UNIVERSITY OF CALICUT
SCHOOL OF DISTANCE EDUCATION

STUDY MATERIAL

B.Sc. MATHEMATICS

(2011 Admission onwards)

III SEMESTER

COMPLEMENTARY COURSE

(STATISTICS)

STATISTICAL INFERENCE

Prepared by:

Dr. K. ANEESH KUMAR
Department of Statistics
Mahatma Gandhi College, Iritty

Scrutinised by:

Dr. ANIL KUMAR .V.
Reader & Head
Department of Mathematics
University of Calicut

Layout & Settings:

Computer Cell, SDE

©
Reserved
CONTENTS

1. SAMPLING DISTRIBUTIONS 5 – 29
   - Sampling distributions
   - Chi-square distribution
   - Student’s t-distribution
   - Snedecor’s F-distribution

2. THEORY OF ESTIMATION-I (POINT ESTIMATION) 30 - 51
   - Statistical Inference
   - Point estimation
   - Desirable properties of good estimator
   - Methods of estimation

3. INTERVAL ESTIMATION 52 – 68
   - Interval estimation
   - Confidence interval for the mean of a normal population
   - Confidence interval for the difference of means
   - Confidence interval for the variance of a normal population
   - Confidence interval for large samples

4. TESTING OF HYPOTHESIS 69 – 89
   - Statistical hypothesis
   - Testing of hypothesis
   - Errors in testing of hypothesis
   - Steps in testing of hypothesis
   - Most powerful test
   - Neyman-Pearson theorem

5. LARGE SAMPLE TESTS 90 – 124
   - Large sample tests
   - Testing mean of a population
   - Testing the inequalities of means of two populations
   - Testing the population proportion
   - Testing the equality of proportion of two populations
   - Goodness of fit
   - $\chi^2$ – test of independence

6. SMALL SAMPLE TESTS 125 – 139
   - Small sample tests
   - Tests based on normally distributed test statistics
   - Tests based on statistics following t – distribution
   - Tests based on statistics following $\chi^2$ - distribution
   - Tests based on statistics following F - distribution
1.1. Sampling Distribution

The characteristics of a large group of individuals are studying in any statistical investigation. This group is referred to as the population under investigation. Let us study only one characteristic, say $X$ – the height of the individuals in the population. Corresponding to each individual of the population we get a number denoting magnitude of the characteristic considered. The set of such numbers are called the statistical population. This statistical population is considered as the set of admissible values of the variable $X$ (here height). The distribution of the values of $X$ is known as the distribution of this statistical population (or simply saying population now onwards).

Any function of the statistical population values are called population parameter. For eg., mean, variance, median etc., of the variable considered.

The process of making inference about the population based on samples taken from the population is known as statistical inference.

Consider a random sample taken from a population then the function of sample values like sample mean, sample variance sample moments etc., are known as statistic.

The distribution of statistic is known as the sampling distribution of that statistic.

Sampling distribution of means of random samples taken from a normal population $N(\mu, \sigma)$:

Let $X_1, X_2, ..., X_n$ be the random samples taken from $N(\mu, \sigma)$. Then,

Sample mean $\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}$

$mgf$ of $\bar{X}$, $M_{\bar{X}}(t) = M_{\frac{X_1 + X_2 + ... + X_n}{n}}(t) = M_{\frac{X_1 + X_2 + ... + X_n}{n}}\left(\frac{t}{n}\right)$

$= M_{\frac{X_1}{n}}\left(\frac{t}{n}\right)M_{\frac{X_2}{n}}\left(\frac{t}{n}\right)....M_{\frac{X_n}{n}}\left(\frac{t}{n}\right)$ ($\because X_i$'s are ind.)
We have for $X \sim N(\mu, \sigma)$,

$$M_X(t) = e^{\mu t + \frac{t^2 \sigma^2}{2}}$$

Therefore,

$$M_{\overline{X}}(t) = e^{\mu \frac{t}{n} + \frac{t^2 \sigma^2}{2n^2}} = e^{\mu \frac{t}{n} + \frac{t^2 \sigma^2}{2n^2}}$$

$$= (\mu \frac{t}{n} + \frac{t^2 \sigma^2}{2n^2})^n = e^{\mu \frac{t}{n} + \frac{t^2 \sigma^2}{2n^2}}$$

It is the m.g.f. of a normal population with parameters $\mu$ and $\frac{\sigma}{\sqrt{n}}$. This implies that

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Problem 1: A random sample of size 25 is taken from a normal population with mean 1 and variance 9. What is the probability that the sample mean is negative?

