample proportion (p) and population proportion (P)	$\sqrt{PQ/n}$
Difference between two sample means ($\bar{x}_1 - \bar{x}_2$)	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Difference between two sample standard deviations ($s_1 - s_2$)	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
Difference between two sample proportions ($p_1 - p_2$)	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

In Table 4.1, n is the sample size, σ^2 the population variance, s the sample standard deviation, P the population proportion, p the sample proportion, $Q = 1 - P$.

known and their estimates given by the corresponding samples (sample statistics) are used for the entire of population. However, sample statistics based on different samples can vary from one sample to another. Sampling determines the reliability of these estimates.

4.3 SAMPLING DISTRIBUTION

Let all possible samples of size n be drawn from a population at random. Then, we compute some statistic, say mean ($\bar{x}$), for each of the samples. The means of different samples will not be the same.

If these different means are grouped according to their frequencies, the frequency distribution thus obtained is called the **sampling distribution of mean**. Similarly, we can obtain the **sampling distribution of variance**, etc.

A sample having size $n \geq 30$ is called a **large sample**, otherwise a **small sample**. If the sample is large, then the sampling distribution of a statistic approaches a normal distribution. We observe that population may or may not be normal.

If we draw a sample from the population and make some measurement on it and put it back in the population before drawing another sample so that the parent population remains unchanged, then this is called **sampling with replacement**. Henceforth, we will only use the concept of sampling with replacement.

4.3.1 Standard Error

The standard deviation of the sampling distribution of a statistic is called the **standard error** (SE) of the statistic. Thus, the standard error of means is the standard deviation of the sampling distribution of means. Standard error is used to evaluate the difference between the sample statistic and the corresponding population parameter and between two sample statistics. The reciprocal of standard error is called **precision**.

For large samples, the standard errors of some well-known statistics are given in Table 4.1.

Table 4.1 Standard errors of some well-known statistics

Statistic	Standard error
Difference between sample mean ($\bar{x}$) and population mean (μ)	$\sigma/\sqrt{n}$
Difference between sample standard deviation (s) and population standard deviation (σ)	$\sigma/\sqrt{2n}$
Difference between sample proportion (p) and population proportion (P)	$\sqrt{PQ/n}$
Difference between two sample means ($\bar{x}_1 - \bar{x}_2$)	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Difference between two sample standard deviations ($s_1 - s_2$)	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
Difference between two sample proportions ($p_1 - p_2$)	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

In Table 4.1, n is the sample size, σ^2 the population variance, s the sample standard deviation, P the population proportion, p the sample proportion, $Q = 1 - P$.

4.4 TEST OF SIGNIFICANCE

In sampling theory, we are mainly concerned with the study of test of significance. This test enables us to decide, on the basis of sample results, whether the difference between the observed sample statistic and hypothetical parameter value or the difference between two independent sample statistics is significant or might be attributed due to chance or fluctuations of sampling.

4.4.1 Null Hypothesis and Alternative Hypothesis

For applying the test of significance, we set up a hypothesis which is tested for possible rejection under the assumption that it is true. Such hypothesis is called the **null hypothesis** and is denoted by H_0.

The hypothesis complementary to null hypothesis is called the **alternative hypothesis** and is denoted by H_1.

Suppose we want to test the null hypothesis that the population has an assumed value of mean μ_0. Then, we have $H_0: \mu = \mu_0$. The three possible alternative hypotheses will be

(i) $H_1: \mu \neq \mu_0$ $(\mu > \mu_0 \text{ or } \mu < \mu_0)$ —This alternative hypothesis is called the **two-tailed alternative hypothesis** or the **two-tailed test.**

(ii) $H_1: \mu > \mu_0$ —This alternative hypothesis is called the **right-tailed alternative hypothesis** or the **single-tailed test.**

(iii) $H_1: \mu < \mu_0$ —This alternative hypothesis is called the **left-tailed alternative hypothesis** or the **single-tailed test.**

4.4.2 Critical Region and Level of Significance

A region R, corresponding to a sample statistic t, in the sample space which amounts to the rejection of the null hypothesis H_0 is called the **critical region** or the **region of rejection.** The complementary region $\bar{R}$ which amounts to the acceptance of H_0 is called the **acceptance region.**

The probability of the value of the variate falling in the critical region is called the **level of significance.** If t is the value of statistics obtained using a random sample of size n and R is the critical region, then the probability α that a random value of statistic t belongs to the critical region is the level of significance and is given by $P(t \in R \mid H_0) = \alpha$.

The level of significance is always fixed in advance before studying the characteristics of random sample. It is usually expressed as a percentage, and the total area of the critical region is written as $\alpha\%$ level of significance. The two popular values of the level of significance which are usually employed in testing the hypothesis are 5% and 1%.

4.4.3 Single-Tailed and Two-Tailed Tests

A test of any statistical hypothesis in which the alternative hypothesis is single tailed (right tailed or left tailed) is called a **single-tailed test.** A test of any statistical hypothesis in which the alternative hypothesis is two tailed is called a **two-tailed test.**

4.4.4 Critical Values

In case of large samples, if t is any statistics and $E(t)$ is the corresponding population mean, then the vari

$z = \dfrac{t - E(t)}{SE(t)}$ is normally distributed with mean 0 and variance unity. Value of test statistics z which separ

the critical region and the accepted region is called the **critical value** or the **significant value** of z. We den

this value by z_α, where α is the level of significance. Critical value depends on

1. the prescribed level of significance and
2. the alternative hypothesis, whether it is a two-tailed test or a single-tailed test.

The critical value z_α of the test statistic for a two-tailed test is given by

$$P(|z| > z_\alpha) = \alpha \qquad (4$$

i.e., the total area of the critical region lying at both the ends under the probability curve is α (see Fig 4.1

Since a normal curve is symmetrical, $P(z > z_\alpha) = P(z < -z_\alpha)$. Now, from (4.1), we ha

$P(z > z_\alpha) + P(z < -z_\alpha) = \alpha$ or $2P(z > z_\alpha) = \alpha$ or $P(z > z_\alpha) = \alpha/2$, i.e., the area under each tail is $\alpha/2$. T

value $z = z_\alpha$ is called the **upper critical value**, and the value $z = -z_\alpha$ is called the **lower critical value**. T

acceptance region is given by $(-z_\alpha, z_\alpha)$.

