UNIT 4 ESTIMATION AND TESTS OF HYPOTHESES, STATISTICAL HYPOTHESES

PART-A
SHORT QUESTIONS WITH SOLUTIONS

Q1. Define population and sample with examples.

Model Paper-II, Q1(g)

Answer :

Population

The term population refers to information of group of observations about which inferences are to be made. Population size denoted as 'N' represents the number of objects or observations in the population. Population may be finite or infinite depending upon N being finite or infinite.

Examples

- Engineering students in Telangana
- Budget of India.

Sample

The term sample refers to a finite subset of the population. Sample size is represented as 'n' denoting the number of objects or observation in the sample.

Examples

- Engineering students of Deccan College
- Budget of Telangana.

Q2. A random sample of size 100 has a standard deviation of 5. What can you say about the maximum error with 95% confidence?

Model Paper-I, Q1(h)

Answer :

Given that,

Standard deviation, $\sigma = 5$

Samples size, $n = 100$

$Z_{\alpha/2}$ for 95% confidence = 1.96

Maximum error is expressed as,

$$E = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad \ldots (1)$$

Substituting the corresponding values in equation (1),

$$E = 1.96 \times \frac{5}{\sqrt{100}}$$

$$= 0.98$$

$$\therefore E = 0.98$$

Q3. It is desired to estimate the mean number of hours of continuous use until a certain computer will first require repairs. If it can be assumed that $\sigma = 48$ hours, how large a sample be needed so that one will be able to assert with 90% confidence that the sample mean is off by at most 10 hours?

Answer : Model Paper-II, Q1(h)

Given that,

Maximum error = 10 hours = E

$\sigma = 48$ hours

$Z_{\alpha/2} = 1.645$ [For 90%]

$$n = \left[\frac{Z_{\alpha/2}\sigma}{E}\right]^2 = \left[\frac{1.645 \times 48}{10}\right]^2$$

$$= 62.3$$

$\therefore \quad n \cong 62.$

Q4. Define hypothesis.

Answer :

Hypothesis means a mere assumption or some supposition to be proved or disproved. But for a researcher, hypothesis is a formal question that he intends to solve.

Thus, a hypothesis is defined as "A proposition or a set of propositions set forth as an explanation for the occurrence of some specified group of phenomena either asserted merely as a provisional conjecture to guide some investigation or accepted as highly probable in light of established facts.

Q5. What is statistical hypothesis and explain the types of it?

Answer : Model Paper-I, Q1(g)

Statistical Hypothesis

Statistical hypothesis are statements about the probability distributions of the populations. There are two types of hypothesis. They are,

(i) Null hypothesis

(ii) Alternative hypothesis.

(i) Null Hypothesis

To decide whether one procedure is better than another, we form the hypothesis that there is no difference between the procedures. Such hypothesis are known as 'null hypothesis' or 'simply hypothesis', are denoted by symbol H_0.

(ii) Alternative Hypothesis

Any hypothesis that differs from a given null hypothesis is called an 'alternative hypothesis'. The alternative hypothesis is denoted by symbol H_1.

Q6. Explain the level of significance.

Answer :

The level of significance refers to the probability of containing a statistic random value 't' in the critical region.

or

The level of significance also refers to the size of type I error.

The most commonly used levels of significance during the testing of hypothesis are 1% and 5%. Before gathering the information about the samples, the level of significance must be set up.

Q7. What do you mean by confidence interval?

Answer :

An estimate of a population parameter given by two numbers between which the parameter may be considered to be is called interval estimate of the parameter. The interval estimate or confidence interval consists of an upper confidence limit and lower confidence limit and we assign a probability that this interval contains the true population value. Confidence interval indicates how much confidence we want this interval will contain the population value.

Q8. Define Type-I and Type-II errors intesting of hypothesis.

Answer : Model Paper-III, Q1(g)

Type 1 Error

It involves rejection of null hypothesis when it is true.

Reject H_0 when it is true i.e., rejecting a correct hypothesis.

$\therefore \quad P(\text{Reject } H_0 \text{ when it is true})$

$= P(\text{Reject } H_0/H_1) = \alpha$

$= P(\text{Type I error}) = \alpha$

Type II Error

It involves acceptance of the null hypothesis when it is false and should be rejected.

Accept H_0 when it is wrong i.e., accepting a wrong hypothesis.

$\therefore \quad P(\text{Accept } H_0 \text{ when it is wrong})$

$= P(\text{Accept } H_0/H_1) = \beta$

$= P(\text{Type II error}) = \beta$

The size of type I and type II errors are also called as producer's risk and consumer's risk respectively.

Q9. A sample of 64 students have a mean weight of 70 kgs. Can this be regarded as a sample from a population with mean weight of 65 kgs and standard deviation of 25 kgs?

Answer :

Given that,

1. **Null Hypothesis H_0**

   $\mu = 70$ (sample of 64 students have mean weight of 70 kgs)

   Mean weight 65 kgs on standard deviation 25 kgs.

2. **Alternative Hypothesis**

   Cannot be regarded.

3. **Level of Significance**

   0.05 (not given, we have to assume)

4. **Computation**

   Test statistic,

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

   $\bar{x}$ = Mean of the sample = 70 kgs

   μ = Mean of the population = 65 kgs

   σ = Standard deviation of the population = 25 kgs

   n = Sample size = 64

$$Z = \frac{70 - 65}{\frac{25}{\sqrt{64}}} = 5 \times \frac{8}{25} = 1.6$$

5. **Decision**

   Accept the null hypothesis H_0, since $|Z| < Z_{\alpha/2}$.

Q10. Define point estiamtion.

Answer : Model Paper-III, Q1(h)

If a sample derived from an unknown population parameter, the calculation of a single value is done as an estimate. The procedure for determining the parameter is called '*point estimation*'. It also include concepts like estimator, estimate etc. The point estimate of a population parameter (θ) is a single numerical value.

A point estimator $\hat{\theta}$ is an unbiased estimator of the parameter θ if $\mu_{\hat{\theta}} = E(\hat{\theta}) = \theta$. In case of two unbiased estimators $\hat{\theta}_1$, $\hat{\theta}_2$ of the same population parameter θ, the estimator with small variance of the sampling distribution is chosen, i.e., if $\sigma^2_{\hat{\theta}_2} < \sigma^2_{\hat{\theta}_1}$ then $\hat{\theta}_2$ is the more efficient estimator of θ as compared to $\hat{\theta}_1$. Thus, out of all possible unbiased estimators of any parameter θ, the estimator with the smallest variance is the most efficient estimator of θ.

PART-B

ESSAY QUESTIONS WITH SOLUTIONS

4.1 ESTIMATION AND TESTS OF HYPOTHESES

4.1.1 Introduction, Statistical Inference

Q11. State and explain central limit theorem.

Answer :

Statement

If $\bar{x}$ is the random sample mean of size n taken from a population having the mean μ and finite variance σ^2, then the limiting form of the distribution of $Z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ as $n \to \infty$, is the standard normal distribution $N(0, 1)$.

The central limit theorem is one of the important concepts in statistics. It states that the distribution of sample means tends to be always normal distribution. This is true and also is free from shape of the population distribution, from where the sample is taken. The mean of sampling distribution is equal to the population mean and is independent of its sample size whether the population is normal (or) not.

As the sample size increases, the sampling distribution of the mean will approach normality. The relationship between shape of sampling distribution of mean and shape of population distribution is called as Central Limit Theorem.

If the random sample comes from a normal population, the sampling distribution of the mean is normal regardless of the size of the sample.

Q12. Discuss briefly about statistical inference?

Answer : Model Paper-III, Q8(a)

Statistical Inference

Statistical Inference or Inferential Statistics refers to the method that derives properties of the probability distributions with the help of analyzing the data. Its main goal is to study the characteristics of the population from a sample. The statistical inferences make use of random sample which is drawn from a population for defining/describing and making inferences regarding the population. Now a days, the trend is to differentiate methods such as classical method of estimation and Bayesian method. Among these two methods, the former one is a method of estimation of population parameter where in inferences are dependent on evidence collected from a random sample of the population. While the latter one utilizes knowledge regarding the probability distribution of unknown parameters along with evidence collected from sample data.

The two significant problems associated with the inferences are estimation and testing of hypothesis. Estimation is the statement/judgement made to determine an unknown population parameter where as testing of hypothesis is performed to decide whether one procedure is better than the another.

Consider the below examples to differentiate estimation and tests of hypothesis.

Example for Estimation

A member belonging to an association wants to estimate the true proportion of voters who are favourable to him by collecting evidence from a sample of one thousand voters. In the random sample, the fraction of voters who are favourable to that particular member can be utilized as an estimate of true proportion and the knowledge regarding sampling distribution of proportion allows to form degree of accuracy for that particular estimate.

Example for Tests of Hypothesis

Consider two machines A_1 and A_2. Among these two machines, the member needs to determine which machine is more efficient over another machine. The person can hypothesize that machine A_1 is more efficient than machine A_2 and can either accept or reject the hypothesis after performing required testing. Here, instead of estimating a parameter, a decision regarding prestated hypothesis is taken.

4.1.2 Classical Methods of Estimation

Q13. Explain briefly the following,

(i) Point estimation

(ii) Interval estimation.

Answer :

(i) Point Estimation

If a sample derived from an unknown population parameter, the calculation of a single value is done as an estimate. The procedure for determining the parameter is called '*point estimation*'. It also include concepts like estimator, estimate etc. The point estimate of a population parameter (θ) is a single numerical value.

A point estimator $\hat{\theta}$ is an unbiased estimator of the parameter θ if $\mu_{\hat{\theta}} = E(\hat{\theta}) = \theta$. In case of two unbiased estimators $\hat{\theta}_1$, $\hat{\theta}_2$ of the same population parameter θ, the estimator with small variance of the sampling distribution is chosen, i.e., if $\sigma_{\hat{\theta}_2}^2 < \sigma_{\hat{\theta}_1}^2$ then $\hat{\theta}_2$ is the more efficient estimator of θ as compared to $\hat{\theta}_1$. Thus, out of all possible unbiased estimators of any parameter θ, the estimator with the smallest variance is the most efficient estimator of θ. This is shown in the figure (1).

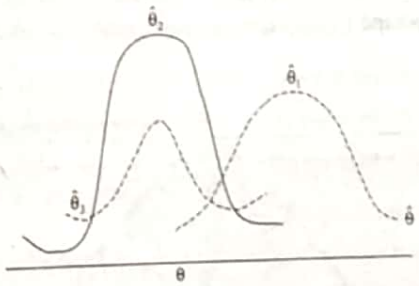


Figure (1): Sampling Distribution of Different Estimations of θ

The figure (1) depicts sampling distribution of the three estimators namely $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$ all of which estimates θ. Out of those three estimators, the distribution of $\hat{\theta}_2$ and $\hat{\theta}_3$ are centered at θ and hence are unbiased. Among $\hat{\theta}_2$ and $\hat{\theta}_3$, $\hat{\theta}_2$ has smaller variance than $\hat{\theta}_3$ and is finally considered as more efficient compared to $\hat{\theta}_2$.

But, even the efficient estimators cannot exactly estimate the population parameter. Thus, an interval is determined in which the value of the parameter is expected.

The sample mean estimate is rarely equal to the mean of population. Therefore, a point estimation is accompanied with the statement of error which provides the difference between an estimate and estimator. This is given by, $\bar{X} - \mu$.

For large values of 'n', the random variable $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is a normal variable.

$$\therefore \quad P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$$

$$P\left(-Z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow \quad -Z_{\alpha/2} < \frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < Z_{\alpha/2}$$

Let 'E' be the maximum error of estimates of $|\bar{X} - \mu|$

Thus, maximum error of estimates, $E = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

It is applicable to large samples.

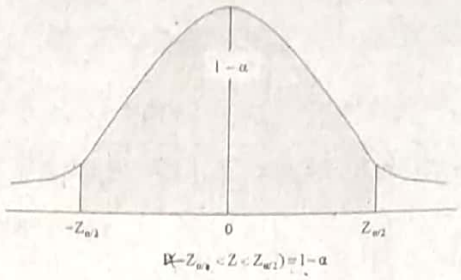


Figure (2)

For small samples, the maximum error of estimate can be taken as follows,

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$$

$t_{\alpha/2}$ is the t-value within $(n-1)$ degrees of freedom above which an area $\alpha/2$ is found.

$$P\left(-t_{\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1-\alpha \qquad \left(\because T = \frac{\bar{X}-\mu}{S/\sqrt{n}}\right)$$

Where,

S = Standard deviation

n = Size of sample.

Maximum error, $E = t_{\alpha/2} \cdot \dfrac{S}{\sqrt{n}}$

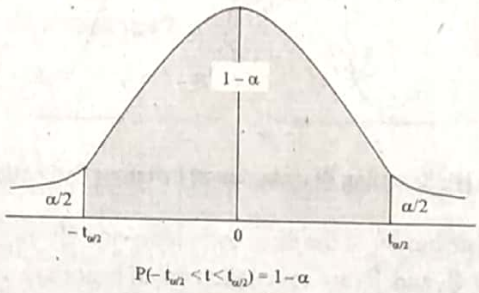


Figure (3)

(ii) Interval Estimation

Interval estimates of a population helps in estimating the interval containing the actual parameter value.

Let, 'α' be the probability that the estimated interval does not include the actual parameter value, then the probability that the estimated interval includes the actual parameter value will be $(1-\alpha)$.

If θ is a parameter and l, u are the lower and upper limits of the estimated interval. Then,

$$P(l < \theta < u) = 1 - \alpha$$

Where,

Interval (l, u) – Confidence interval

$(u - l)$ – Measure of precision of the estimate

$(1-\alpha)$ – Measure of reliability.

Q14. Show that S^2 is an unbiased estimator of the parameter σ^2.

Answer :

Model Paper-II, Q8(a)

Let, $x_1, x_2, x_3, \ldots, x_n$ be the sample taken from sample size n having mean $\bar{x}$, then

Finite population, $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ ... (1)

Taking expectation on both sides of equation (1), we get,

$$E(S^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$= E\left[\frac{1}{n-1}\cdot\frac{n}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$= E\left[\frac{n}{n-1}\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$= E\left[\frac{n}{n-1}\cdot S^2\right] \qquad \left[\because S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$= \left[\frac{n}{n-1}\right]\cdot E[S^2]$$

$$= \left[\frac{n}{n-1}\right]\cdot \sigma^2\left[\frac{n-1}{n}\right] \qquad \left[\because E(S^2) = \sigma^2\left[\frac{n-1}{n}\right]\right]$$

$$= \sigma^2$$

$$\therefore E(S^2) = \sigma^2$$

Hence, finite population S^2 is an unbiased estimator of the parameter σ^2.