Solution:

Given sample size $n = 25$, $\mu = 1$ and $\sigma = 3$.

We have $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

The required probability is $P(\overline{X} < 0)$

$$= P\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{0 - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= P\left(Z < -\frac{1}{3} / \frac{3}{\sqrt{25}}\right), \text{ where } Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

$$= P(Z < -1.67) = P(Z > 1.67)$$

$$= P(Z > 0) - P(0 < Z < 1.67)$$

($\because$ Standard normal curve is symmetric about the y-axis)

$$= 0.5 - 0.4525 \text{ (from standard normal table)}$$

$$= 0.0475$$

Problem 2: A random sample is taken from a normal population with mean 10 and variance 9. How large a sample should be taken if the sample mean is to lie between 9 and 11 with probability 0.95?
Solution:

To find the minimum number of samples so as $P(9 < \bar{X} < 11) = 0.95$

Let $n$ samples are taken from the given population $N(10, 3)$,

then we have, $\bar{X} \sim N(10, \frac{3}{\sqrt{n}})$

ie., $Z = \frac{\bar{X} - 10}{\frac{3}{\sqrt{n}}} \sim N(0,1)$

$P(9 < \bar{X} < 11) = 0.95 \Rightarrow P\left(\frac{9-10}{\frac{3}{\sqrt{n}}} < \frac{\bar{X} - 10}{\frac{3}{\sqrt{n}}} < \frac{11-10}{\frac{3}{\sqrt{n}}}\right) = 0.95$

ie., $P\left(\frac{-1}{3}\sqrt{n} < Z < \frac{1}{3}\sqrt{n}\right) = 0.95$

ie., $P(0 < Z < \frac{1}{3}\sqrt{n}) = 0.475$

For $Z \sim N(0,1)$, $P(0 < Z < \frac{1}{3}\sqrt{n}) = 0.475$ happens when

$\frac{1}{3}\sqrt{n} = 1.96 \Rightarrow \sqrt{n} = 5.88 \Rightarrow n = 34.58$

So the minimum number of sample for the required probability condition is 35.

1.2. Chi-square distribution ($\chi^2$-distribution)

A random variable $X$ with pdf,

$$f(x) = \begin{cases} \frac{(\frac{1}{2})^n}{\Gamma(n/2)} e^{-\frac{x}{2}} x^{n/2 - 1}, & 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$$

is said to follow $\chi^2$ distribution with $n$ degrees of freedom, where $n$ is the parameter of the distribution, denoted by $X \sim \chi^2(n)$.

**Moment generating function of $X \sim \chi^2(n)$:**

Observe that, $M_X(t) = E(e^{tX})$
\[
\int_0^\infty e^{tx} \left( \frac{1}{2} \right)^{\frac{n}{2}} e^{-\frac{x}{2}n^{-1}} \, dx = \left( \frac{1}{2} \right)^{\frac{n}{2}} \int_0^\infty e^{-x(\frac{1}{2} - t)\frac{n}{2}^{-1}} \, dx
\]

\[
= \left( \frac{1}{2} \right)^{\frac{n}{2}} \int_0^\infty e^{- \frac{nt}{2(1-t)}} \, dx, \quad (\because \frac{\sqrt{n}}{2} = \int_0^\infty e^{-kx} x^{n-1} \, dx)
\]

This implies that

\[
M_x(t) = (1 - 2t)^{-\frac{n}{2}}
\]

**Mean and Variance:**

Observe that,

\[
E(X) = \frac{d}{dt} M_X(t) \bigg|_{t = 0}
\]

\[
= \left[ \frac{d}{dt} (1 - 2t)^{-\frac{n}{2}} \right] \bigg|_{t = 0}
\]

\[
= -\left[ \frac{2n}{2} (1 - 2t)^{-\frac{n}{2}} - 1 \right] \bigg|_{t = 0} = 0 = n.
\]

Hence, \(E(X) = n\)