The critical value z_α of the test statistic for a right-tailed test is given by

$$P(z > z_\alpha) = \alpha \qquad (4.2$$

i.e., the total area of the critical region α is the area of the right tail under the probability curve.

The critical value z_α of the test statistic for a left-tailed test is given by

$$P(z < -z_\alpha) = \alpha \qquad (4.3$$

i.e., the total area of the critical region α is the area of the left tail under the probability curve.

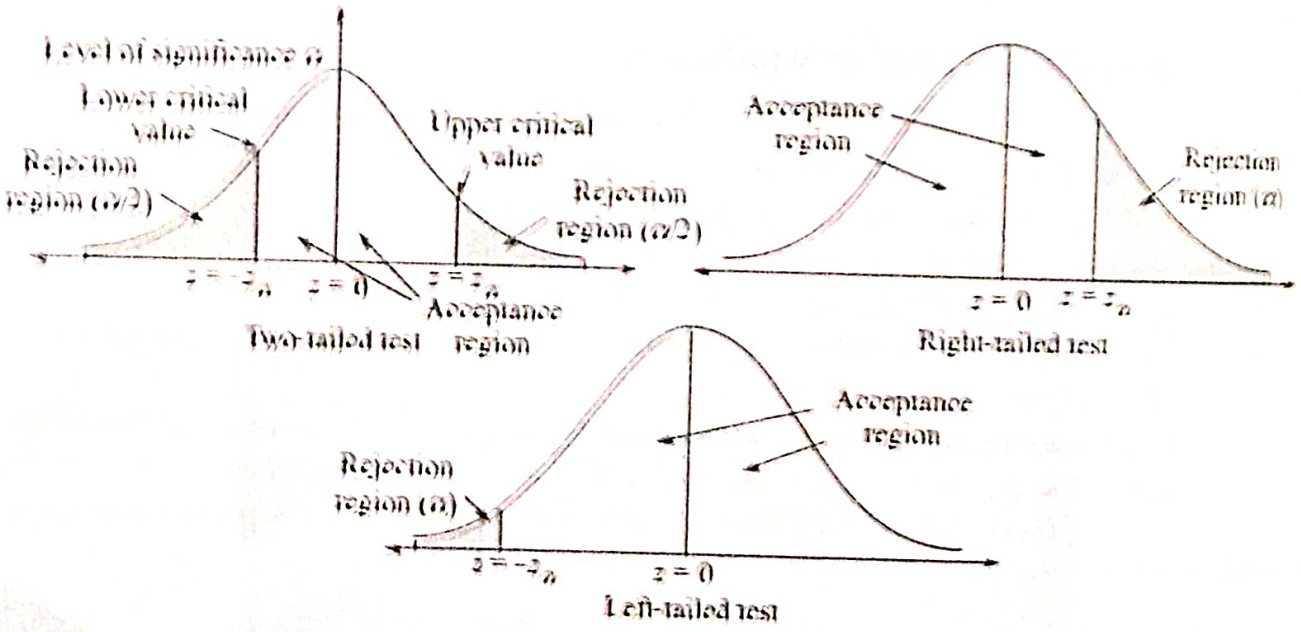


Fig. 4.1

Due to symmetry, we have

$$P(|z| > z_\alpha) = P(z > z_\alpha) + P(z < -z_\alpha) = P(z > z_\alpha) + P(z > z_\alpha) = 2\alpha \qquad \text{[Using (4.2)]}$$

Mon T
3·30 3

ATTENDANCE SHEET
CSE DEPARTMENT

Name of Staff:	Month & Year: Aug 22	Lectures
Class: SY. B.Tech. (CSE) (A) 2022-23-I	Subject:	Lectures

DATE →	23/8	24	26/8	28/8		5/9	8/9	12/9	13/9
ISO	SIGN	SIGN	SIGN	SIGN	SIGN				

Thus, the critical value of z for a single-tailed test at the level of significance α is the same as the critical value of z for the two-tailed test at the level of significance 2α.

The critical values of z at commonly used level of significance for these tests are listed in Table 4.2.

Table 4.2

Test	Critical value	Level of significance		
		1%	5%	10%
Two tailed	$\lvert z_\alpha \rvert$	2.58	1.96	1.645
Right tailed	z_α	2.33	1.645	1.28
Left tailed	z_α	−2.33	−1.645	−1.28

4.4.5 Confidence Limits

The interval in which a population parameter is supposed to lie is called the **confidence interval** for that population parameter. The end points of this interval are called the **confidence limits** or the **fiducial limits.** The probability that is associated with the confidence interval is called the **confidence level.** It is usually written as $1 - \alpha$, where α is the level of significance. Confidence level indicates the confidence that the experimenter has that the population parameter actually lies within the confidence interval.

Consider that the sampling distribution of statistic t is normal with mean μ and standard deviation σ. The sample statistic t can be expected to lie in the interval $(\mu - 1.96\sigma, \mu + 1.96)$ for 95% times, i.e., we can be confident of finding μ in the interval $(t - 1.96\sigma, t + 1.96\sigma)$ in 95% cases. Thus, we call $(t - 1.96\sigma, t + 1.96\sigma)$ the 95% confidence interval for the estimation of μ. The ends of this interval (i.e., $t \pm 1.96\sigma$) are called 95% confidence limits (or fiducial limits) for t. Similarly, $S \pm 2.58\sigma$ are 99% confidence limits. The number 1.96 is called the confidence coefficient. The values of confidence coefficients corresponding to various levels of significance can be found from the normal curve area table given in the Appendix.

4.4.6 Errors in Testing a Hypothesis

The main aim of sampling theory is to draw valid inferences about the population parameters on the basis of sample results. Because of these reasons, we are liable to commit the following two types of errors:

Type I error: We reject H_0, when it is true. If we write $P[\text{Reject } H_0 \mid H_0] = \alpha$, then α (level of significance) is called the **size of type I error.** It is also referred to as **producer's risk.**

Type II error: We accept H_0, when it is not true, i.e., accept H_0 when H_1 is true. If we write $P[\text{Accept } H_0 \mid H_1] = \beta$, then β is called the **size of type II error.** It is also referred to as **consumer's risk.**

The statistical testing of hypothesis aims to limit the type I error to preassigned values (say, 5% or 1%) and to minimise the type II error. Both these errors can be reduced by increasing the size of the sample (if possible).