Q15. Give the differences between point estimation and internal estimation.

Answer :

	Point Estimation		Interval Estimation
1.	Point estimation is an estimation of a single value for the unknown parameter.	1.	Interval estimation is an estimation of an interval for a range of values in which the unknown parameter lies.
2.	Point estimator is a statistical estimator whose value is geometrically represented by a point.	2.	Interval estimator is a statistical estimator whose value is geometrically represented by a set of points.
3.	Point estimation rarely coincides with the true value of the parameter, i.e., it does not provide accuracy of the estimate.	3.	Interval estimation provides accuracy of the estimate by providing an interval, i.e., likely to contain the true value parameter.
4.	Sample mean is an example of point estimation.	4.	Confidence interval is an example of interval estimation.

4.1.3 Estimating the Mean

Q16. Discuss in detail estimating the mean with single sample.

Model Paper-I, Q8(a)

Answer :

For large samples, i.e., $n \geq 30$, the random variable 'z' becomes approximately equal to the standard normal variate i.e.,

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \qquad \ldots (1)$$

Where,

σ – Standard deviation

n – Sample size

$\bar{x}$ – Sample mean

α – Population mean.

If $z_{\alpha/2}$ is some value of z, such that the area of the intervals $(z_{\alpha/2} < z < \infty)$ and $(-\infty < z < -z_{\alpha/2})$ is $\alpha/2$ and the area between the interval $(-z_{\alpha/2} < z < z_{\alpha/2})$ is $(1 - \alpha)$. Then, the probability that z lies in the interval $(-z_{\alpha/2}, z_{\alpha/2})$ is $(1 - \alpha)$.

$\Rightarrow \quad P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha$

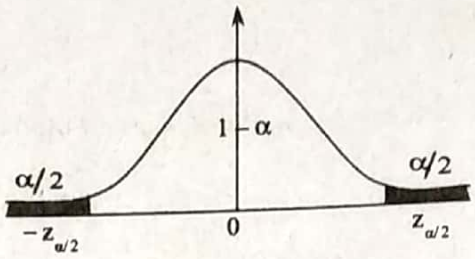


Figure: Sampling Distribution of Random Variables 'z'

As, $P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha$

As, $P\left(-z_{\alpha/2} < \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} < z_{\alpha/2}\right) = 1 - \alpha$

$P\left(-z_{\alpha/2} < \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} < z_{\alpha/2}\right) = 1 - \alpha$ [From equation (1)]

Multiplying by $\left(\frac{\sigma}{\sqrt{n}}\right)$ and subtracting $\bar{x}$ to each term in the inequality, we get,

$$P\left[-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) - \bar{x} < \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}\left(\frac{\sigma}{\sqrt{n}}\right) - \bar{x} < z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\bar{x}\right] = 1 - \alpha$$

Multiplying with '–' sign,

$$P\left[\bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < -\bar{x} + \mu + \bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = 1 - \alpha$$

$$P\left[\bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = 1 - \alpha$$

Thus, when the probability is $(1 - \alpha)$. Then, $(1 - \alpha)$ 100% confidence interval of α is given by,

$$\bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \quad \ldots (2)$$

Equation (2) is the relation used to find the confidence interval of large samples

Therefore,

$$\left[\bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\ \bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = \text{Confidence interval}$$

$$\bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) = \text{Upper confidence limit}$$

$$\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) = \text{Lower confidence limit}$$

Q17. Discuss in detail the estimation of mean of a population when the variance is unknown.

Answer :

Confidence Interval for Small Samples

For small sample, i.e., $n < 30$, the random variable 't' approximately becomes equal to the t-distribution.

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} \quad \ldots (1)$$

Where,

$\bar{x}$ – Sample mean

m – Population mean

s – Standard deviation of sample

$$s^2 - \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$

n – Random sample of a normal population.

If $t_{\alpha/2}$ is some value of t, such that the area of the intervals $(t_{\alpha/2} < t < \infty)$ and $(-\infty < t < -t_{\alpha/2})$ is $\alpha/2$ and the area between the interval $(-t_{\alpha/2} < t < t_{\alpha/2})$ is $(1 - \alpha)$, then the probability that t lies in the interval $(-t_{\alpha/2}, t_{\alpha/2})$ is $(1 - \alpha)$.

$$\Rightarrow \quad P(-t_{\alpha/2} < t < t_{\alpha/2}) = 1 - \alpha$$

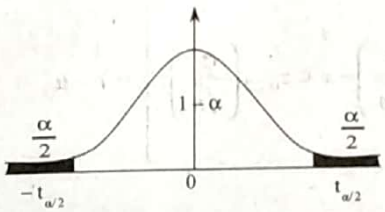


Figure: Sampling Distribution of Random Variable 't'

As, $P(-t_{\alpha/2} < t < t_{\alpha/2}) = 1 - \alpha$

$$\Rightarrow \quad P\left[-t_{\alpha/2} < \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} < t_{\alpha/2}\right] = 1 - \alpha \qquad \text{[From equation (1)]}$$

Multiplying by $\left(\frac{s}{\sqrt{n}}\right)$ and subtracting $\bar{x}$ to each term in the inequality, we get,

$$P\left[-t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) - \bar{x} < \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}\left(\frac{s}{\sqrt{n}}\right) - \bar{x} < t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) - \bar{x}\right] = 1 - \alpha$$

Now, multiply with '–' sign

$$P\left[\bar{x} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < -\bar{x} + \mu + \bar{x} < \bar{x} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right] = 1 - \alpha$$

$$P\left[\bar{x} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{x} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right] = 1 - \alpha$$

Thus, when the probability is $1-\alpha$, then $(1-\alpha)$ 100% confidence interval of α is given by,

$$\bar{x} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{x} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \quad \ldots(2)$$

Equation (2) is the relation used to find the confidence interval of small samples.

Therefore,

$$\left[\bar{x} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right), \bar{x} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right] = \text{Confidence interval}$$

$$\bar{x} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) = \text{Upper confidence limit}$$

$$\bar{x} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) = \text{Lower confidence limit.}$$

PROBLEMS

Q18. The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 gms. per ml. find the 95% and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gm per ml.

Solution :

For 95% Confidence Interval

Given that,

$n = 36$

$\sigma = 0.3$

$z_{\alpha/2} = 1.96$ (For 95%)

$\bar{x} = 2.6$

$\therefore$ The confidence interval is given by,

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

$$\therefore \quad z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \frac{1.96 \times 0.3}{\sqrt{36}}$$

$$= \frac{1.96 \times 0.3}{6}$$

$$= 0.098$$

$\Rightarrow \quad \bar{x} - 0.098 < \mu < \bar{x} + 0.098$

$\Rightarrow \quad 2.6 - 0.098 < \mu < 2.6 + 0.098$

$\Rightarrow \quad 2.502 < \mu < 2.698$

$\therefore \quad 2.502 < \mu < 2.698$

$\therefore$ The confidence interval for mean is $2.502 < \mu < 2.698$.

For 99% Confidence Interval

$n = 36$

$\sigma = 0.3$

$z_{\alpha/2} = 2.58$ (For 99%)

$\bar{x} = 2.6$

$\therefore$ The confidence interval is given by,

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

$$\therefore \quad z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \frac{2.58 \times 0.3}{\sqrt{36}}$$

$$= \frac{2.58 \times 0.3}{6}$$

$$= \frac{0.774}{6}$$

$$= 0.129$$

$\Rightarrow \quad \bar{x} - 0.129 < \mu < \bar{x} + 0.129$

$\Rightarrow \quad 2.6 - 0.129 < \mu < 2.6 + 0.129$

$\Rightarrow \quad 2.471 < \mu < 2.729$

$\therefore \quad 2.471 < \mu < 2.729$

$\therefore$ The confidence interval for mean is $2.471 < \mu < 2.729$.

Q19. How large a sample is required to be 95% confident that our estimate of μ is off by less than 0.05. Assume that the population standard deviation is 0.3.

Model Paper-I, Q8(b)

Solution :

Given that,

Population standard deviation, $\sigma = 0.3$

$$e = 0.05$$

The 95% confidence limit is,

$\Rightarrow \quad (1 - \alpha)\ 100 = 95$

$\Rightarrow \quad 1 - \alpha = \frac{95}{100}$

$\Rightarrow \quad 1 - \alpha = 0.95$

$\Rightarrow \quad \alpha = 0.05$

$$\frac{\alpha}{2} = 0.025$$

From the table of areas, the value of σ, $z_{0.025}$ is 1.96

$\therefore \quad z_{\alpha/2} = 1.96$

If $\bar{x}$ is an estimate of 'μ' then we can be $100\ (1 - \alpha)\%$ confident that the error will not exceed a specified amount 'e' when the sample size is,

$$n = \left(\frac{z_{\alpha/2}.\sigma}{e}\right)^2$$

$$n = \left(\frac{z_{0.025}.\sigma}{e}\right)^2$$

$$n = \left(\frac{(1.96)(0.3)}{0.05}\right)^2$$

$$n = \left(\frac{0.588}{0.05}\right)^2$$

$$n = (11.76)^2$$

$$\boxed{n = 138.3}$$

Hence, when the sample size is 138 then we can be 95% confident that an estimate of 'μ' will not exceed 0.05.

Q20. In a psychological testing experiment, 25 subjects are selected randomly and their reaction time in seconds to a particular stimulus is measured. Past experience suggests that the variance in reaction times to these types of stimuli is 4 sec² and that the distribution of reaction times is approximately normal. The average time for the subjects is 6.2 second. Give an upper 95% bound for the mean reaction time.

Solution :

Given that,

Sample size, $n = 25$

Standard deviation, $\sigma = 4$

Mean, $\bar{x} = 6.2$ seconds

An upper 95% bound for the mean reaction time is,

$$\bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} \quad \ldots (1)$$

The 95% confidence limit is,

$\Rightarrow (1-\alpha)100 = 95$

$\Rightarrow 1-\alpha = \frac{95}{100}$

$\Rightarrow 1-\alpha = 0.95$

$\Rightarrow \alpha = 0.05$

From table of areas the value of $z_{0.05}$ is 1.645

$z_\alpha = 1.645$

Substituting the values of $\bar{x}, z_\alpha, \sigma$ and n in equation (1), we get,

$\Rightarrow 6.2 + (1.645)\frac{4}{\sqrt{25}}$

$\Rightarrow 6.2 + 1.316$

$\Rightarrow 7.516$ seconds

Therefore, we are 95% confident that the mean reaction time is less than 7.516 seconds.

Q21. The contents of 7 similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2 and 9.6 liters. Find a 95% confidence interval for the mean of all such containers, assuming an approximate normal distribution.

Solution :

Given that,

Sample size, $n = 7$

Mean of the sample

$$\bar{x} = \frac{\Sigma x}{n}$$

$$\Rightarrow \bar{x} = \frac{9.8+10.2+10.4+9.8+10.0+10.2+9.6}{7}$$

$$\Rightarrow \bar{x} = \frac{70}{7} = 10$$

$\therefore \bar{x} = 10$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
9.8	−0.2	0.04
10.2	0.2	0.04
10.4	0.4	0.16
9.8	−0.2	0.04
10.0	0	0
10.2	0.2	0.04
9.6	−0.4	0.16
$\Sigma x_i = 70$		$\Sigma(x_i - \bar{x})^2 = 0.48$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{0.48}{7-1} = \frac{0.48}{6}$$

$\Rightarrow s^2 = 0.08$

$\Rightarrow s = \sqrt{0.08}$

$\therefore s = 0.28$

Confidence limit = 95%

$\Rightarrow (1-\alpha)100 = 95\%$

$\Rightarrow 1-\alpha = 95/100$

$\Rightarrow 1-\alpha = 0.95$

$\Rightarrow \alpha = 1 - 0.95$

$\Rightarrow \alpha = 0.05$

$\Rightarrow \frac{\alpha}{2} = \frac{0.05}{2}$

$\therefore \frac{\alpha}{2} = 0.025$

Tabulated value of $t_{\alpha/2}$ for $(n-1)$ *i.e.*, $(7-1) = 6$ d.f is 2.44

$$t_{\alpha/2} = 2.44$$

The confidence interval is,

$$\left[\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

$$\Rightarrow \left[10 - 2.44\times\frac{0.28}{\sqrt{7}}, 10 + 2.44\times\frac{0.28}{\sqrt{7}}\right]$$

$$\Rightarrow \left[10 - \frac{0.6832}{2.645}, 10 + \frac{0.6832}{2.645}\right]$$

$\Rightarrow (9.741, 10.258)$

$\therefore$ The confidence interval is, (9.741,10.28).

Q22. Scholastic aptitude test (SAT) mathematics scores of a random sample of 500 high school seniors in the state of Texas are collected and the sample mean and standard deviation are found to be 501 and 112, respectively. Find a 99% confidence interval on the mean SAT mathematics score for seniors in the state of Texas.

Solution :

Given that,

Sample mean, $\bar{x} = 501$

Sample size, $n = 500$

Standard deviation, $s = 112$

Normal approximation is used as the size of random sample is large

Confidence limit = 99%

$$(1 - \alpha) = \frac{99}{100}$$

$$\Rightarrow \quad 1 - \alpha = 0.01$$

$$\Rightarrow \quad \alpha = 0.01$$

$$\frac{\alpha}{2} = 0.005$$

From table areas under the normal curve $z_{\alpha/2}$ i.e., $z_{0.005}$ is 2.575.

$$z_{\alpha/2} = 2.575$$

Confidence interval for large samples is given by,

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Therefore, 95% confidence interval for μ is,

$$\Rightarrow \quad 501 \pm (2.575)\left(\frac{112}{\sqrt{500}}\right)$$

$$\Rightarrow \quad 501 \pm (2.575)\left(\frac{112}{22.36}\right)$$

$$\Rightarrow \quad 501 \pm (2.575)\,(5)$$

$$\Rightarrow \quad 501 \pm 12.9$$

Which reduces to $488.1 < \mu < 513.9$.

4.1.4 Standard Error of a Point Estimate, Prediction Intervals, Tolerance Limits

Q23. Describe the following,

(i) Standard error of point estimate

(ii) Prediction intervals

(iii) Tolerance limits.