4.4.7 Steps for Testing a Hypothesis

1. **Null hypothesis:** Define the null hypothesis H_0.

2. **Alternative hypothesis:** Define the alternative hypothesis H_1 so as to decide the test to be two-tailed test or single-tailed test.
3. **Level of significance:** Depending on the problem, fix the appropriate level of significance in advance. This level of significance is fixed before drawing the random sample.
4. **Critical value:** Obtain the value of z_α at the level of significance α.
5. **Test statistic:** Compute the test statistic $z = \dfrac{t - E(t)}{SE(t)}$ under the null hypothesis.
6. **Conclusion:** Compare $|z|$ with z_α. If $|z| > z_\alpha$ then we reject H_0 and accept H_1 for the level of significance α. This implies that the difference $|t - E(t)|$ is significant for the level of significance α. If $|z| < z_\alpha$ we accept H_0 and reject H_1 for the level of significance α. This implies that the difference is due to some fluctuations in sampling, and hence, the difference is not significant for the level of significance α.

4.5 TEST OF SIGNIFICANCE FOR LARGE SAMPLES

We know that the sample having size $n \geq 30$ is called a **large sample**. For such samples, distributions such as binomial, Poisson, negative binomial and hypergeometric, approach normal distribution assuming population is normal.

4.5.1 Test for Difference Between Sample Proportion and Population Proportion

Let P and p be population proportion and sample proportion, respectively. Let a sample of size n be drawn from the population. Clearly, this is the same as a series of n independent trials with constant probability of success.

If X is the number of successes in n independent trials with constant probability P of success for each, then $X \sim B(n, P)$. Therefore,

$$E(X) = nP,\ V(X) = nPQ,\ Q = 1 - P$$

Since $p = X/n$ is the observed proportion of success in the sample, we have

$$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{nP}{n} = P,\ V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{n(PQ)}{n^2} = \frac{PQ}{n}$$

Hence, $SE(p) = \sqrt{\dfrac{PQ}{n}}$ and $z = \dfrac{t - E(p)}{SE(p)} = \dfrac{p - P}{\sqrt{PQ/n}} \sim N(0, 1)$

This z is used to test the significant difference between the population proportion and the sample proportion.

Note:

(i) The probable limits for the observed proportion p of successes are $P \pm z_\alpha\sqrt{PQ/n}$.

(ii) If P is not known, then take $P = p$. Then and the approximate limits for the proportion of population are $p \pm z_\alpha\sqrt{pq/n}$, where $q = 1 - p$.

Example 4.1 A coin was tossed 400 times and the head turned up 216 times. Test the hypothesis that the coin is unbiased.

2. **Alternative hypothesis:** Define the alternative hypothesis H_1 so as to decide the test to be two-tailed test or single-tailed test.
3. **Level of significance:** Depending on the problem, fix the appropriate level of significance in advance. This level of significance is fixed before drawing the random sample.
4. **Critical value:** Obtain the value of z_α at the level of significance α.
5. **Test statistics:** Compute the test statistic $z = \dfrac{t - E(t)}{\text{SE}(t)}$ under the null hypothesis.
6. **Conclusion:** Compare $|z|$ with z_α. If $|z| > z_\alpha$, then we reject H_0 and accept H_1 for the level of signific α. This implies that the difference $|t - E(t)|$ is significant for the level of significance α. If $z_c <$ we accept H_0 and reject H_1 for the level of significance α. This implies that the difference $t -$ is due to some fluctuations in sampling, and hence, the difference is not significant for the le significance α.

4.5 TEST OF SIGNIFICANCE FOR LARGE SAMPLES

We know that the sample having size $n \geq 30$ is called a **large sample**. For such samples, distribu such as binomial, Poisson, negative binomial and hypergeometric, approach normal distribution assumin population is normal.

4.5.1 Test for Difference Between Sample Proportion and Population Proportion

Let P and p be population proportion and sample proportion, respectively. Let a sample of size n be d from the population. Clearly, this is the same as a series of n independent trials with constant probabil of success.

If X is the number of successes in n independent trials with constant probability P of success for each then $X \sim B(n, P)$. Therefore,

$$E(X) = nP,\ V(X) = nPQ,\ Q = 1 - P$$

Since $p = X/n$ is the observed proportion of success in the sample, we have

$$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{nP}{n} = P,\ V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{n(PQ)}{n^2} = \frac{PQ}{n}$$

Hence, $\text{SE}(p) = \sqrt{\dfrac{PQ}{n}}$ and $z = \dfrac{p - E(p)}{\text{SE}(p)} = \dfrac{p - P}{\sqrt{PQ/n}} \sim N(0, 1)$

This z is used to test the significant difference between the population proportion and the sa proportion.

Note:

(i) The probable limits for the observed proportion p of successes are $P \pm z_\alpha\sqrt{PQ/n}$.

(ii) If P is not known, then take $P \approx p$. Then and the approximate limits for the proportion of popula are $p \pm z_\alpha\sqrt{pq/n}$, where $q = 1 - p$.

Example 4.1 A coin was tossed 400 times and the head turned up 216 times. Test the hypothesis th the coin is unbiased.

2. **Alternative hypothesis:** Define the alternative hypothesis H_1 so as to decide the test
two-tailed test or single-tailed test.
3. **Level of significance:** Depending on the problem, fix the appropriate level of signific
advance. This level of significance is fixed before drawing the random sample.
4. **Critical value:** Obtain the value of z_α at the level of significance α.
5. **Test statistics:** Compute the test statistic $z = \dfrac{t - E(t)}{\mathrm{SE}(t)}$ under the null hypothesis.
6. **Conclusion:** Compare $|z|$ with z_α. If $|z| > z_\alpha$, then we reject H_0 and accept H_1 for the level of sig
α. This implies that the difference $|t - E(t)|$ is significant for the level of significance α. If z
we accept H_0 and reject H_1 for the level of significance α. This implies that the difference
is due to some fluctuations in sampling, and hence, the difference is not significant for the
significance α.

4.5 TEST OF SIGNIFICANCE FOR LARGE SAMPLES

We know that the sample having size $n \geq 30$ is called a **large sample**. For such samples, distri
such as binomial, Poisson, negative binomial and hypergeometric, approach normal distribution assu
population is normal.

4.5.1 Test for Difference Between Sample Proportion and Population Proportion

Let P and p be population proportion and sample proportion, respectively. Let a sample of size n b
from the population. Clearly, this is the same as a series of n independent trials with constant prob
of success.

If X is the number of successes in n independent trials with constant probability P of success for
then $X \sim B(n, P)$. Therefore,

$$E(X) = nP,\ V(X) = nPQ,\ Q = 1 - P$$

Since $p = X/n$ is the observed proportion of success in the sample, we have

$$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{nP}{n} = P,\ V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{n(PQ)}{n^2} = \frac{PQ}{n}$$

Hence, $\mathrm{SE}(p) = \sqrt{\dfrac{PQ}{n}}$ and $z = \dfrac{p - E(p)}{\mathrm{SE}(p)} = \dfrac{p - P}{\sqrt{PQ/n}} \sim N(0, 1)$

This z is used to test the significant difference between the population proportion and
proportion.

Note:

(i) The probable limits for the observed proportion p of successes are $P \pm z_\alpha\sqrt{PQ/n}$.

(ii) If P is not known, then take $P \approx p$. Then and the approximate limits for the proportion o
are $p \pm z_\alpha\sqrt{pq/n}$, where $q = 1 - p$.

Example 4.1 A coin was tossed 400 times and the head turned up 216 times. Test the hyp
the coin is unbiased.

Solution: Here, $n = 400$, X = Number of success = 216

p = Proportion of success in the sample = X/n = 216/400 = 0.54

P = Population proportion (i.e., getting head or tail) = 0.5 and $Q = 1 - P = 1 - 0.5 = 0.5$

H_0: The coin is unbiased, i.e., $P = 0.5$

H_1: The coin is not unbiased, i.e., $P \neq 0.5$ (two tailed)

Under H_0, test statistic $z = \dfrac{p - P}{\sqrt{PQ/n}} = \dfrac{0.54 - 0.5}{\sqrt{0.5 \times 0.5/400}} = 1.6$

Since $|z| < 1.96$, the hypothesis is accepted at 5% level of significance and we may conclude that the coin is unbiased at 5% level of significance.

Example 4.2 A cubical die is thrown 9000 times and a throw of 3 or 4 is observed 3240 times. Show that the die cannot be regarded as an unbiased one and find the extreme limits between which the probability of a throw of 3 or 4 lies.

Solution: Here, $n = 9000$, X = Number of success = 3240

p = Probability of success (i.e., getting 3 or 4 on die) = 2/6 = 1/3, $Q = 1 - 1/3 = 2/3$

P = Probability of success in sample X/n = 2340/9000 = 0.36

H_0: The die is unbiased, i.e., $P = 1/3$

and H_1: $P \neq 1/3$ (two-tailed test)

Under H_0, test statistic $z = \dfrac{p - P}{\sqrt{PQ/n}} = \dfrac{0.36 - 0.33}{\sqrt{(1/3) \times (2/3) \times (1/9000)}} = 0.03496$

Since $|z| = 0.03496 < 1.96$, the hypothesis is accepted at 5% level of significance and we may conclude that the die is unbiased at 5% level of significance.

We have to find 95% confidence limits of the proportion. It is given by

$$P \pm z_\alpha \sqrt{\frac{PQ}{n}} = 0.33 \pm 1.96 \sqrt{\frac{0.33 \times 0.67}{9000}} = 0.33 \pm 0.0097 = 0.3203 \text{ and } 0.3397$$

Example 4.3 A manufacturer claimed that at least 95% of the equipments which he supplied to a factory conformed to the specifications. An examination of a sample of 200 pieces of the equipment revealed that 18 were faulty. Test this claim at a significant level of **(i)** 0.05 and **(ii)** 0.01.

Solution: Here, $n = 200$, X = Number of success = 200 − 18 = 182, p = Proportion of success $= X/n = 182/200 = 0.91$, P = Probability of success = 0.95, $Q = 1 - 0.95 = 0.05$

H_0: The proportion of success in the lot is 95% at least, i.e., $P \geq 0.95$

and H_1: $P < 0.95$ (left-tailed test)

Under H_0, test statistic $z = \dfrac{p - E(p)}{\text{SE}(p)} = \dfrac{p - P}{\sqrt{PQ/n}} = \dfrac{0.91 - 0.95}{\sqrt{(0.95 \times .05)/200}} = \dfrac{-0.04}{\sqrt{0.95 \times 0.05/200}}$

$$= \frac{-0.04}{\sqrt{0.0002375}} = \frac{-0.04}{0.0154} = -2.6$$

(i) Since $z = -2.6 < -1.645$, the hypothesis is rejected for left-tailed test at 5% level of significance and we can conclude that manufacturer's claim is rejected at 5% level of significance.

(ii) Since $z = -2.6 < -2.33$, the hypothesis is rejected for left-tailed test at 1% level of significance and we can conclude that manufacturer's claim is rejected at 1% level of significance.

Example 4.4 Twenty people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate of those attacked by this disease is 85%?

Solution: Here, $n = 20$, X = Number of persons who survived = 18, p = Proportion of people who survived = $18/20 = 0.9$, P = Probability of people who survived = 0.85, $Q = 1 - 0.85 = 0.15$

H_0: The proportion of people who survived the attack is 85%, i.e., $P = 0.85$

and H_1: $P > 0.85$ (right-tailed test)

Under H_0, the test statistic $z = \dfrac{p - P}{\sqrt{PQ/n}} = \dfrac{0.90 - 0.85}{\sqrt{(0.85 \times 0.15)/20}} = \dfrac{0.05}{0.08} = 0.625$

Since $|z| = 0.625 < 1.656$, i.e., the calculated value $|z|$ is less than the tabulated value of z, i.e., z_α at the level of significance for right-tailed test, the hypothesis is accepted.