Model Paper-I, Q9(a)

Answer :

(i) Standard Error of Point Estimate

Consider the estimator $\overline{X}$ of μ with σ known. It is known that measure of the quality of an unbiased estimator is indeed its variance. The variance of $\overline{X}$ is given by,

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

Therefore, the standard deviation (standard error) of $\overline{X}$ is,

$$\text{s.e} = \sqrt{\sigma^2_{\overline{X}}}$$

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

Standard Error (s.e) = $\frac{\sigma}{\sqrt{n}}$

Case (i)

For $\overline{X}$, confidence limits on μ with σ^2 known is given by,

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

It can also be written as $\overline{X} \pm z_{\alpha/2}\, s.e(\bar{x})$ where "s.e" represents standard error.

Case (ii)

For $\overline{X}$, confidence interval on μ with σ^2 unknown is given by,

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

It can also be written as $\bar{x} \pm t_{\alpha/2}\, s.e(\bar{x})$ where "s.e" represents standard error.

(ii) Prediction Interval

Prediction Interval (PI) is one form of confidence interval which is employed along with predictions in regression analysis. PI represents a range that forecosts the new observation based on the current model.

A $100(1-\alpha)\%$ prediction interval of a future observation 'x_0' for a normal distribution of measurements with unknown mean 'μ' and known variance 'σ^2' is given by,

$$\overline{X} - z_{\alpha/2}\, \sigma \sqrt{1+\frac{1}{n}} < x_0 < \bar{x} + z_{\alpha/2}\, \sigma \sqrt{1+\frac{1}{n}}$$

Where "$z_{\alpha/2}$" represents the z-value that leaves an area of $\frac{\alpha}{2}$ to the right side.

(iii) Tolerance Limits

Tolerance limits for a normal distribution of measurements with unknown mean 'μ' and standard deviation 'σ' are given by,

$$\bar{x} \pm ks$$

Where value 'k' is calculated in such a way that an analyst can declare with $100(1-\gamma)\%$ confidence that tolerance limits covers minimum proportion $(1-\alpha)$ of the population measurements.

PROBLEMS

Q24. Due to the decrease in interest rates, the First Citizens Bank received a lot of mortage applications. A recent sample of 50 mortage loans resulted in an average loan amount of $257,300. Assume a population standard deviation of $25,000. For the next customer who fills out a mortage application, find a 95% prediction interval for the loan amount.

Solution :

Model Paper-II, Q8(b)

Given that,

Mean, $\bar{x} = \$257{,}300$

Standard deviation, $\sigma = \$25{,}000$

Sample size, $n = 50$

A $100(1-\alpha)\%$ prediction interval is,

$$\bar{x} - z_{\alpha/2}\, \sigma \sqrt{\frac{1+1}{n}} < x_0 < \bar{x} + z_{\alpha/2}\, \sigma \sqrt{1+\frac{1}{n}}$$

Confidence limit = 95%

$\Rightarrow (1-\alpha)100 = 95$

$\Rightarrow 1-\alpha = \frac{95}{100}$

$\Rightarrow 1-\alpha = 0.95$

$\Rightarrow \alpha = 0.05$

$\Rightarrow \frac{\alpha}{2} = 0.025$

From table ares of normal curve $z_{\alpha/2}$ i.e., $z_{0.025}$ is 1.96

$\therefore z_{\alpha/2} = 1.96.$

Therefore, 95% prediction interval for the future loan amount is,

$$257300 - (1.96)(25{,}000)\sqrt{1+\frac{1}{50}} < x_0 < 257300 + (1.96)(25{,}000)\sqrt{1+\frac{1}{50}}$$

$$257300 - (1.96)(25{,}000)\sqrt{1.02} < x_0 < 257300 + (1.96)(25{,}000)\sqrt{1.02}$$

$$257300 - (1.96)(25{,}000)(1.009950494) < x_0 < 257300 + (1.96)(25{,}000)(1.009950494)$$

$$257300 - 49487.57421 < x_0 < 257300 + 49487.57421$$

$$207{,}812.43 < x_0 < 306{,}787.57$$

Therefore, the prediction interval is ($207,812.43, $306,787.5).

Q25. A meat inspector has randomly selected 30 packs of 95% lean beef. The sample resulted in a mean of 96.2% with a sample standard deviation of 0.8%. Find a 99% prediction interval for the leanness of a new pack. Assume normality.

Solution :

Given that,

Mean, $\bar{x} = 96.2\%$

Standard deviation, $s = 0.8\%$

Sample size, $n = 30$

Confidence limit = 99%

$\Rightarrow (1-\alpha)100 = 99$

$\Rightarrow (1-\alpha) = \frac{99}{100}$

$\Rightarrow (1-\alpha) = 0.99$

$\Rightarrow \alpha = 0.01$

$\frac{\alpha}{2} = 0.005$

The critical value of $t_{0.005}$ at $v = 29$ degrees of freedom using t-distribution is 2.756.

We have, $100(1-\alpha)\%$ prediction interval for a normal distribution of measurements with unknown mean 'μ' and unknown variance 'μ^2'. It is given by,

$$\bar{x} - z_{\alpha/2}\, s\sqrt{1+\frac{1}{n}} < x_0 < \bar{x} + z_{\alpha/2}\, s\sqrt{1+\frac{1}{n}} \quad \ldots (1)$$

Therefore, 99% prediction interval for x_0 is obtained by substituting the values of $\bar{x}$, s, n and $t_{\alpha/2}$ in equation (1),

$$96.2 - (2.756)(0.8)\sqrt{1+\frac{1}{30}} < x_0 < 96.2 + (2.756)(0.8)\sqrt{1+\frac{1}{30}}$$

$$96.2 - 2.24124 < x_0 < 96.2 + 2.24124$$

Which reduces to,

$$93.96 < x_0 < 98.44$$

Therefore, the prediction interval is, (93.96, 98.44).

Q26. A meat inspector has randomly selected 30 packs of 5% lean beef. The sample resulted in a mean of 96.2% with a sample standard deviation of 0.8%. Find tolerance interval that gives two sided 95% bounds on 90% of the distribution of packages of 95% lean beef. Assume the data came from an approximately normal distribution.

Solution :

Given that,

Sample size, $n = 30$

Mean, $\bar{x} = 96.2\%$

Standard deviation, $s = 0.8\%$

We have, Tolerance limits for a normal distribution of measurements with unknown mean 'μ' and unknown standard deviation 'σ' are given by,

$$\bar{x} \pm ks \quad \ldots (1)$$

The value of k from Table Tolerance factors for Normal distribution for $1 - \alpha = 0.90$, $\gamma = 0.05$ and $n = 30$ is 2.14 for two-sided limits.

On substituting the values of $\bar{x}$, k and s in equation (1), we get,

$\Rightarrow$ $96.2 \pm (2.14)\,(0.8)$

$\Rightarrow$ 96.2 ± 1.712

The upper bound and lower bound are 97.9 94.5 respectively.

Therefore, we can be 95% confident that the resulting range covers the central 90% of the distribution of 95% lean beef packages.

Q27. A machine produces metal pieces that are cylindrical in shape. A sample of these pieces is taken and the diameters are found to be 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01 and 1.03 centimeters. Use these data to calculate three interval types and draw interpretations that illustrate the distinction between them in the context of the system for all computations assume an approximately normal distribution. The sample mean and standard deviation for the given data are $\bar{x} = 1.0056$ and $s = 0.0246$.

(a) Find a 99% confidence interval on the mean diameter

(b) Compute 99% prediction interval on a measured diameter of a single metal piece taken from the machine.

(c) Find the 99% tolerance limits that will contain 95% of the metal pieces produced by this machine.

Solution :

Given that,

Mean, $\bar{x} = 1.0056$

Standard deviation, $s = 0.0246$

Sample size, $n = 9$

(a) The 99% Confidence Interval on the Mean Diameter

The 99% confidence interval for the mean diameter is given by,

$$\bar{x} \pm t_{0.005} \frac{s}{\sqrt{n}} \quad \ldots (1)$$

Confidence limit = 99%

$\Rightarrow$ $(1 - \alpha) = \frac{99}{100}$

$\Rightarrow$ $\alpha = 0.01$

$\Rightarrow$ $\frac{\alpha}{2} = 0.005$

The critical value of $t_{\alpha/2}$ i.e., $t_{0.005}$ for t-distribution is 3.355. Substituting the values of $\bar{x}$, s, $t_{\alpha/2}$ and 'n' in equation (1), we get,

$\Rightarrow$ $1.0056 \pm (3.355)\,\frac{(0.0246)}{\sqrt{9}}$

$\Rightarrow$ $1.0056 \pm \frac{(3.355)\,(0.0246)}{3}$

$\Rightarrow$ 1.0056 ± 0.0275

Therefore, 99% confidence bounds are 0.9781 and 1.0331.

(b) The 99% Prediction Interval on a Measured Diameter of a Single Metal Piece taken from Machine

The 99% prediction interval for future observation is given by,

$$\bar{x} \pm t_{0.005} s \sqrt{1 + \frac{1}{n}} \quad \ldots (2)$$

Confidence limit = 99%

$\Rightarrow \quad (1-\alpha) = \frac{99}{100}$

$\Rightarrow \quad \alpha = 0.01$

$\Rightarrow \quad \frac{\alpha}{2} = 0.005$

The critical value of $t_{\alpha/2}$ i.e., $t_{0.005}$ for t-distribution at 8 degrees of freedom is 3.355.

On substituting the values of $\bar{x}$, $t_{\alpha/2}$, s and n in equation (2), we get,

$\Rightarrow \quad 1.0056 \pm (3.355)\ (0.0246\ \sqrt{1+\frac{1}{9}}$

$\Rightarrow \quad 1.0056 \pm (3.355)\ (0.0246)\ \sqrt{1.11}$

$\Rightarrow \quad 1.0056 \pm (3.355)\ (0.0246)\ (1.054)$

$\Rightarrow \quad 1.0056 \pm 0.0869$

Therefore, the upper bound and lower bound are 1.0926 and 0.9187, respectively.

(c) The 99% Tolerance Limits that will Contain 95% of the Metal Pieces Produced by this Machine

The 99% tolerance limits are given by,

$$\bar{x} \pm ks \qquad \ldots (3)$$

The value of k from Table Tolerance factors for normal distribution for $n = 9$, $1 - \gamma = 0.99$ and $1 - \alpha = 0.95$ is 4.550 for two-sided limits

On substituting the values of $\bar{x}$, k and s in equation (3), we get,

$\Rightarrow \quad 1.0056 \pm (4.550)\ (0.0246)$

$\Rightarrow \quad 1.0056 \pm 0.1119$

Therefore, the upper bound and lower bound are 1.1175 and 0.8937, respectively. We are 99% confident that the above range covers the central 95% of the distribution of diameters produced.

4.1.5 Estimating the Variance

Q28. Explain about estimation of variance with single sample.

Answer : **Model Paper-II, Q9(a)**

The sample variance S^2 is computed from a random sample of size 'n' drawn from a normal population associated with variance σ^2. The statistic S^2 serves as a point estimate of σ^2. Therefore, S^2 is referred to as an estimate of σ^2.

An interval estimate of σ^2 is created by the statistic.

$$X^2 = \frac{(n-1)S^2}{\sigma^2} \qquad \ldots (1)$$

When samples are drawn from a normal population, the statistic χ^2 contains chi-squared distribution associated with $(n-1)$ degrees of freedom.

$$P\left(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}\right) = 1 - \alpha \qquad \ldots (2)$$

Where, $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ represents chi-squared distribution values associated with $(n-1)$ degrees of freedom. The for more one leaves an area of $1 - \frac{\alpha}{2}$ to the right side and the latter one leaves an area of $\frac{\alpha}{2}$ to the right side.

The figure illustrating $P\left(\chi^2_{1-\alpha/2} \le X^2 < \chi^2_{\alpha/2}\right) = 1-\alpha$ is given below,

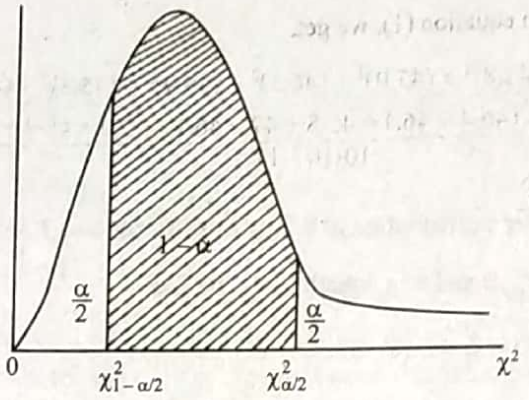


Figure: $P\left(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}\right) = 1-\alpha$

Now, substituting the value of X^2 in (2), we get,

$$P\left[\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}\right] = 1-\alpha$$

Now, dividing every term on the L.H.S by $(n-1)S^2$, we get,

$$\Rightarrow \quad P\left[\frac{\chi^2_{1-\alpha/2}}{(n-1)S^2} < \frac{(n-1)S^2}{\sigma^2(n-1)S^2} < \frac{\chi^2_{\alpha/2}}{(n-1)S^2}\right] = 1-\alpha$$

$$P\left[\frac{\chi^2_{1-\alpha/2}}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{\alpha/2}}{(n-1)S^2}\right] = 1-\alpha$$

Now, reversing each term on L.H.S we get,

$$P\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right] = 1-\alpha$$

Confidence Interval for σ^2

When S^2 is the variance of sample size 'n' drawn from normal population then a $100(1-\alpha)\%$ confidence interval is given by,

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$$

PROBLEM

Q29. The following are the weights, in decagrams of 10 packages of grass seed distributed by a certain company 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2 and 46.0. Find a 95% confidence interval for the variance of the weights of all such packages of grass seed distributed by this company, assuming a normal population.

Solution :

Given that,

Sample size, $n = 10$

We have, $\boxed{s^2 = \dfrac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}}$

Confidence limit = 95%

$\Rightarrow \quad (1-\alpha)100 = 95$

$\Rightarrow \quad (1-\alpha) = \dfrac{95}{100}$

$\Rightarrow \quad \alpha = 1 - 0.95$

$\Rightarrow \quad \alpha = 0.05$

$\dfrac{\alpha}{2} = 0.025$

The critical value of $\chi^2_{0.025}$ and $\chi^2_{0.975}$ using chi-squared distribution table at $9(v = n = 1)$ degrees of freedom are 19.023 and 2.700.