Example 4.5 A large consignment contains bad apples, the exact number of which is not known. A random sample of 500 apples was taken from the consignment and 60 were found to be bad. Obtain the 98% confidence limit for the percentage of bad apples in the consignment.

Solution: Here, $n = 500$, X = Number of bad apples in the sample = 60, p = Proportion of bad apples in the sample = $60/500 = 0.12$

Since the critical value of z at 2% level of significance is 2.33, 98% confidence limits (2% level of significance) for population proportion are

$$p \pm 2.33\sqrt{(pq/n)} = 0.12 \pm 2.33\sqrt{(0.12 \times 0.88/500)}$$

$$= 0.12 \pm 2.33 \times \sqrt{0.000212} = 0.12 \pm 2.33 \times 0.01453$$

$$= 0.12 \pm 0.03386 = 0.0861, 0.015386$$

Hence, 98% confidence limit for bad apples in the consignment is (8.61,15.38).

EXERCISE 4.1

1. About 325 men out of 600 men chosen from a big city were found to be smokers. Does this information support the conclusion that the majority of men in the city are smokers?
2. A die is tossed 960 times and it falls with 5 upwards 184 times. Is the die biased?
3. A sample of 900 days is taken from meteorological records of a certain district and 100 of them are found to be foggy. What is the probable limit to the percentage of foggy days?
4. A random sample of 500 fuses was taken from a large consignment and 65 were found to be defective. Show that the percentage of defectives in the consignment almost certainly lies between 8.5 and 17.5.

5. A manufacturer claims that only 4% of the products supplied by him are defective. A random sample of 600 products contained 36 defectives. Test the claim of the manufacturer.
6. In a random sample of 200 people in a city, 108 like to purchase imported watches and the remaining like to purchase local watches. Can we conclude that both the imported and the local watches are popular in the city? (Use 2% level of significance.)
7. A bag contains defective articles, the exact number of which is not known. A sample of 100 from the bag gives 10 defective articles. Find the limits for the proportion of defective articles in the bag.
8. From a large lot of mangoes, a random sample of 600 mangoes was drawn and 60 were found to be bad. Find the standard error of the proportion of bad mangoes in this sample. Hence, find the 3σ limits for the percentage of bad mangoes in this lot.
9. About 400 apples are taken at random from a large basket and 40 are found to be bad. Estimate the proportion of bad apples in the basket.
10. A sample of 600 persons selected at random from a large city shows that the percentage of males in the sample is 53. It is believed that the ratio of males to the total population in the city is 0.5. Test whether the belief is confirmed by the observation.
11. Balls are drawn from a bag containing equal number of black and white balls, each ball being replaced before drawing another. In 2250 drawings, 1018 black and 1232 white balls are drawn. Do you suspect some bias on the part of the drawer?
12. A manufacturer claims that only 10% of the articles produced are below the standard quality. Out of a random inspection of 300 articles, 37 are found to be of poor quality. Test the manufacturer's claim at 5% level of significance.

4.5.2 Test for Difference Between Two Sample Proportions

Let two samples of sizes n_1 and n_2 be drawn from two populations with population proportions P_1 and P_2, respectively. Also, let X_1 and X_2 be the number of successes and p_1 and p_2 be the observed proportions of successes in these samples, respectively. Then $p_1 = X_1/n_1$ and $p_2 = X_2/n_2$. Defining $Q_1 = 1 - P_1$ and $Q_2 = 1 - P_2$, we have

$$E(p_1) = P_1,\ E(p_2) = P_2,\ V(p_1) = \frac{P_1Q_1}{n_1} \text{ and } V(p_2) = \frac{P_2Q_2}{n_2}$$

Since for large samples, p_1 and p_2 are normally distributed, their difference is also normally distributed. Therefore,

$$E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2$$

$$V(p_1 - p_2) = V(p_1) + V(p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}$$

Hence, $$SE(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}} \quad (4.4)$$

and the standardised variable corresponding to $p_1 - p_2$ is given by

$$z = \frac{p_1 - p_2 - E(p_1 - p_2)}{SE(p_1 - p_2)} = \frac{p_1 - p_2 - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

We set up the null hypothesis H_0 that there is no significant difference between two population proportions i.e., $H_0: P_1 = P_2 = P$.

Under H_0, the standard error and test statistics are, respectively, given by

$$SE(p_1 - p_2) = \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \text{ and } z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Here, we use unbiased estimate of the common population proportion P provided by two samples taken together which is given by

$$P = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} \text{ and } Q = 1 - P$$

Under H_0: $P_1 = P_2$, i.e., the sample proportions are equal, the standard error is the same as (4.4) and the test statistics is given by

$$z = \frac{P_1 - P_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1) \qquad \text{(Taking absolute value)}$$

Example 4.6 In a sample of 600 men from a certain city, 450 are found smokers. In another sample of 900 men from another city, 450 are smokers. Do the data indicate that the cities are significantly different with respect to the habit of smoking among men?

Solution: Here, $n_1 = 600$ men, $X_1 = 450$, $p_1 = \frac{X_1}{n_1} = \frac{450}{600} = 0.75$

$n_2 = 600$ men, $X_2 = 450$, $p_2 = \frac{X_2}{n_2} = \frac{450}{900} = 0.50$

$$P = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{600 \times 0.75 + 900 \times 0.5}{600 + 900} = \frac{900}{1500} = 0.60 \text{ and } Q = 1 - P = 1 - 0.6 = 0.4$$

H_0: There is no significant difference with respect to the habit of smoking among the men of two cities, i.e., $p_1 = p_2$

and $H_1: p_1 > p_2$ (right tailed)

Under H_0, test statistic $z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.75 - 0.50}{\sqrt{0.60 \times 0.4\left(\frac{1}{600} + \frac{1}{900}\right)}} = 9.682$

The critical value of z at 5% level of significance for the right-tailed test is 1.645 and that at 1% level of significance for the right-tailed test is 2.33. Since $|z| > 1.645$ at 5% level of significance and also $|z| > 2.33$

ATTENDANCE SHEET
CSE DEPARTMENT

Name of Staff:	Month & Year: Aug 22								
Class: SY. B.Tech. (CSE) (A) 2022-23-I	Subject:								
DATE →	22/8	24	26/8	26/8		30/8	31/8	1/9	
ISO	SIGN	SIGN	SIGN	SIGN					

at 1% level of significance, it is significant at both levels of significance and we can reject the hypothesis at both levels of significance.

Example 4.7 In two large populations, there are 30% and 25% fair-haired people. Is this difference likely to be hidden in samples of 1200 and 900, respectively, from the two populations?

Solution: P_1 = Proportion of fair-haired people in the first population = 30% = 0.3 and $Q_1 = 0.7$

P_2 = Proportion of fair-haired people in the second population = 25% = 0.25 and $Q_2 = 0.75$

H_0: No significant difference in population proportions is likely to be hidden in sampling, i.e., $P_1 = P_2$

and $H_1: P_1 \neq P_2$ (two-tailed test)

Under H_0, the test statistic $$z = \frac{P_1 - P_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} = \frac{0.3 - 0.25}{\sqrt{\frac{0.3 \times 0.7}{1200} + \frac{0.25 \times 0.75}{900}}} = 2.5376$$

The critical value of z at 5% level of significance for the two-tailed test is 1.96. Since $|z| = 2.5376 > 1.96$, the hypothesis is rejected and we can conclude that a significant difference in population proportions is likely to be hidden in sampling at 5% level of significance.

Also, the critical value of z at 1% level of significance for the two-tailed test is 2.58. Since $|z| = 2.5376 < 2.58$, the hypothesis is accepted and we can conclude that no significant difference in population proportions is likely to be hidden in sampling at 1% level of significance.

EXERCISE 4.2

1. Before an increase in excise duty on tea, 400 people out of a sample of 500 people were found to be tea-drinkers. After an increase in duty, 400 people were tea-drinkers in a sample of 600 people. Using standard error of proportions, state whether there is a significant decrease in the consumption of tea. Take **(i)** $\alpha = 0.05$ and **(ii)** $\alpha = 0.01$.
2. In a city A, 20% of a random sample of 900 school boys had a certain slight physical defect. In another city B, 18.5% of a random sample of 1600 school boys had the same defect. Is the difference between the proportions significant?
3. A soft drink manufacturer has two brands of soft drinks: A and B. The company claims that brand A outsells brand B by 10%. Two independent random samples of people who regularly consume soft drinks are chosen. In the first sample of size 300, 80 people preferred brand A, and in the second sample of size 200, 50 people preferred brand B. Can we accept the manufacturer's claim of 10% difference in sales of A over B?
4. One type of aircraft is found to develop engine trouble in 5 flights out of a total of 100 and another type in 7 flights out of 200 flights. Is there a significant difference in the two types of aircrafts so far as engine defects are concerned?
5. Before a big increase in the price of petrol, 400 persons out of a sample of 1000 persons were found to purchase big sized cars. After the increase in the price of petrol, 280 persons out of a sample of 800 persons were found to purchase big sized cars. Find whether there is a significant decrease in the purchase of big cars. Test 5% and 2% levels of significance.
6. The percentage of officials in two big PSU's with computer knowledge is 30 and 25, respectively. Is this likely to be hidden in samples of size 1000 and 800, respectively, from the two PSU's?

4.5.3 Test for Difference Between Sample Mean and Population Mean

Let a sample of size n be drawn from a population with mean μ and variance σ^2. Also, let this sample have observations $x_1, x_2, \ldots, x_n$ which are independent and identically distributed. Then each x_i is an independent normal variate with mean μ_i and variance σ_i^2.

We have sample mean $(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}$ and sample variance $(s^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Therefore,

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu_i = \frac{1}{n}(n\mu) = \mu$$

$$V(\bar{x}) = V\left[\frac{x_1 + x_2 + \cdots + x_n}{n}\right] = \frac{1}{n^2}[V(x_1) + V(x_2) + \cdots + V(x_n)]$$

($\because x_i$'s are independent, the covariance terms are zero)

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

To test whether the given sample of size n has been drawn from a population with mean μ, i.e., to test whether the difference between the sample mean and the population mean is significant or not. Under the null hypothesis H_0 that there is no difference between the sample mean and the population mean, the standard error and test statistic are given by

$$\text{SE}(\bar{x}) = \frac{\sigma}{\sqrt{n}} \text{ and } z = \frac{\bar{x} - E(\bar{x})}{\text{SE}(\bar{x})} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Thus, mean $\bar{x}$ of the n independent normal variates (with mean μ and variance σ^2) follows normal distribution $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$.

If σ is not known, then approximate $\sigma \approx s$, where s is the standard deviation of the sample, and the approximate test statistic is given by $z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$.

If the level of significance is α and the critical value is z_α such that $-z_\alpha \le |z| = \left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right| \le z_\alpha$, then the confidence limits for the population mean μ are $\bar{x} \pm z_\alpha \frac{\sigma}{\sqrt{n}}$. Thus, at 5% level of significance, 95% confidence interval for μ is $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$, and at 1% level of significance, 99% confidence interval for μ is $\bar{x} - 2.58\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}$.

Central Limit Theorem

The sampling distribution of $\bar{x}$ is normally distributed with mean μ and variance σ^2/n. This result holds good even if the population is not normal for the large sample n. This result is established by the central limit theorem stated below.

ATTENDANCE SHEET
CSE DEPARTMENT

Name of Staff:					Month & Year: Aug 22			
Class: SY. B.Tech. (CSE) (A) 2022-23-I					Subject:			
DATE →	22/8	24	26/8	28/8		6/9	8/9	12/9
ISO	SIGN	SIGN	SIGN	SIGN				

If the variable x_i, $i = 1, 2, \ldots, n$, has a non-normal distribution with mean μ_i and variance σ_i^2, then $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is a random variable whose distribution approaches standard normal distribution [$N(0, 1)$] when $n \to \infty$. There is only one restriction upon the distribution of x_i that it has finite mean and variance.

Example 4.8 The mean weight obtained from a random sample of size 100 is 64 g. The standard duration of the weight distribution of the population is 3 g. Test the statement that the mean weight of the population is 67 g at 5% level of significance. Also, set up 99% confidence limits of the mean weight of the population.