Substituting the values of x_i and n in equation (1), we get,

$$s^2 = \frac{10\left((46.4)^2 + (46.1)^2 + (45.8)^2 + (47.0)^2 + (46.1)^2 + (45.9)^2 + (45.8)^2 + (46.9)^2 + (45.2)^2 + (46.0)^2\right) - (46.4 + 46.1 + 45.8 + 47 + 46.1 + 45.9 + 45.8 + 46.9 + 45.2)^2}{10(10-1)}$$

$$s^2 = \frac{10(21273.12) - (461.2)^2}{(10)(9)}$$

$$s^2 = \frac{212731.2 - 212705.44}{90}$$

$$s^2 = \frac{25.76}{90}$$

$$s^2 = 0.286$$

We have, $100(1-\alpha)\%$ confidence interval for σ^2 is,

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}} \quad \ldots (2)$$

Therefore, 95% confidence interval for σ^2 is obtained by substituting the values of n, s, $\chi^2_{0.025}$ and $\chi^2_{0.975}$ in equation (2), we get,

$$\Rightarrow \quad \frac{(10-1)(0.286)}{19.023} < \sigma^2 < \frac{(10-1)(0.286)}{2.7}$$

$$\Rightarrow \quad \frac{2.574}{19.023} < \sigma^2 < \frac{2.574}{2.7}$$

$$0.135 < \sigma^2 < 0.953$$

Therefore, the 95% confidence interval for σ^2 is reduced to, $0.135 < \sigma^2 < 0.953$.

4.1.6 Estimating a Proportion for Single Mean

Q30. Explain about estimation of proportion for single sample.

Answer :

In a binomial experiment, a point estimator of the proportion p is,

$$\hat{p} = \frac{X}{n}$$

Where 'X' denotes the number of successes in 'n' trials.

Hence, the sample proportion $\hat{p}$ serves as the point estimate of parameter p.

When unknown proportion p is not extremely close to 0 or 1 then a confidence interval for $\hat{p}$ can be formal by using the sampling distribution of $\hat{p}$.

According to central limit theorem, for large value of n, $\hat{p}$ is normally distributed with mean,

$$\mu_{\hat{p}} = E(\hat{p})$$

$$= E\left(\frac{X}{n}\right)$$

$$= \frac{np}{n}$$

$$= p$$

Variance, $\sigma_{p^2} = \sigma^2_{X/n}$

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

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

$$= \frac{pq}{n}$$

Hence,

$$P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha \quad \ldots (1)$$

Where, $z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}}$

Here, $z_{\alpha/2}$ represents an area of $\frac{\alpha}{2}$ under the standard normal curve.

Substituting the value of z in equation (1), we get,

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{pq/n}} < z_{\alpha/2}\right) = 1 - \alpha \quad \ldots (2)$$

For large values of n, substitution of point estimate $\hat{p} = \frac{x}{n}$ in place of p produces a very small error. Therefore, equation (2) can be written as,

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) \approx 1 - \alpha$$

Solving for p in the above equation (2), we get,

$$-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}} < z_{\alpha/2}$$

This is an alternative form of the confidence interval for p associated with limits.

$$\frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n}}{1 + \frac{z_{\alpha/2}^2}{n}} \pm \frac{z_{\alpha/2}}{1 + \frac{z_{\alpha/2}^2}{n}}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z_{\alpha/2}^2}{4n^2}}$$

Large Sample Confidence Interval for P

When $\hat{p}$ represents the proportion of successes in a sample of size 'n' and $\hat{q} = 1 - \hat{p}$ then an approximate $100(1 - \alpha)\%$ confidence interval for the binomial parameter p is,

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Where, $z_{\alpha/2}$ represents a z-value that leaves an area of $\frac{\alpha}{2}$ to the right side.

PROBLEMS

Q31. A certain machine makes electrical resistors having a mean resistance of 40 ohms and a standard deviation of 2 ohms. What is the probability that a random sample of 36 of these resistors will have a combined resistance of more than 1458 ohms?

Solution :

Given that,

$\mu = 40$ ohms

$\sigma = 2$ ohms

$n = 36$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{2}{\sqrt{36}} = \frac{2}{6}$$

$$= \frac{1}{3}$$

$$P\left(\sum_{i=1}^{n} X_i < 1458\right) = P\left(\frac{\sum_{i=1}^{n} X_i}{n} > \frac{1458}{36}\right)$$

$$= P(\overline{X} > 40.5)$$

$$= P\left(\frac{\overline{X} - \mu}{\sigma_{\bar{x}}} > \frac{40.5 - 40}{\frac{1}{3}}\right)$$

$$= P\left(\frac{\overline{X} - \mu}{\sigma_{\bar{x}}} > \frac{0.5}{\frac{1}{3}}\right)$$

$$= P(Z > 1.5)$$

$$= 1 - P(Z < 1.5) = 1 - 0.9332$$

$$= 0.0668.$$

Q32. In a random sample of n = 500 families owing television sets in the city of Hamilton, Canada, it is found that x = 340 subscribe to HBO. How large a sample is required if we want to be 95% confident that our estimate of P is within 0.02 of the true value?

Solution : **Model Paper-III, Q8(b)**

Given that,

Preliminary sample, $n = 500$

$e = 0.02$

The point estimation of proportion,

$$\hat{p} = \frac{X}{n}$$

$$= \frac{340}{500}$$

$$\boxed{\hat{p} = 0.68}$$

$$\hat{q} = 1 - \hat{p}$$

$$= 1 - 0.68$$

$$\boxed{\hat{q} = 0.32}$$

If $\hat{p}$ is an estimate of p, then we can be $100(1 - \alpha)\%$ confident that the error will be less than specified amount e when the sample is given by,

$$\boxed{n = \frac{Z_{\alpha/2}^2 \, \hat{p} \, \hat{q}}{e^2}} \quad \ldots (1)$$

Confidence limit = 95%

$\Rightarrow$ $(1 - \alpha)100 = 95$

$$1 - \alpha = \frac{95}{100}$$

$$1 - \alpha = 0.95$$

$$\alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

The value of $z_{0.025}$ using Table Areas under Normal Curve is 1.96.

Now, substituting the values of $\hat{p}$, $\hat{q}$, e and $z_{0.025}$ in equation (1), we get,

$$\Rightarrow \quad n = \frac{(1.96)^2 (0.68)(0.32)}{(0.02)^2}$$

$$= \frac{0.8359}{4 \times 10^{-4}} = 2089.7$$

$$\approx 2090$$

Hence, if the sample size is 2090 then we are 95% confident that sample proportion will not exceed from the true proportion by more than 0.02.

Q33. In a random sample of n = 500 families owing television sets in the city of Hamilton, Canada, it is found that x = 340 subscribe to HBO. How large a sample is required if we want to be atleast 95% confident that our estimate of p is within 0.02 of the true value?

Solution :

Given that,

$n = 500$

$X = 340$

$e = 0.02$

Confidence limit = 95%

$\Rightarrow \quad (1-\alpha) = 95$

$\Rightarrow \quad 1-\alpha = \frac{95}{100}$

$\Rightarrow \quad 1-\alpha = 0.95$

$\Rightarrow \quad \alpha = 0.05$

$\therefore \quad \frac{\alpha}{2} = 0.025$

The value of $z_{0.025}$ from table areas under normal curve is 1.96.

We have, sample size, $n = \frac{z_{\alpha/2}^2}{4e^2}$ ... (1)

Substituting the values of $z_{\alpha/2}^2$ and e in equation (1), we get,

$$n = \frac{(1.96)^2}{4 \times (0.02)^2}$$

$$= \frac{3.842}{4 \times 0.0004}$$

$$= \frac{3.842}{0.0016}$$

$$= 2401$$

Therefore, the size of the required large sample is 2401.

4.1.7 Estimating the Difference Between Two Means for Two Samples

Q34. Discuss in detail about estimating the difference between two means for two samples.

Answer :

Let $\bar{x}_1$ be the mean of a random sample of size n_1 drawn from a population with mean μ_1 and variance σ_1^2 and $\bar{x}_2$ be the mean of a random sample of size n_2 drawn from a population with mean μ_2 and variance σ_2^2. The statistic $\overline{X_1} - \overline{X_2}$ is the point estimator of the difference between two means μ_1 and μ_2. The difference of the sample means $\bar{x}_1 - \bar{x}_2$ is computed by finding a point estimate of $\mu_1 - \mu_2$.

The sampling distribution of $\overline{X_1} - \overline{X_2}$ is approximately normally distributed with mean,

$\mu_{\overline{X_1} - \overline{X_2}} = \mu_1 - \mu_2$

And standard deviation

$$\sigma_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Hence, $P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha$ ... (1)

Where,

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ with a probability of $1 - \alpha$

Now, substitute the value of 'z' in equation (1), we get,

$$P\left(-z_{\alpha/2} < \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) = 1 - \alpha \quad \ldots (2)$$

Thus, the above equation leads to the following $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$.

Confidence Interval for $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Known

When $\bar{x}_1$ and $\bar{x}_2$ are means of two independent samples of sizes n_1 and n_2 with variances σ_1^2 and σ_2^2 are known then a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Where, $z_{\alpha/2}$ represents z-value that leaves an area of $\frac{\alpha}{2}$ to the right side.

PROBLEMS

Q35. A study was conducted in which two types of engines, A and B were compared. Gas mileage, in miles per gallon was measured. Fifty experiments were conducted using engine type A and 75 experiments were done with engine type B. The gasoline used and other conditions were held constant. The average gas mileage was 36 miles per gallon for engine A and 42 miles for gallon for engine B. Find a 96% confidence interval on $\mu_B - \mu_A$ where μ_A and μ_B are population mean gas mileages for engines A and B, respectively. Assume that the population standard deviations are 6 and 8 for engines A and B, respectively.

Solution :

Given that,

Number of experiments conducted by engine A,

$$n_A = 50$$

Number of experiments conducted by engine B

$$n_B = 75$$

Population standard deviation for engine A,

$$\sigma_A = 6$$

Population standard deviation for engine B,

$$\sigma_B = 8$$

Mean of type A engine,

$$\overline{x_A} = 36$$

Mean of type B engine,

$$\overline{x_B} = 42$$

Confidence limit = 96%

$\Rightarrow (1-\alpha) = \frac{96}{100}$

$\Rightarrow 1 - \alpha = 0.96$

$\Rightarrow \alpha = 0.04$

$\frac{\alpha}{2} = 0.02$

$\therefore$ The value of $z_{\alpha/2}$ i.e., $z_{0.02}$ from Table Areas under the Normal curve is 2.05.

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

$$(\overline{x_B} - \overline{x_A}) - z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} < (\mu_B - \mu_A) < \overline{x_B} - \overline{x_A} + z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \quad \ldots (1)$$

The point estimate of $\mu_B - \mu_A$ is,

$\overline{x_B} - \overline{x_A}$

$\Rightarrow$ $42 - 36$

$\Rightarrow$ 6

Substituting the values of $n_A, n_B, z_{0.02}$ in equation (1), we get,

$$6 - (2.05)\sqrt{\frac{(6)^2}{50} + \frac{(8)^2}{75}} < (\mu_B - \mu_A) < 6 + 2.05\sqrt{\frac{(6)^2}{50} + \frac{(8)^2}{75}}$$

$$6 - (2.05)\sqrt{\frac{36}{50} + \frac{64}{75}} < (\mu_B - \mu_A) < 6 + 2.05\sqrt{\frac{36}{50} + \frac{64}{75}}$$

$$6 - (2.05)\sqrt{0.72 + 0.85} < (\mu_B - \mu_A) < 6 + 2.05\sqrt{0.72 + 0.85}$$

$$6 - (2.05)(1.25) < (\mu_B - \mu_A) < 6 + (2.05)(1.25)$$

$$6 - 2.56 < (\mu_B - \mu_A) < 6 + 2.56$$

$$3.44 < \mu_B - \mu_A < 8.56$$

Therefore, the 96% confidence interval on $\mu_B - \mu_A$ is, $3.44 < \mu_B - \mu_A < 8.56$.

Q36. The article "Macroinvertebrate Community Structure as an Indicator of Acid Mine Pollution", published in the Journal of Environmental Pollution, reports on an investigation undertaken in Cane Creek, Alabama, to determine the relationship between selected physiochemical parameters and different measures of macro invertebrate community structure. One facet of the investigation was an evaluation of the effectiveness of a numerical species diversity index to indicate aquatic degradation due to acid mine drainage. Conceptually, a high index of macroinvertabrate species diversity should indicate an unstressed aquatic system, while a low diversity index should indicate a stressed aquatic system. Two independent sampling stations were chosen for this study, one located downstream from the acid mine discharge point and the other located upstream. For 12 monthly samples collected at the downstream station, the species diversity index had a mean value $\bar{x}_1 = 3.11$ and a standard deviation $s_1 = 0.771$, while 10 monthly samples collected at the upstream station had a mean index value $\bar{x}_2 = 2.04$ and a standard deviation $s_2 = 0.448$. Find a 9% confidence interval for the difference between the population means for the two locations, assuming that the populations are approximately normally distributed with equal variances.

Solution :

Given that,

$n_1 = 12$

$n_2 = 10$

$\overline{x_1} = 3.11$

$\overline{x_2} = 2.04$

$s_1 = 0.771$

$s_2 = 0.448$

Let μ_1 and μ_2 be the population means for the species diversity indices at the downstream and upstream stations respectively.

A $100(1 - \sigma)\%$ confidence interval for $\mu_1 - \mu_2$ is given by,

$$(\overline{x_1} - \overline{x_2}) - t_{\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < \overline{x_1} - \overline{x_2} + t_{\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \ldots (1)$$

Where, $\overline{x_1}$, $\overline{x_2}$ are the means of two independent random samples of sizes n_1 and n_2 respectively from approximately normal populations with unknown but equal variances.

The point estimate of $\mu_1 - \mu_2$ is,

$\overline{x_1} - \overline{x_2}$

$\Rightarrow$ $3.11 - 2.04$

$\Rightarrow$ 1.07

The pooled estimate S_p^2 of the common variance σ^2 is given by,

$$S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$$= \frac{(12-1)(0.771)^2 + (10-1)(0.448)^2}{12+10-2}$$

$$= \frac{11(0.5944) + 9(0.2007)}{20}$$

$$= \frac{6.5384 + 1.8063}{20}$$

$$= \frac{8.3447}{20}$$

$$S_p^2 = 0.4172$$

$$S_p^2 = \sqrt{0.4172}$$

$$\boxed{S_p = 0.646}$$

Confidence limit = 90%

$\Rightarrow$ $(1-\alpha)100 = 90$

$\Rightarrow$ $(1-\alpha) = \frac{90}{100}$

$\Rightarrow$ $(1-\alpha) = 0.9$

$\Rightarrow$ $\alpha = 0.1$

$\frac{\alpha}{2} = 0.05$

The critical value of $t_{0.05}$ for t-distribution at 20 ($v = n_1 + n_2 - 2$) degrees of freedom is 1.725.

On substituting the values of $t_{0.05}$, $\overline{x_1} - \overline{x_2}$, S_p, n_1 and n_2 in equation (1), we get,

$$1.07 - (1.725)(0.646)\sqrt{\frac{1}{12}+\frac{1}{10}} < \mu_1 - \mu_2 < 1.07 + (1.725)(0.646)\sqrt{\frac{1}{12}+\frac{1}{10}}$$

$$1.07 - (1.725)(0.646)\sqrt{0.0833+0.1} < \mu_1 - \mu_2 < 1.07 + (1.725)(0.646)\sqrt{0.0883+0.1}$$

$$1.07 - 0.4770 < 1.07 + 0.4770$$

Which reduces to $0.593 < \mu_1 - \mu_2 < 1.547$

Therefore, the 90% confidence interval for the difference between the population mean for the two locations is,

$$0.593 < \mu_1 - \mu_2 < 1.547.$$

Q37. A study was conducted by the Department of Zoology at the Virginia Tech to estimate the difference in the amounts of the chemical orthophosphorus measured at two different stations on the James River. Orthophosphorus was measured in milligrams per liter. Fifteen samples were collected from station 1 and 12 samples were obtained from station 2. The 15 samples from station 1 had an average orthophosphorus content of 3.84 milligrams per liter and a standard deviation of 3.07 milligrams per liter, while the 12 samples from station 2 had an average content of 1.49 milligrams per liter and a standard deviation of 0.80 milligram per liter. Find a 95% confidence interval for the difference in the true average orthophosphorus contents at these two stations, assuming that the observations came from normal populations with different variances.

Solution :

Given that,

Number of samples collected from station 1, $n_1 = 15$

Number of samples collected from station 2, $n_2 = 12$

Mean of orthophosphorus content from station 1 samples, $\overline{x_1} = 3.84$,

Mean of orthophosphorus content from station 2 samples, $\overline{x_2} = 1.49$

Standard deviation of samples at station 1, $s_1 = 3.07$

Standard deviation of samples at station 2, $s_2 = 0.8$

The point estimate of $\mu_1 - \mu_2$ is,

$$\overline{x_1} - \overline{x_2}$$

$\Rightarrow$ $3.84 - 1.49$

$\Rightarrow$ 2.35

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

$$(x_1 - x_2) - t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < \overline{x_1} - \overline{x_2} + t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \ldots (1)$$

The 'v' degrees of freedom is obtained by,

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\frac{\frac{s_1^2}{n_1}}{(n_1 - 1)}\right] + \left[\frac{\frac{s_2^2}{n_2}}{(n_2 - 1)}\right]}$$

$$= \frac{\left(\frac{(3.07)^2}{15} + \frac{(0.8)^2}{12}\right)^2}{\left[\frac{\left(\frac{(3.07)^2}{15}\right)^2}{(15-1)}\right] + \left[\frac{\left(\frac{(0.8)^2}{12}\right)^2}{(12-1)}\right]}$$

$$= \frac{\left(\frac{9.4249}{15} + \frac{0.64}{12}\right)^2}{\frac{\left(\frac{9.4249}{15}\right)^2}{14} + \frac{\left(\frac{0.64}{12}\right)^2}{11}}$$

$$= \frac{(0.6283 + 0.0533)^2}{\frac{(0.6283)^2}{14} + \frac{(0.0533)^2}{11}} = \frac{0.4646}{0.02819 + 2.58 \times 10^{-4}}$$

$$= \frac{0.4646}{0.02844}$$

$$= 16.33$$

$$\cong 16$$

Confidence limit = 95%

$\Rightarrow$ $(1-\alpha)100 = 95$

$\Rightarrow$ $1 - \alpha = \frac{95}{100}$

$\Rightarrow$ $1 - \alpha = 0.95$

$\Rightarrow$ $\alpha = 0.05$

$\frac{\alpha}{2} = 0.025$

The critical value of $t_{0.025}$ for t-distribution at $v = 16$ degrees of freedom is 2.120.

Therefore, 95% confidence interval for $\mu_1 - \mu_2$ is obtained by substituting the values of $\overline{x_1} - \overline{x_2}$, s_1, s_2, n_1, n_2 and $t_{0.025}$ in equation (1),

$\Rightarrow$ $2.35 - 2.120 \sqrt{\frac{3.07^2}{15} + \frac{0.8^2}{12}} < \mu_1 - \mu_2 < 2.35 + 2.120 \sqrt{\frac{3.07^2}{15} + \frac{0.8^2}{12}}$

$\Rightarrow$ $2.35 - 2.120 \sqrt{0.68166} < \mu_1 - \mu_2 < 2.35 + 2.120 \sqrt{0.68166}$

$\Rightarrow$ $2.35 - 2.120\,(0.8256) < \mu_1 - \mu_2 < 2.35 + 2.120\,(0.8256)$

$\Rightarrow$ $2.35 - 1.75 < \mu_1 - \mu_2 < 2.35 + 1.75$

Which reduces to $0.6 < \mu_1 - \mu_2 < 4.1$

Therefore, we are 95% confident that the above range covers the difference of the true average orthophosphorus contents for the two stations.

4.1.8 Estimating the Difference Between Two Proportions for Two Samples

Q38. Discuss in detail about estimating the difference between two proportions for two samples.

Answer :

Consider that there are two distinct populations. Independent samples of sizes n_1 and n_2 are selected at random from two populations and proportion of successes p_1 and p_2 for two samples is computed. The statistic $\hat{p}_1 - \hat{p}_2$ is the point estimator of the difference between two proportions $p_1 - p_2$.

The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is used for establishing a confidence interval for $p_1 - p_2$.

The construction of confidence intervals for p_1 and p_2, for n_1 and n_2 is sufficiently large, the difference of two means $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed with mean $(\mu_{\hat{p}_1 - \hat{p}_2}) = p_1 - p_2$ and

Variance $\left(\sigma^2_{\hat{p}_1 - \hat{p}_2}\right) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

Therefore, we can conclude that,

$$P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha \quad \ldots (1)$$

Where,

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

Here, $z_{\alpha/2}$ represents the value of Z that leaves an area of $\frac{\alpha}{2}$ to the right side.

On substituting the value of Z in equation (1), we get,

$$P\left[-z_{\alpha/2} < \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} < z_{\alpha/2}\right] = 1 - \alpha$$

Large Sample Confidence Interval for $p_1 - p_2$

When $\hat{p}_1$ is the proportion of success of random sample of size n_1 and $\hat{p}_2$ is the proportion of success of random sample of size n_2, $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$ then an approximate $100(1 - \alpha)\%$ confidence interval for $p_1 - p_2$ is,

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Where, $z_{\alpha/2}$ represents 'z' value that leaves an area of $\frac{\alpha}{2}$ to the right side.

PROBLEM

Q39. A certain change in a process for manufacturing component parts is being considered samples are taken under both the existing and the new process so as to determine if the new process results in an improvement. If 75 of 1500 items from the existing process are found to be defective and 80 of 2000 items from the new process are found to be defective, find a 90% confidence interval for the true difference in the proportion of defectives between the existing and the new process.

Solution :

Given that,

$$n_1 = 1500, n_2 = 2000, X_1 = 75, X_2 = 80$$

Let p_1 be the true proportion of defectives for the existing process

$$\hat{p}_1 = \frac{75}{1500} \qquad \left[\because p = \frac{X}{n}\right]$$

$$= 0.05$$

$$\hat{q}_1 = 1 - \hat{p}_1$$

$$= 1 - 0.05$$

$$= 0.95$$

Let p_2 be the true proportion of defectives for the new process.

$$\hat{p}_2 = \frac{80}{2000}$$
$$= 0.04$$
$$\hat{q}_2 = 1 - \hat{p}_2$$
$$= 1 - 0.04$$
$$= 0.96$$

The point estimate of $p_1 - p_2$ is given by,

$$\hat{p}_1 - \hat{p}_2 = 0.05 - 0.04$$
$$= 0.01$$

Confidence limit = 90%

$$\Rightarrow (1 - \alpha)100 = 90$$
$$\Rightarrow 1 - \alpha = \frac{90}{100}$$
$$\Rightarrow 1 - \alpha = 0.9$$
$$\Rightarrow \alpha = 0.1$$
$$\frac{\alpha}{2} = 0.05$$

The value of $z_{0.05}$ using Table Areas under the Normal curve is 1.645.

We have, $100(1 - \alpha)\%$ confidence interval for the difference of two binomial parameters, $p_1 - p_2$ is,

$$\hat{p}_1 - \hat{p}_2 - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \quad \ldots (1)$$

Therefore, 90% confidence interval is obtained by substituting the values of $z_{0.05}$, p_1, q_1, n_1, n_2 and $\hat{p}_1 - \hat{p}_2$ in equation (1), we get,

$$\Rightarrow 0.01 - 1.645 \sqrt{\frac{(0.05)(0.95)}{1500} + \frac{(0.04)(0.96)}{2000}} < p_1 - p_2 < 0.01 + 1.645 \sqrt{\frac{(0.05)(0.95)}{1500} + \frac{(0.04)(0.90)}{2000}}$$
$$\Rightarrow 0.01 - 1.645\,(7.132 \times 10^{-3}) < p_1 - p_2 < 0.01 + 1.645\,(7.132 \times 10^{-3})$$
$$\Rightarrow 0.01 - 0.0117 < p_1 - p_2 < 0.01 + 0.0117$$
$$\Rightarrow -0.0017 < p_1 - p_2 < 0.0217$$

Therefore, 90% of the confidence interval is, $-0.0017 < p_1 - p_2 < 0.0217$.

4.1.9 Maximum Likelihood Estimation

Q40. Explain the method of maximum likelihood estimation.

Answer :

Method of Maximum Likelihood Estimation (MLE)

The method of maximum likelihood estimation is used to obtain the best estimate of a population parameter of the density function pertaining to some distribution.

The general procedure of finding the maximum likelihood estimate is as follows,

1. Initially, determine the density function of the given distribution and its population parameter.
2. Denote the density function as $g(t, \lambda)$ where λ is the population parameter.
3. Assume that there are 'n' independent observations ranging from t_1 to t_n.
4. The joint density function for all observations is then calculated. Denote it by J. The value of J is calculated as,

   $J = g(t_1, \lambda), g(t_2, \lambda), \ldots, g(t_n, \lambda)$

   Here, J is said to be the likelihood.
5. Obtain the maximum likelihood by calculating the derivative of J with respect to the population parameter λ.

6. Set $\frac{dJ}{d\lambda} = 0$

7. Logarithms can be used for convenience.

8. Find solution for λ in terms of t_i. This solution is the maximum likelihood estimator of λ.

Thus, we find the maximum likelihood estimate of a given distribution.

Now,

Consider the example of estimating maximum likelihood for exponential distribution.

Let the density function of exponential distribution be $g(t, \lambda)$.

$\Rightarrow g(t, \lambda) = \lambda e^{-\lambda t}$

Where,

λ = Population parameter of the distribution.

Let, $t_1, t_2, ..., t_n$ be the 'n' independent observations.

Then, the joint density function is,

$$J = g((t_1, t_2, ..., t_n), \lambda)$$

$$\Rightarrow \quad J = \prod_{i=1}^{n} g(t_i, \lambda)$$

$$\Rightarrow \quad J = \prod_{i=1}^{n} \lambda e^{-\lambda t_i}$$

$$\Rightarrow \quad J = \prod_{i=1}^{n} \lambda . \prod_{i=1}^{n} e^{-\lambda t_i}$$

$$\Rightarrow \quad J = \lambda^n . e^{-\lambda \sum_{i=1}^{n} t_i} \qquad \text{... (1)}$$

Applying log on both sides of equation (1), we get,

$$\log J = \log (\lambda^n . e^{-\lambda \sum_{i=1}^{n} t_i})$$

$$\Rightarrow \quad \log J = \log \lambda^n + \log e^{-\lambda \sum_{i=1}^{n} t_i}$$

$$\Rightarrow \quad \log J = n \log \lambda + \left(\log_e e^{-\lambda \sum_{i=1}^{n} t_i} \right)$$

$$\Rightarrow \quad \log J = n \log \lambda - \left(\lambda \sum_{i=1}^{n} t_i \right) \log_e e$$

$$\Rightarrow \quad \log J = n \log \lambda - \left(\lambda \sum_{i=1}^{n} t_i \right) \quad [\because \log_e e = 1] \qquad \text{... (2)}$$

Differentiating both sides of equation (2), with respect to λ, we get,

$$\frac{d}{d\lambda}(\log J) = \frac{d}{d\lambda}\left(n \log \lambda - \left(\lambda \sum_{i=1}^{n} t_i \right) \right)$$

$$\Rightarrow \quad \frac{1}{J} \cdot \frac{dJ}{d\lambda} = \frac{d}{d\lambda}(n \log \lambda) - \frac{d}{d\lambda}(\lambda) \cdot \sum_{i=1}^{n} t_i$$

$$\Rightarrow \quad \frac{1}{J} \cdot \frac{dJ}{d\lambda} = n . \frac{1}{\lambda} - 1 . \sum_{i=1}^{n} t_i$$

$$\Rightarrow \quad \frac{1}{J} \cdot \frac{dJ}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i$$

$$\Rightarrow \quad \frac{dJ}{d\lambda} = J \left[\frac{n}{\lambda} - \sum_{i=1}^{n} t_i \right]$$

Set $\frac{dJ}{d\lambda} = 0$, solve for λ,

$$\Rightarrow \quad J\left[\frac{n}{\lambda} - \sum_{i=1}^{n} t_i\right] = 0$$

$$\Rightarrow \quad \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0$$

$$\Rightarrow \quad \frac{n}{\lambda} = \sum_{i=1}^{n} t_i$$

$\therefore$ The maximum likelihood estimate of γ i.e., $\gamma = \frac{n}{\sum_{i=1}^{n} t_i}$

PROBLEMS

Q41. Consider a poisson distribution with probability mass function.

$$f\left(\frac{x}{\mu}\right) = \frac{e^{-\mu}\mu^{x}}{x!}, \; x = 0, 1, 2, \ldots.$$

Suppose that a random sample $x_1, x_2, \ldots, x_n$ is taken from the distribution, what is the maximum likelihood estimate of μ?