Solution: Here, $n = 100$, $\mu = 67$, $\bar{x} = 64, \sigma = 3$

H_0: There is no significant difference between sample and population mean, i.e., $\mu = 67$

and H_1: $\mu \neq 67$ (two-tailed test)

Under H_0, test statistic $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{64 - 67}{3/\sqrt{100}} = -10$

The critical value of z at 5% level of significance for the two-tailed test is 1.96. Since $|z| = 10 > 1.96$, H_0 is rejected and we can conclude that the sample is not drawn from the population with mean 67.

Now, we have to find 99% confidence limits. They are given by

$$\bar{x} \pm 2.58\sigma/\sqrt{n} = 64 \pm 2.58(3/\sqrt{100}) = 64.774, 63.226$$

Example 4.9 If e is the permissible error for estimating the population parameter μ, then prove that the minimum sample size n required for estimating μ with 95% confidence is given by $n = (1.96\sigma/e)^2$, where σ^2 is the population variance. Hence, find the minimum sample size required at 95% confidence if the permissible error is 0.05 and $\sigma = 0.32$.

Solution: For large n, the test statistic z is given by $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

We know that $P(|z| \leq 1.96) = 0.95$ or $P\left(\left|\dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right| \leq 1.96\right) = 0.95$ or $P\left\{|\bar{x} - \mu| \leq 1.96\dfrac{\sigma}{\sqrt{n}}\right\} = 0.95$ (1)

We need n such that $P\{|\bar{x} - \mu| < e\} > 0.95$ (2)

Comparing (1) and (2), we get $\min e = \dfrac{1.96\sigma}{\sqrt{n}}$. Thus, $\dfrac{1.96\sigma}{\sqrt{n}} \leq e$, which gives $n \geq \left(\dfrac{1.96\sigma}{e}\right)^2$. Hence,

$$\min n = \left(\frac{1.96\sigma}{e}\right)^2$$

The minimum sample size n required at 95% confidence for $\sigma = 0.32$ and $e = 0.05$ is

$$n = \left(\frac{196\sigma}{e}\right)^2 = \left\{\frac{(1.96)(0.32)}{0.05}\right\}^2 = 157.35 = 158$$

Note: For 99% confidence, $\min\ n = \left(\frac{2.58\sigma}{e}\right)^2$

Example 4.10 The mean of a certain normal population is equal to the standard error of the mean of the samples of 100 from that distribution. Find the probability that the mean of the sample of 25 from the distribution will be negative?

Solution: Let μ be the mean and σ be the SD of the distribution. Then

$$\text{SE of the sample means} = \mu = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{100}} = \frac{\sigma}{10}$$

Also, for a sample of size 25, we have

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{25}} = \frac{\bar{x} - \sigma/10}{\sigma/5} = \frac{5\bar{x}}{\sigma} - \frac{1}{2}$$

Since $\bar{x}$ is negative, $z < -1/2$.

Therefore, the probability that a normal variate $z < -\frac{1}{2}$ $= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1/2} e^{-(1/2)z^2}\, dz = -\frac{1}{\sqrt{2\pi}} \int_{1/2}^{\infty} e^{-(1/2)z^2}\, dz$

$= 0.5 - 0.1915$

(Values from the Normal table in the Appendix)

$= 0.3085$

EXERCISE 4.3

1. A sample of 900 members is found to have a mean of 3.4 cm. Can it be reasonably regarded as a true random sample from a large population with mean 3.25 cm and SD 1.61 cm?
2. The average marks in mathematics of a sample of 100 students were 51 with SD of 6 marks. Could this have been a random sample from a population with average marks 50?
3. Let E be the permissible error in the means $\bar{x}$ and μ of sample and population, respectively. Let the standard deviation of the population be σ. Find the minimum sample size n if $P(|\bar{x} - \mu| < E) > 0.95$. Hence, find the minimum sample size to estimate the mean within 5 units of the true mean if $\sigma = 3$ and with 98% confidence.
4. A normal population has mean 0.2 and standard deviation 2. Find the probability that the mean of a sample of size 625 will be negative.
5. If the mean breaking strength of copper wire is 575 lbs with a standard deviation of 8.3 lbs, how large must a sample be used in order that there is one chance in 100 that the mean breaking strength of the sample is less than 572 lbs?
6. The mean value of a random sample of 144 items is 75 with standard deviation 15. Find 95% confidence limits for the population mean. Assume normal approximation to the sample. Also, find the minimum sample size to estimate the mean within 4 units of the true mean at 95% confidence limits.
7. An unbiased coin is thrown n times. It is desired that the relative frequency of the appearance of head should lie between 0.49 and 0.51. Find the smallest value of n that will ensure this result with 90% confidence.

ATTENDANCE SHEET
CSE DEPARTMENT

Name of Staff: Month & Year: Aug 22 Lecture

Class: SY. B.Tech. (CSE) (A) 2022-23-I Subject: Lectures

DATE → | 22/8 | 24 | 26/8 | 28/8 | | 5/9 | 8/9 | 12/9 | 13/9

8. A research worker wishes to estimate the mean of a population by using sufficiently large sample. The probability is 95% that the sample mean will not differ from the true mean by more than 25% of the SD. How large a sample should be taken?
9. The guaranteed average life of a certain type of bulbs is 1000 h with an SD of 125 h. It is decided to sample the output so as to ensure that 90% of the bulbs do not fall short of the guaranteed average by more than 2.5%. What must be the minimum size of the sample?
10. The density function of a random variable x is $f(x) = ke^{-2x^2+10}$. Find the upper 5% point of the distribution of mean of the random sample of size 25 from the above population.