Solution :

The likelihood function is given by,

$$L(x_1, x_2, \ldots x_n; \mu) = \prod_{i=1}^{n} f\left(\frac{x_i}{\mu}\right)$$

$$= \frac{e^{-n\mu}\mu \sum_{i=1}^{n} x_i}{\prod_{i=1}^{n} x_i!}$$

Applying logarithms on both sides, we get

$$\ln L(x_1, x_2, x_3, \ldots, x_n, \mu) = \ln\left(\frac{e^{-n\mu}\sum_{\mu_i=1}^{n} x_i}{\prod_{i=1}^{n} x_i!}\right)$$

$$\ln L(x_1, x_2, x_3, \ldots, x_n, \mu) = \ln\left(e^{-n\mu}\right) + \ln\left(\mu \sum_{i=1}^{n} x_i\right) - \ln \prod_{i=1}^{n} x_i!$$

$$\ln L(x_1, x_2, x_3, \ldots, x_n, \mu) = \ln\left(e^{-n\mu}\right) + \ln\left(\mu \sum_{i=1}^{n} x_i\right) - \ln \prod_{i=1}^{n} x_i!$$

$$\ln L(x_1, x_2, x_3, \ldots, x_n, \mu) = -n\mu + \sum_{i=1}^{n} \ln \mu - \ln \prod_{i=1}^{n} x_i!$$

Applying derivative w.r.t 'μ' on both sides, we get,

$$\frac{\partial}{\partial \mu}\left(\ln L(x_1, x_2, x_3, \ldots, x_n; \mu)\right) = \frac{\partial}{\partial \mu}(-n\mu) + \frac{\partial}{\partial \mu}\left(\sum_{i=1}^{n} \ln \mu\right) - \frac{\partial}{\partial \mu}\left(\ln \prod_{i=1}^{n} x_i!\right)$$

$$\frac{\partial}{\partial \mu}\left(\ln L(x_1, x_2, x_3, \ldots, x_n; \mu)\right) = -n + \sum_{i=1}^{n} \frac{x_i}{\mu} - 0 \qquad \ldots (1)$$

In the above equation, the derivative has to set to zero inorder to solve for maximum likelihood estimator, μ.

$$-n+\sum_{i=1}^{n}\frac{x_i}{\mu}=0$$

$$\sum_{i=1}^{n}\frac{x_i}{\mu}=n$$

$$\mu=\sum_{i=1}^{n}\frac{x_i}{n}=\bar{x}$$

Applying second derivative on log likelihood function becomes negative, which means that solution attained is a maximum. Indeed, the sample average looks like a reasonable estimator as 'μ' is the mean of poison distribution.

Q42. Consider a random sample $x_1, x_2, \ldots\ldots x_n$ from a normal distribution $N(\mu, \sigma)$. Find the maximum likelihood estimators for μ and σ^2.

Solution : Model Paper-I, Q9(b)

The likelihood function for the normal distribution is given by,

$$L(x_1, x_2, \ldots. x_n; \mu, \sigma^2)=\frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}}\exp\left[\frac{-1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] \quad \ldots (1)$$

Applying logarithms on both sides, we get,

$$\ln L\left(x_1, x_2, x_3, \ldots\ldots, x_n; \mu, \sigma^2\right)=\ln\left(\frac{1}{2\pi^{n/2}}\right)+\ln\left(\frac{1}{(\sigma^2)^{n/2}}\right)+\log\left(e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)$$

$$\ln L\left(x_1, x_2, x_3, \ldots\ldots, x_n; \mu, \sigma^2\right)=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\ln e$$

Applying derivative with respect to 'μ', we get,

$$\frac{\partial}{\partial\mu}\ln\left(x_1, x_2, x_3, \ldots\ldots x_n; \mu, \sigma^2\right)=\frac{\partial}{\partial\mu}\left(\frac{-n}{2}\ln 2\pi\right)+\frac{\partial}{\partial\mu}\left(\frac{-n}{2}\ln\sigma^2\right)+\frac{\partial}{\partial\mu}\left(\frac{-1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)$$

$$\frac{\partial}{\partial\mu}\ln L=0+0-\frac{1}{2}\sum_{i=1}^{n}2\left(\frac{x_i-\mu}{\sigma}\right)\frac{-1}{\sigma}$$

$$\frac{\partial}{\partial\mu}\ln L=\sum_{i=1}^{n}\frac{(x_i-\mu)}{\sigma^2}$$

Now, applying derivative with respect to σ^2, we get,

$$\frac{\partial}{\partial\sigma^2}\ln L\left(x_1, x_2, x_3, \ldots\ldots x_n; \mu, \sigma^2\right)=\frac{\partial}{\partial\sigma^2}\left(\frac{-n}{2}\ln 2\pi\right)+\frac{\partial}{\partial\sigma^2}\left(\frac{-n}{2}\ln\sigma^2\right)+\frac{\partial}{\partial\sigma^2}\left(\frac{-1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)$$

$$\frac{\partial}{\partial\sigma^2}\ln L=0-\frac{n}{2}\frac{1}{\sigma^2}-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\left(\frac{-1}{(\sigma^2)^2}\right)$$

$$\frac{\partial}{\partial\sigma^2}\ln L=-\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2(\sigma^2)}\sum_{i=1}^{n}(x_i-\mu^2)$$

Now, setting the derivative $\frac{\partial}{\partial\mu}\ln L$ equal to zero, we get,

$$\frac{\partial}{\partial\mu}\ln L=0$$

$$\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2}=0$$

$$\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\mu=0$$

$$\sum_{i=1}^{n} x_i - n\mu = 0$$

$$\sum_{i=1}^{n} x_i = n\mu$$

Therefore, the maximum likelihood estimator of μ is given by,

$$\boxed{\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}}$$

Now, setting up the derivative $\frac{\partial}{\partial \sigma^2} \ln L$ equal to zero, we get,

$$\frac{\partial}{\partial \sigma^2} \ln L = 0$$

$$\frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

$$\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{n}{2(\sigma^2)}$$

$$\sum_{i=1}^{n}(x_i - \mu)^2 = n\sigma^2$$

$$\sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Therefore, the maximum likelihood estimator of σ^2 is,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\boxed{\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad [\because \mu = \bar{x}]$$

Q43. It is known that a sample consisting of the values 12, 112, 13.5, 12.3, 13.8 and 11.9 comes from a population with density function $f(x, \theta) = \begin{cases} \frac{\theta}{x^{\theta+1}} ; & x > 1 \\ 0 ; & \text{elsewhere} \end{cases}$ **where** $\theta > 0$.

Find the maximum likelihood estimate of θ.

Solution :

The likelihood function of 'n' observations from the population is given by,

$$L(x_1, x_2, \ldots\ldots, x_{10}; \theta) = \prod_{i=1}^{n} \frac{\theta}{x_i^{\theta+1}}$$

$$= \frac{\theta^n}{\left(\prod_{i=1}^{n} x_i\right)^{\theta+1}}$$

Applying logarithms on both sides, we get

$$\ln L(x_1, x_2, \ldots\ldots x_{10}; \theta) = \ln\left[\frac{\theta^n}{\left(\prod_{i=1}^{n} x_i\right)^{\theta+1}}\right]$$

$$= \ln(\theta^n) - \ln\left(\prod_{i=1}^{n} x_i\right)^{\theta+1}$$

$$= n \ln \theta - (\theta+1)\sum_{i=1}^{n} \ln x_i$$

Applying derivative with respect to 'θ', we get,

$$\frac{\partial}{\partial\theta}\ln L(x_1, x_2, x_3, \ldots\ldots, x_{10};\theta) = \frac{\partial}{\partial\theta}(n\ln\theta) - \frac{\partial}{\partial\theta}\left((\theta+1)\sum_{i=1}^{n}\ln x_i\right)$$

$$= \frac{n}{\theta} - \sum_{i=1}^{n}\ln x_i \frac{\partial}{\partial\theta}(\theta+1)$$

$$= \frac{n}{\theta} - \sum_{i=1}^{n}\ln x_i(1+\theta)$$

$$\frac{\partial}{\partial\theta}\ln L(x_1, x_2, x_3, \ldots\ldots, x_{10};\theta) = \frac{n}{\theta} - \sum_{i=1}^{n}\ln x_i$$

Now, setting the derivative $\frac{\partial}{\partial\theta}\ln L$ to zero, we get,

$$\frac{n}{\theta} - \sum_{i=1}^{n}\ln x_i = 0$$

$$\sum_{i=1}^{n}\ln x_i = \frac{n}{\theta}$$

$$\boxed{\hat{\theta} = \frac{n}{\sum_{i=1}^{n}\ln x_i}}$$

$$\hat{\theta} = \frac{6}{\ln(12)+\ln(11.2)+\ln(13.5)+\ln(12.3)+\ln(13.8)+\ln(11.9)}$$

$$= \frac{6}{2.48+2.41+2.60+2.5+2.62+2.47}$$

$$= \frac{6}{15.08}$$

$$\boxed{\hat{\theta} = 0.3978}$$

Applying second derivative on log-likelihood function yields a negative value. Therefore, the solution obtained is maximum at value θ.

Q44. Suppose 10 rats are used in a biochemical study where they are injected with cancer cell and then given a cancer drug that is designed to increase their survival rate. The survival times, in months are 14, 17, 27, 18, 12, 8, 22, 13, 19 and 12. Assume that the exponential distribution applies. Give a maximum likelihood estimate of the mean survival time.

Solution :

The probability density function for the exponential random variable 'X' is given by,

$$f(x,\beta) = \begin{cases}\frac{1}{\beta}e^{-x/\beta}, & x>0\\ 0 & \text{, elsewhere}\end{cases}$$

Given that, $n = 10$,

The log-likelihood function for the data is,

$$GL(x_1, x_2, x_3, \ldots x_{10};\beta) = -10ln\beta - \frac{1}{\beta}\sum_{i=1}^{10}x_i$$

Applying derivative with respect to β, we get,

$$\frac{\partial}{\partial\beta}GL(x_1, x_2, \ldots x_{10};\beta) = \frac{\partial}{\partial\beta}(-10\ln\beta) - \frac{\partial}{\partial\beta}\left(\frac{1}{\beta^2}\sum_{i=1}^{10}x_i\right)$$

$$= \frac{-10}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{10}x_i = 0$$

Setting up the derivative $\frac{\partial}{\partial\beta} lnL$ equal to zero, we get,

$$\frac{-10}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{10} x_i = 0$$

$$\frac{1}{\beta^2}\sum_{i=1}^{10} x_i = 0$$

$$\beta = \frac{1}{10}\sum_{i=1}^{10} x_i = \bar{x}$$

$$= \frac{1}{10}(x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10})$$

$$= \frac{1}{10}(14 + 17 + 27 + 18 + 12 + 8 + 22 + 13 + 9 + 12)$$

$$= \frac{162}{10}$$

$$= 16.2$$

Applying second derivative log-likelihood function 'L' becomes a negative value. Therefore, the maximum likelihood estimator and the population mean is the sample average ($\bar{x}$).

4.2 STATISTICAL HYPOTHESES: GENERAL CONCEPTS

Q45. What is a statistical hypothesis? Write about null hypothesis and alternative hypothesis.

Answer :

Statistical Hypothesis

Statistical hypothesis are statements about the probability distributions of the populations. There are two types of hypothesis. They are,

(i) Null hypothesis

(ii) Alternative hypothesis.

(i) Null Hypothesis

Null hypothesis is the hypothesis which is tested for possible under the assumption that it is true. Usually, it is denoted by H_0.

Example

To decide whether a given computer performs well, we form the hypothesis that the computer is doing well, i.e., $p = 100\%$. Where, p is probability of success.

(ii) Alternative Hypothesis

Any hypothesis that differs from a given null hypothesis is called an 'alternative hypothesis'. The alternative hypothesis is denoted by symbol H_1.

Example

If null hypothesis is $P = 100\%$, possible alternative hypothesis are,

$P \neq 100\%$ or $P > 100\%$ or $P < 100\%$.

Therefore, in order to test the null hypothesis that the population has a specified mean μ_0 i.e., $H_0 = \mu = \mu_0$, then the alternative hypothesis would be,

(i) $H_1 : \mu \neq \mu_0$ (Two tailed alternative)

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

(iii) $H_1 : \mu < \mu_0$ (Left tailed).

4.2.1 Testing a Statistical Hypothesis

Q46. What is a hypothesis? Explain the procedure for testing a statistical hypothesis.

Model Paper-III, Q9(a)

Answer :

Hypothesis

Statistical hypothesis is an assumption about the parameters of the population and sometimes it also concerns the type and nature of the distribution.

Example

(i) The average height of soldiers in the army is 165 cm.

(ii) A given drug cures 80% of the patients taking it.

(iii) A given machine has an effective life of 20 years.

All these hypotheses may be verified on the basis of certain sample tests. Procedures or tests which enable us to decide whether to accept or reject the hypothesis are called *tests of hypothesis* or *tests of significance*.

Use of Hypothesis Testing

The test of hypothesis discloses the fact whether the difference between the computed statistic and hypothesical parameter is significant or otherwise.

Procedure of Hypothesis Testing

Hypothesis Testing

To test a hypothesis means to tell on the basis of the data, whether or not the hypothesis seems to be valid. Procedure for hypothesis testing refers to all those steps that we undertake for making a choice between two actions i.e., rejection and acceptance of a null hypothesis.

Steps in Hypothesis Testing

1. Making a Formal Statement

This step consists of making a formal statement of the null hypothesis (H_o) and also of alternatively hypothesis (H_a). This means that the hypothesis should be clearly stated, considering the nature of the research problem.

Example

In a population of 30 industrial units of same size, labour turnover is analyzed. The past records show that mean of turnover is 320 employees. A sample of 5 of these units is taken at random which gave a mean of 300. The labour minister wants to know if there is a significant difference between the sample and population.

Null hypothesis $H_o : \mu = 320$

Alternative hypothesis $H_a : \mu \neq 320$

In the above example, if the union claims that the turnover is above 320, and if the claim has to tested then,

Null hypothesis $H_o : \mu = 320$

Alternative hypothesis $H_a : \mu > 320$

We should decide whether to use a two tailed test or one-tailed test.

2. Selecting a Significance Level

The hypothesis are tested on a predetermined level of significance and as such the same should be specified. Generally, in practice, either 5% level or 1% is adopted for the purpose.

The factors that affect the level of significance are,

- The size of the samples
- Magnitude of the difference between sample means.
- Variability of measurements within samples.
- Whether the hypothesis is directional or non-directional.

The level of significance must be adequate in the context of the purpose and nature of inquiry.

3. Deciding the Distribution to Use

The next step is to determine the appropriate sampling distribution. The choice generally remains between normal distribution and the t-distribution.

4. Selecting a Random Sample and Compu-ting an Appropriate Value

Another step is to select a random sample (s) and compute an appropriate value from the sample data concerning the test statistic utilizing the relevant distribution.

5. Calculation of Probability

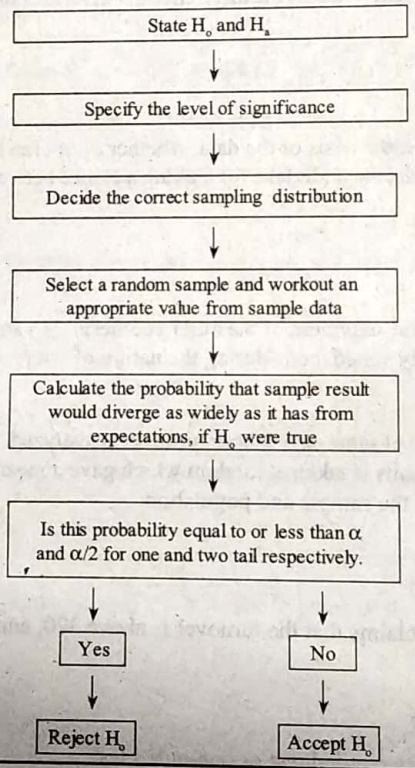
Calculate the probability that the sample result would diverge as widely as it has from expectations, if null hypothesis were is fact true.

6. Comparing the Probability

Now, compare the probability that has been calculated with the specified value for α, the level of significance.

If the calculated probability is less than or equal to α value in case of one tailed test and $\frac{\alpha}{2}$ value in case of two tailed test, then reject H_o and accept H_a and vice versa.

Flow diagram for hypothesis testing.



Q47. Explain various concepts of hypothesis and its applications.

Answer :

1. Null Hypothesis

A definite statement about the populations parameter for applying the tests of significance is called the null hypothesis, which is usually a hypothesis of an difference.

For example, if we want to decide whether one procedure is better than another, we formulate the null hypothesis that there is no difference between the procedure. It is usually denoted by H_0. According to Prof. R.A. Fisher, "Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true".

Example : $H_0 \mu = \mu H_0$

2. Alternate Hypothesis

To be able to construct suitable criteria for testing statistical hypothesis, it is necessary that we also formulate alternative hypothesis. Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis and is usually denoted by the symbol H_1. For example if we want, to test the null hypothesis that the average height of the soldiers is 165 cm

Example : $H_a : \mu \neq \mu_{H_0}, \mu > \mu_{H_0}, \mu < \mu_{H_0}$

3. Errors in Sampling

The main objective in sampling theory is to draw valid inferences about the population parameters on the basis of the sample results. There is always a chance of making error. There are two possible types of errors in the test of hypothesis.

(i) Type I error

(ii) Type II error.

(i) Type I Error

If we reject the null hypothesis when it should be accepted, we say that a type I error has been made. The probability of committing a type I error is denoted by α i.e,

P[Reject H_0 when it is true] = α

(ii) Type II Error

If we accept the null hypothesis when it should be rejected, we say that a type II error has been made. The probability of committing a type II error is denoted by β i.e.,

P[Accept H_0 when it is wrong] = β

Actual	Decision	
	Accept H_0	Reject H_0
H_0 is true	Correct decision (No error) probability $= 1 - \alpha$	Wrong decision (Type I error) probability $= \alpha$
H_0 is wrong	Wrong decision (Type II error) probability $= \beta$	Correct decision (No error) probability $1 - \beta$

Table: Dichotomous

α error is also known as producer's risk because a good lot is getting rejected.

β error is also known as consumer's risk because a bad lot is getting accepted.

Tradeoff Between Type I and Type II Error

Type I error can be controlled by fixing it at a lower significance level (say 1% instead of 5%). But with a fixed sample size n, when we try to reduce type I error, the probability of committing type II error increases. Both types of errors cannot be reduced simultaneously. Hence here is a trade off between these two types of errors.

4. Test Statistic

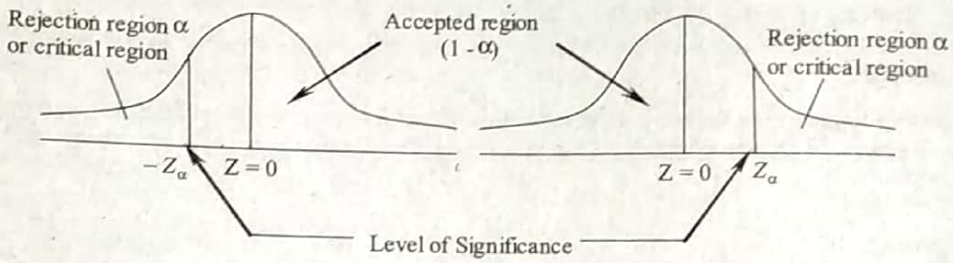
To test whether the null hypothesis setup should be accepted or rejected we use a test statistic which is based one appropriate probability distribution. For this purpose z-distribution under normal curve is used for large samples i.e, ($n \geq 30$) and t-distribution is used for small samples i..e, ($n < 30$).

5. Level of Significance (LOS)

In testing of a given hypothesis the maximum probability with which we would be willing to risk a type I error is called the level of significance of the test.

It is denoted by 'α' and in other words it is the maximum probability of making type I error. α is generally taken as 0.05 or 0.01 (i.e, 5% and 1%). $\alpha = 0.01$ is used for high precision and $\alpha = 0.05$ for moderate precision.

If we adopt 5% level of significance, it implies that there are 5 chance in 100, we are likely to reject a correct null hypothesis H_0. In other words, we are 95% confident that we have made a correct decision. The level of significance is also called the critical level.



6. Power of a Test

Power of a test is the probability of rejecting H_0 given that a specific alternative is true.

The power of a test can be computed as $1 - \beta$. Often different types of tests are compared by contrasting power properties.

7. Critical Region (C.R)

The area under the probability curve is divided into two regions.

(i) Region of rejection (significant region) where N.H. is rejected.

(ii) Region of acceptance (nonsignificant region) where N.H. is accepted.

Critical region is the region of rejection of N.H. The area of the critical region equals to the level of significance α.

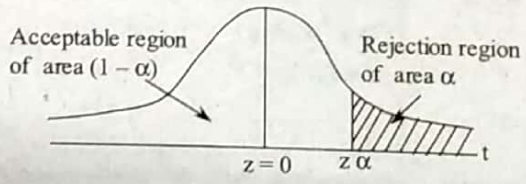
8. One Tailed and Two Tailed Tests

A test of any statistical hypothesis where the alternative hypothesis is expressed by the symbol (<) or the symbol (>) is called a cone tailed test since the entire critical region lies in one tail of the distribution of the test statistic.

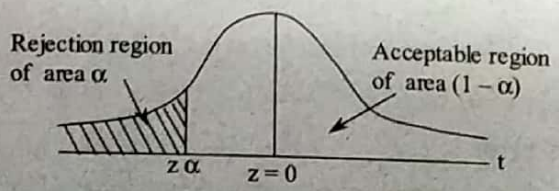
The critical region for all alternative hypothesis containing the symbol (>) lies entirely in the right tail of the distribution while the critical region for an alternative hypothesis containing the symbol (>) lies entirely in the left tail. The symbol indicates the direction where the critical region lies.

A test of any statistical hypothesis where the alternative is written with a symbol ($\neq$) is called a two tailed test, since region is split into two equal parts, one in each tail of the distribution of the test statistic.

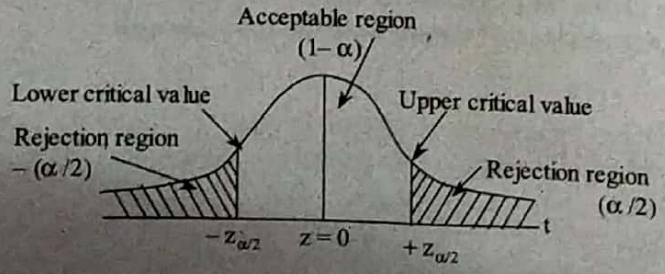
Right Tailed Test



Left Tailed Test



Two Tailed Test



The critical region lies on both sides of the right and left tails of the curve such that the critical region of area $\alpha/2$ lies on the right tail and critical region of area $\alpha/2$ lies on the left tail as shown in two tailed test figure above.

Application of One and Two Tailed Tests

A two tailed test rejects the null hypothesis if say the sample mean is significantly higher or lower than the hypothesized value of the mean of the population. Such a test is appropriate when the null hypothesis is some specified value and alternative hypothesis is a value not equal to the specified value of null hypothesis.

A one tailed test would be used when we are to test, say, whether the population mean is either lower than or higher than some hypothesized value. When we are interested in knowing whether, say, population mean is lower than the hypothesized mean, then it is a left tail test. But when we want to know whether, say, population mean is higher than the hypothesized mean, then it is a right tail test.

9. Critical Value of Significant Values

The value of test statistic which separates the area under the probability curve into critical (or rejection) region and noncritical (or acceptance) region. It depends upon the following,

(i) Level of significance (L.O.S)

(ii) Alternative hypothesis.

We know that $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

Here, Z is test statistic.

10. Decision Rule or Test of Hypothesis

Given a hypothesis H_0 and an alternate hypothesis H_a, we make a rule which is known as decision rule according to which we accept H_0 or reject H_0.

By comparing the critical value and the calculated value we reject or accept the null hypothesis at 5% or 1% level of significance as follows,

If Z is calculated value and z_α is critical value or table value then.

(a) Two Tailed Test

if $|Z| \geq z_\alpha$ at 5% or 1% L.O.S Reject H_0

if $|Z| < z_\alpha$ at 5% at 5% or 1% L.O.S Accept H_0.

(b) Right Tailed Test

if $|Z| \geq z_\alpha$ at 5% or 1% L.O.S Reject H_0

if $Z < z_\alpha$ at 5% or 1% L.O.S accept H_0.

(c) Left Tailed Test

if $|Z| \leq z_\alpha$ at 5% or 1% L.O.S Reject H_0

if $Z \geq z_\alpha$ at 5% or 1% L.O.S Accept H_0.

PROBLEMS

Q48. A real estate agent claims that 60% of all private residences being built today are 3-bedroom homes. To test this claim, a large sample of new residences is inspected; the proportion of these homes with 3 bedrooms is recorded and used as the test statistic. State the null and alternative hypotheses to be used in this test and determine the location of the critical region.

Solution :

The claim of real estate agent is rejected if test statistic is less than or greater than $P = 0.6$. Then, the hypothesis should be,

Null Hypothesis H_0: $P = 0.6$

Alternate Hypothesis H_1: $P \neq 0.6$

In two-tailed test, if an alternative hypothesis is represented by not equal to ($\neq$) symbol then critical region located equally is in both the tails of the distribution.

Q49. A manufacturer of a certain brand of rice cereal claims that the average saturated fat content does not exceed 1.5 grams per serving. State the null and alternative hypotheses to be used in testing this claim and determine where the critical region is located.

Solution :

The claim of manufacturer should not be accepted if $\mu > 1.5$ grams and should be accepted if $\mu \leq 1.5$ grams.

Null Hypothesis H_0: $\mu \leq 1.5$

Alternate Hypothesis H_1: $\mu > 1.5$

Acceptance of Null hypothesis do not eliminate the values that are less than 1.5 grams.

In one-tailed test, when an alternative hypothesis contains '>' symbol then critical region is located completely in the right tail of the distribution.

4.2.2 Tests Concerning a Single Mean

Q50. Explain the process of testing of hypothesis concerning single mean.

Answer :

To test whether the given sample of size n has been drawn from a population sample with mean μ, we set a null hypothesis H_0, i.e., there is no difference between sample mean $\bar{x}$ and population mean μ. Therefore, test statistic corresponding to $\bar{x}$ is given by,

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

Where,

σ is the standard deviation of the population.

In case, if we do not know S.D, then use the statistics, $Z = \frac{\bar{x} - \mu}{S / \sqrt{n}}$,

Where, S is the sample S.D.

- The 95% confidence limits for μ is, $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$
- The 99% confidence limits for μ is, $\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$
- The 98% confidence limits for μ is, $\bar{x} \pm 2.33 \frac{\sigma}{\sqrt{n}}$.

PROBLEMS

Q51. A random sample of 100 recorded deaths in a country showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is greater than 70 years? Use a 0.05 level of significance.

Solution :

1. **Null Hypothesis (H_0)**

   $\mu = 70$ years

2. **Alternative Hypothesis (H_1)**

   $\mu > 70$ years

3. **Level of Significance**

   $\alpha = 0.05$

4. **Critical Region**

   Reject if $Z > 1.645$

5. Test Statistic

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

Where,

$\bar{x}$ = Mean of sample

= 71.8 years

μ = Mean of population

= 70 years

σ = Standard deviation

= 8.9 years

n = Sample size = 100

$$Z = \frac{71.8 - 70}{\frac{8.9}{\sqrt{100}}}$$

$$Z = \frac{1.8}{\frac{8.9}{10}}$$

$$Z = \frac{1.8}{0.89}$$

$$Z = 2.022$$

6. Decision

Since the calculated value of Z is greater than the tabulated value i.e., $2.022 > 1.645$

Therefore the null hypothesis is rejected. Hence, the mean life span today is greater than 70 years.

Q52. A manufacturer claims that the mean breaking strength of a cable is 8 kgs, with a standard deviation of 0.5 kg. Test the hypothesis that the mean $\mu = 8$ kgs against $\mu \neq 8$ if a random sample of 50 cables is tested and found to have a mean breaking strength of 7.8 kgs. Use a 0.01 level of significance.

Model Paper-II, Q9(b)

Solution :

Given that,

1. Null Hypothesis (H_0)

$\mu = 8$ kilograms

2. Alternative Hypothesis (H_1)

$\mu \neq 8$ kilograms

3. Level of Significance (α)

$\alpha = 0.01$

4. Computation

Test statistics,

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

$\bar{x}$ = Mean of sample = 7.8 kilograms

μ = Mean of the company = 8 kilograms

σ = Standard deviation = 0.5 kilograms

n = Sample size = 50

$$\therefore \quad Z = \frac{0.5}{\sqrt{50}}$$

$$= \frac{0.5}{7.071}$$

$$= \frac{-0.2}{0.0707}$$

$$= -2.828$$

The calculated value of $|Z| = -2.828$

5. Decision

$\alpha = 0.01$

$$|Z_{\alpha/2}| = 2.33$$

$$|Z|_{cal} > Z_{\alpha/2}$$

$$2.82 > +2.33$$

Since, the calculated value of Z i.e., 2.82 is greater than the tabulated value of Z i.e., 2.33. Hence, H_0 is rejected.