4.5.4 Test for Difference Between Two Sample Means

Let $\bar{x}_1$ be the mean of the sample of size n_1 from a population with mean μ_1 and variance σ_1^2. Also, let $\bar{x}_2$ be the mean of the independent sample of size n_2 from another population with mean μ_2 and variance σ_2^2. Then

$$E(\bar{x}_1) = \mu_1,\ E(\bar{x}_2) = \mu_2,\ V(\bar{x}_1) = \frac{\sigma_1^2}{n_1} \text{ and } V(\bar{x}_2) = \frac{\sigma_2^2}{n_2}$$

Since for large samples $\bar{x}_1$ and $\bar{x}_2$ are normal variates, their difference is also a normal variate.
For developing the test of significance for the difference between sample proportions, the sample statistic is

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2,\ V(\bar{x}_1 - \bar{x}_2) = V(\bar{x}_1) + V(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Hence, $$\text{SE}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{4.5}$$

and the standardised variable corresponding to $\bar{x}_1 - \bar{x}_2$ is

$$z = \frac{\bar{x}_1 - \bar{x}_2 - E(\bar{x}_1 - \bar{x}_2)}{\text{SE}(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

We set up the hypothesis H_0 that there is no significant difference between the two population means, i.e., $H_0: \mu_1 = \mu_2$.
Under H_0, the standard error is the same as (4.5) and the test statistic is given by

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Under H_0: $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2 = \sigma$, the standard error and test statistic are given by

$$\text{SE}(\bar{x}_1 - \bar{x}_2) = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \text{ and } z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

Note:

(i) If $\sigma_1^2 \neq \sigma_2^2$ and σ_1^2, σ_2^2 are not known, then the test statistic is given by

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

where s_1^2 and s_2^2 are the variance of the samples of sizes n_1 and n_2, respectively.

(ii) If $\sigma_1^2 = \sigma_2^2 = \sigma$ and σ is not known, then the test statistic is given by

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$

Example 4.11 The means of simple samples of sizes 1000 and 2000 are 67.5 cm and 68.0 cm, respectively. Can the samples be regarded as drawn from the same population of standard deviation 2.5 cm?

Solution: Here, $\bar{x}_1 = 67.5, \bar{x}_2 = 68.0$

$n_1 = 1000, n_2 = 2000$

H_0: The samples are drawn from the same population with SD $\sigma = 2.5$, i.e., $\mu_1 = \mu_2$ with SD $\sigma = 2.5$

and $H_1: \mu_1 \neq \mu_2$ (two-tailed test)

Under H_0, test statistic $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \dfrac{67.5 - 68.0}{2.5\sqrt{\left(\frac{1}{1000} + \frac{1}{2000}\right)}} = \dfrac{-0.5}{2.5 \times 0.0387} = \dfrac{-0.5}{0.09675} = -5.17$

The critical value of z at 5% level of significance for the two-tailed test is 1.96. Since $|z| = 5.17 > 1.96$, H_0 is rejected and we can conclude that the samples cannot be regarded as drawn from the same population.

Example 4.12 The sizes, means and standard deviations of two samples are 500, 400; 28.57, 29.62 and 1.25, 1.42, respectively. Can we conclude that the samples are drawn from the same population with standard deviation σ?

Solution: Here, $n_1 = 500$, $\bar{x}_1 = 28.57$, $s_1 = 1.25$

$n_2 = 400$, $\bar{x}_2 = 29.62$, $s_2 = 1.42$

H_0: The samples are drawn from the same population with SD σ, i.e., $\mu_1 = \mu_2$ with SD σ

and $H_1: \mu_1 \neq \mu_2$ (two-tailed test)

Under H_0, test statistic $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Mon Tue
3:30 3:30

ATTENDANCE SHEET
CSE DEPARTMENT

Name of Staff: Month & Year: Aug 22 Lectures A

Class: SY. B.Tech. (CSE) (A) 2022-23-I Subject: Lectures Tak

DATE →	23/8	24	26/8	28/8			6/9	8/9	12/9	13/9	
ISO	SIGN	SIGN	SIGN							GN	S

Since σ^2 is not given, we have $\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} = \frac{500(1.25)^2 + 400(1.42)^2}{500 + 400} = 1.764$

Thus, we have $z = \frac{28.57 - 29.62}{1.328\sqrt{\frac{1}{500} + \frac{1}{400}}} = \frac{-1.05}{0.0891} = -11.78$

The critical value of z at 5% level of significance is 1.96. Since $|z| = 11.78 > 1.96$, H_0 is rejected and we may conclude that the samples cannot regarded as drawn from the same population.

Example 4.13 For sample I, $n_1 = 1000, \Sigma x_1 = 49,000, \Sigma(x_1 - \bar{x}_1)^2 = 7,84,000$. For sample II, $n_2 = 1500, \Sigma x_2 = 70,500, \Sigma(x_2 - \bar{x}_2)^2 = 24,00,000$. Discuss the significance of the difference of the sample means.

Solution: Here, H_0: There is no significant difference between the sample means, i.e., $H_0: \bar{x}_1 = \bar{x}_2$ and $H_1: \bar{x}_1 \neq \bar{x}_2$ (two-tailed test). Now,

$$s_1^2 = \frac{1}{n_1}\sum(x_1 - \bar{x}_1)^2 = \frac{784000}{1000} = 784, \quad s_2^2 = \frac{1}{n_2}\sum(x_2 - \bar{x}_2)^2 = \frac{1}{1500}(2400000) = 1600$$

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{49000}{1000} = 49, \bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{70500}{1500} = 47$$

Under H_0, test statistic $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{49 - 47}{\sqrt{\frac{784}{1000} + \frac{1600}{1500}}} = 1.470$

The critical value of z at 5% level of significance for the two-tailed test is 1.96. Since $|z| = 1.47 < 1.96$, H_0 is accepted and we can conclude that there is no significant difference between the sample means.

EXERCISE 4.4

1. The average income of people was ₹ 210 with an SD of ₹ 10 in a sample of 100 people of a city. For another sample of 150 people, the average income was ₹ 220 with an SD of ₹ 12. Test whether there is any significant difference between the average income of the localities.
2. A random sample of 200 villages from Coimbatore district gives the mean population per village as 485 with an SD of 50. Another random sample of the same size from the same district gives the mean population per village as 510 with an SD of 40. Is the difference between the mean values given by the two samples statistically significant? Justify your answer.
3. A simple sample of heights of 6400 Englishmen has a mean of 170 cm and an SD of 6.4 cm, while a simple sample of heights of 1600 Americans has a mean of 172 cm and an SD of 6.3 cm. Does the data indicate that Americans are, on an average, taller than Englishmen?
4. Two random samples of sizes 1000 and 2000 farms gave average yields of 2000 kg and 2050 kg, respectively. The variance of wheat farms in the country may be taken as 100 kg. Examine whether the two samples differ significantly in yield?