Q53. It is claimed that a vacuum cleaner uses an average of 46 kilowatt hours per year. If a random sample of 12 homes indicates that the vacuum cleaners use an average of 42 kilowatt hours per year with a standard deviation of 11.9 kilowatt hours, does this suggest at 0.05 level of significance that vacuum cleaners use on average less than 46 kilowatt hours annually (assume that the population of kilowatt hours is normal).

SOlution :

1. Null Hypothesis (H_0)

$\mu = 46$ kilowatt hours

2. Alternative Hypothesis (H_1)

$\mu < 46$ kilowatt hours

3. Level of Significance

$\alpha = 0.05$

4. Computations

$\bar{x} = 42$ kilowatt hours

$s = 11.9$ kilowatt hours

$n = 12$

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{42 - 46}{\frac{11.9}{\sqrt{12}}}$$

$$= \frac{-4}{\frac{11.9}{\sqrt{12}}}$$

$$\therefore \quad z = -1.164$$

5. Decision

The calculated value of z i.e., $z = -1.164$, $|z| = 1.164$

The tabulated value of $z_\alpha = z_{0.05} = 1.96$

Since, the calculated value of z is less than the tabulated value of z.

Hence, H_0 is accepted.

4.2.3 Tests on Two Means

Q54. Explain the procedure for testing difference of two sample means.

Answer :

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

The standard normal variate Z corresponding to test whether there is any significant difference between $\bar{x}_1$ and $\bar{x}_2$ is,

$$Z = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

If samples are drawn from same population with standard deviation σ i.e., $\sigma_1^2 = \sigma_2^2 = \sigma$, then the test statistic becomes,

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In general, the two-sided hypothesis on two means can be written as

$$H_0 : \mu_1 - \mu_2 = d_0$$

The alternative hypothesis can be either one -sided or two sided and the distribution employed is the distribution of test statistic under H_0. The test statistic for σ_1 and σ_2 known is given by,

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Unknown But Equal Variances

Hypothesis Concerning Two Means (σ_1 and σ_2 Unknown; Small Sample)

The more common situations involving tests on two means are those in which variances are unknown. If we assume that distributions are normal and that $\sigma_1 = \sigma_2 = \sigma$. The pooled t-test (often called the two-sample t-test) may be used. The test statistic is given by the following test procedure.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

With n_1+n_2-2 degrees of freedom,

Where,

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

$$S_p^2 = \frac{\Sigma(x_i-\bar{x})^2+\Sigma(y_i-\bar{y})^2}{n_1+n_2-2}$$

Where $\bar{x}_1$, $\bar{x}_2$ are the means of two samples of size n_1 and n_2;

S_1^2, S_2^2 are the variances of two samples of size n_1 and n_2.

The critical region with this t-distribution can be obtained in a similar way.

For example when $A.H.$ is $\mu_1-\mu_2 \neq \delta$, the null hypothesis (H_0) is not rejected when $-t_{\frac{\alpha}{2}, n_1+n_2-2} < t < t_{\frac{\alpha}{2}, n_1+n_2-2}$ and the critical region is $t < -t_{\frac{\alpha}{2}, n_1+n_2-2}$ or $t < t_{\frac{\alpha}{2}, n_1+n_2-2}$

Critical region for testing $H_0 : \mu_1-\mu_2=\delta$

Alternate hypothesis;	Reject null hypothesis if
(i) $\mu_1-\mu_2 \neq \delta$	$t<-t_{\alpha/2}$ or $t>t_{\alpha/2}$
(ii) $\mu_1-\mu_2>\delta$	$t>t_\alpha$
(iii) $\mu_1-\mu_2<\delta$	$t<-t_\alpha$

Note

1. The two-sample t-test can not be used if $\sigma_1 \neq \sigma_2$.
2. The two-sample t-test can not be used for "before and after" kind of data, where the data is naturally paired.

   In other words the samples must be "independent" for two sample t-test.

PROBLEM

Q55. An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units? Assume the populations to be approximately normal with equal variances.

Solution :

Let μ_1 be the population mean for material 1 and μ_2 be the population mean for material 2.

Null Hypothesis H_0: $\mu_1-\mu_2=2$

Alternate Hypothesis H_1: $\mu_1-\mu_2>2$

Level of significance, $\alpha=0.05$

Test statistic $t = \dfrac{(\bar{x}_1-\bar{x}_2)-d_0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

$\bar{x}_1 = 85$ $\bar{x}_2 = 81$

$S_1 = 4$ $S_1 = 5$

$n_1 = 12$ $n_2 = 10$

We have,

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

$$= \sqrt{\frac{(12-1)(4)^2+(10-1)(5)^2}{12+10-2}}$$

$$= \sqrt{\frac{(11)(16)+(9)(25)}{20}}$$

$$= \sqrt{\frac{176+225}{20}}$$

$$= \sqrt{\frac{401}{20}}$$

$$= \sqrt{20.05}$$

$$= 4.478$$

Now, the t =statistic can be computed as, $t = \frac{(\bar{x}_1-\bar{x}_2)-d_0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

$$t = \frac{(85-81)-2}{4.478\left(\frac{1}{12}+\frac{1}{10}\right)} \qquad [\because \mu_1-\mu_2 = d_0 = 2]$$

$$t = \frac{4-2}{4.478\left(\sqrt{0.0833+0.1}\right)}$$

$$t = \frac{2}{4.478\sqrt{0.1833}}$$

$$t = 1.043$$

Computation of p-value is,

$$P = P(T > 1.043)$$

The value of P at 19(20 – 1) degrees of freedom with t = 1.043 is $\cong 0.16$

Therefore, Null Hypothesis is accepted and unable to conclude that abrasive wear of material 1 is two units greater than material 2.

4.2.4 Test on a Single Proportion

Q56. Explain the procedure for testing single proportion.

Answer :

Consider the problem of testing a Null Hypothesis

$$H_0 : P = P_0$$

Where 'p' represents the binomial distribution parameter.

The Alternative Hypothesis H_1 can be either one-sided or-two sided alternatives.

$$P < P_0, P > P_0 \text{ or } P \neq P_0$$

Case (i): $P < P_0$

Null Hypothesis $H_0 : P = P_0$

Alternate Hypothesis $H_1 : P < P_0$

The value of 'P' is computed using the binomial distribution.

$$P = P(X \leq x) \text{ when } P = P_0$$

'x' represents the number of successes in sample of size 'n'. If evaluated value of P is less than or equal to α, then test is significant at α level and null Hypothesis H_0 is rejected in favor of H_1.

Case(ii): $P > P_0$

Null Hypothesis $H_0 : P = P_0$

Alternate Hypothesis $H_1 : P > P_0$

According to Binomial distribution, the value of P is,

$P = P(X \geq x)$ when $P = P_0$

If evaluated value of 'P' is less than or equal to α then null hypothesis H_0 is rejected in favor of H_1.

Cast (iii): $P \neq P_0$

Null Hypothesis H_0: $P = P_0$

Alternate Hypothesis H_1: $P \neq P_0$

As per binomial distribution, we have,

$P = 2P(X \leq x$ when $P = P_0)$ if $x < \mu$

or

$P = 2P(X \geq x$ when $P = P_0)$ if $x > \mu$

If evaluated value of P is less than or equal to α then null Hypothesis H_0 is rejected in favor of H_1.

The steps involved in testing a proportion are as follows,

1. Null Hypothesis H_0: $P = P_0$.
2. Alternate Hypothesis H_1: $P < P_0$, $P > P_0$ or $P \neq P_0$.
3. Select the level of significance that is equal to α.
4. Test statistic: Binomial Variable X is used along $P = P_0$.
5. Computations: Evaluating the value of P by finding the value of x.
6. Decision: Final conclusion is made on the basis of P-value.

PROBLEMS

Q57. A builder claims that heat pumps are installed in 70% of all homes being constructed today in the city of Richmond, Virginia. Would you agree with this claim if a random survey of new homes in this city showed that 8 out of 15 had heat pumps installed? Use a 0.10 level of significance.

Solution :

Given that,

$n = 15$

$P = 0.7$

$x = 8$

Null hypothesis $H_0 : p = 0.7$

Alternate hypothesis $H_1 : p \neq 0.7$

Level of significance $\alpha = 0.10$

Test statistic: Binomial variable X with $p = 0.7$ and $n = 15$

$x = 8$

$\mu P_0 = 15(0.7)$

$= 10.5$

In two tailed test, the P-value is twice the area of the shaded region.

$P = 2P\,(x \leq 8$ when $P = 07)$

$= 2\sum_{x=0}^{8} b(x; 15, 0.7)$

$= 2(0.1311)$

$= 0.2622 > 0.1$

Therefore, null hypothesis is accepted and concluded that there is no proper reason to doubt the claim of builder.

Q58. A commonly prescribed drug for relieving nervous tension is believed to be only 60% effective. Experimental results with a new drug administered to a random sample of 100 adults who were suffering from nervous tension show that 70 received relief. Is this sufficient evidence co conclude that the new drug is superior to the one commonly prescribed? Use a 0.05 level of significance.

Solution :

1. Null hypothesis (H_0), $P = \frac{60}{100} = 0.6$
2. Alternative hypothesis (H_1), $P > 0.6$
3. Level of significance, $\alpha = 0.05$
4. Critical region, Reject if $Z > 1.645$
5. Test statistic, $Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$

Where,

P = Proportion of adult who got relief from nervous tension = 70/100 = 0.7

n = Sample size = 100

$P = 0.6$

$Q = 1 - P$

$Q = 1 - 0.6$

$Q = 0.4$

$$Z = \frac{0.7 - 0.6}{\sqrt{\frac{(0.6)(0.4)}{100}}}$$

$$Z = \frac{0.1}{\sqrt{\frac{0.24}{100}}}$$

$$Z = \frac{0.1}{\sqrt{0.002}}$$

$$Z = \frac{0.1}{0.049}$$

$\therefore Z = 2.041$

$Z_{cal} > Z_{tab}$

$2.041 > 1.645$

6. Decision

Since the calculated value of Z is greater than the tabulated value. The null hypothesis is rejected and alternative hypothesis is accepted.

Therefore, the new drug is superior.

4.2.5 Two Samples: Tests on Two Proportions

Q59. Discuss hypothesis concerning two proportions.

Answer : Model Paper-III, Q9(b)

Hypothesis Concerning Two Proportions

There are many situations in which we wish to test the hypothesis that two proportions are equal. For example, we might want to show evidence that the proportion of engineers in state is equal to the proportion of engineers in another state. A person may decide to give up smoking only if he or she is convinced that the proportion of smokers with lung cancer is more than the proportion of non smokers with lung cancer.

In general, we want to test the null hypothesis that two proportions or binomial parameters are equal. That is, we wish to test the null hypothesis $H_0 : p_1 = p_2$, against one of the alternatives $p_1 < p_2, p_1 > p_2$, or $p_1 \neq p_2$.

Ofcourse, this is equivalent to testing the null hypothesis that $p_1 - p_2 = 0$, against one of the alternatives $p_1 - p_2 < 0, p_1 - p_2 > 0$, or $p_1 - p_2 \neq 0$. $p_1 - p_2$ is a random variable on which we base our discussion. Let us suppose that there are two distinct populations A and B. Independent samples of size n_1 and n_2 are selected at random from this two binomial populations and the proportion of success p_1 and p_2 for the two samples is computed.

We know that from previous discussion of construction of confidence intervals for p_1 and p_2, for n_1 and n_2 sufficiently large, the point estimator.

$p_1 - p_2$ was approximately normally distributed with mean

$$\mu_{p_1 - p_2} = 0 \text{ and variance } \sigma^2_{p_1 - p_2} = \frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}$$

Here an unbiased pooled estimate of the population proportion $\hat{p}$ is,

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

Where x_1 and x_2 are the number of success in each of the two samples.

The Z–value for testing $p_1 = p_2$ is determined from the formula,

$$Z = \frac{p_1 - p_2}{\sigma_{p_1 + p_2}} = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{P_1 - P_2}{\sqrt{\hat{p}\hat{q}\left(\frac{n_1 + n_2}{n_1 n_2}\right)}}$$

The critical region for testing the null hypothesis, $p_1 = p_2$ against the number of alternatives at $\propto$ level of significance.

Null Hypothesis Alternative Hypothesis and Critical Region

1. $H_0 : P_1 = P_2 V_S$
   $H_1 : P_1 \neq P_2$
   C.R.: $Z < -Z_{\alpha/2}$
   $Z > Z_{\alpha/2}$
2. $H_0 : P_1 = P_2$
   $H_1 : P_1 > P_2$
   C.R.: $Z > Z_\alpha$
3. $H_0 : P_1 = P_2$
   $H_1 : P_1 < P_2$
   C.R.: $Z < -Z_\alpha$

PROBLEM

Q60. A vote is to be taken among the residents of a town and the surrounding county to determine whether a proposed chemical plant should be constructed. The construction site is within the town limits and for this reason many voters in the county believe that the proposal will pass because of the large proportion of town voters who favor the construction. To determine if there is a significant difference in the proportions of town voters and county voters favoring the proposal, a poll is taken. If 120 of 200 town voters favor the proposal and 240 of 500 county residents favor it, would you agree that the proportion of town voters favoring the proposal is higher than the proportion of county voters? Use an $\alpha = 0.05$ level of significance.

Solution :

Let P_1 be the true proportion of voters in the town and P_2 be the true proportion of voters in the country

$x_1 = 120, x_2 = 240$

$n_1 = 200, n_2 = 500$

Null hypothesis $H_0 : P_1 = P_2$

Alternate hypothesis $H_1 : P_1 > P_2$

Level of significance, $\alpha = 0.05$

$$\hat{P}_1 = \frac{x_1}{n_1}$$

$$= \frac{120}{200}$$

$$= 0.60$$

$$\hat{P}_2 = \frac{x_2}{n_2}$$

$$= \frac{240}{500}$$

$$= 0.48$$

$$\hat{P} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$= \frac{120 + 240}{200 + 500}$$

$$= \frac{360}{700}$$

$$= 0.51$$

Test statistic,

$$z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.60 - 0.48}{\sqrt{(0.51)(0.49)\left(\frac{1}{200} + \frac{1}{500}\right)}}$$

$$= \frac{0.120}{\sqrt{0.25 \times 0.007}}$$

$$= \frac{0.120}{0.042}$$

$$z = 2.857 \cong 2.9$$

Computation of p-value,

$$P = P(z > 2.9)$$

$$= 1 - 0.9981$$

$$= 0.0019E$$

Therefore, Null hypothesis H_0 is rejected and concluded that proportion of town voters favoring the proposal is greater than the proportion of country voters.