Moreover, \(E(X^2) = \left[ \frac{d^2}{dt^2} (1 - 2t)^{-\frac{n}{2}} \right] \bigg|_{t = 0}
\]

\[
= \left[ -2n \left( \frac{n}{2} - 1 \right) (1 - 2t)^{-\frac{n}{2}} - 1 \right] \bigg|_{t = 0} = n(n + 2).
\]

Hence, \(V(X) = n(n+2) - n^2 = 2n\)

**Additive property of \(\chi^2\) distribution:**

Let \(X_1\) and \(X_2\) are two independent \(\chi^2\) random variables where \(X_1 \sim \chi^2(n_1)\) and \(X_2 \sim \chi^2(n_2)\). Then \(X_1 + X_2\) follow chi-square distribution with \(n_1 + n_2\) degrees of freedom.
Proof: Observe that

\[ X_1 \sim \chi^2(n_1) \implies M_{X_1}(t) = (1-2t)^{-\frac{n_1}{2}} \quad \text{and} \]

\[ X_2 \sim \chi^2(n_2) \implies M_{X_2}(t) = (1-2t)^{-\frac{n_2}{2}}. \]

Therefore, \( M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t) = (1-2t)^{-\frac{n_1}{2}}(1-2t)^{-\frac{n_2}{2}} = (1-2t)^{-\left(\frac{n_1+n_2}{2}\right)} \)

This is the m.g.f. of a \( \chi^2 \) random variable with \( n_1+n_2 \) degrees of freedom. Hence \( X_1 + X_2 \) follow chi-square distribution with \( n_1+n_2 \) degrees of freedom.

In general if, \( X_1, X_2, \ldots, X_k \) are \( n \) independent random variables with \( n_1, n_2, \ldots, n_k \) degrees of freedom respectively, then \( X_1 + X_2 + \ldots + X_k \) follow chi-square distribution with \( n_1 + n_2 + \ldots + n_k \) degrees of freedom.

Tables of chi-square distribution:

Chi-square table gives the values of \( \chi^2(\alpha) \) for a \( \chi^2 \) variable with various degrees of freedom and for various values of \( \alpha \), such that \( P(\chi^2 > \chi^2(\alpha)) = \alpha \).

**Theorem:** If \( X \sim N(0, 1) \), then \( Y = X^2 \) follow chi-square distribution with one degree of freedom.

**Proof:**

Observe that \( M_Y(t) = M_{X^2}(t) = E(e^{X^2}) \)

\[
= \int_{-\infty}^{\infty} e^{tx^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx
\]
\[
\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-x^2}{2}} \, dx = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{\frac{-x^2}{2}} \, dx
\]

(: \: e^{\frac{-x^2}{2}} \text{ is an even function})

\[
= \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{\frac{x^2(t-\frac{1}{2})}{2}} \, dx
\]

\text{put } x^2 = u \Rightarrow dx = \frac{du}{2x} = \frac{du}{2\sqrt{u}}

\[M_y(t) = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{\frac{u(t-\frac{1}{2})}{2}} \frac{du}{2\sqrt{u}} \]

\[= \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-u(\frac{1}{2}-t)} \frac{1}{u^\frac{1}{2}} \, du = \frac{1}{\sqrt{2\pi}} \frac{1}{(\frac{1}{2}-t)^\frac{1}{2}} \]

This implies that \[M_y(t) = (1 - 2t)^{-\frac{x}{2}} \]

But \((1-2t)^{-\frac{1}{2}}\) is the m.g.f. of a chi-square random variable with one degree of freedom. Hence \(Y \sim \chi^2(1)\).

**Theorem:** If \(X_1, X_2, \ldots, X_n\) are \(n\) random samples taken from a standard normal population, then the sum of squares of random sample follows \(\chi^2(n)\)

**Proof:**

Since \(X_i\)'s are random samples from \(N(0, 1)\), they are independent and \(X_i^2 \sim \chi^2(1) \text{ for all } i\). If \(X\) follows \(N(0, 1)\), we have \(X^2 \sim \chi^2(1)\).

Here \(X_i^2 \sim \chi^2(1) \text{ for all } i\). Hence, by additive property of chi-square distribution, the sum of squares of \(X_1, X_2, \ldots, X_n\) follow \(\chi^2(n)\).

**Theorem:** If \(X_1, X_2, \ldots, X_n\) are \(n\) random samples taken from \(N(\mu, \sigma)\), Show that

\[Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)\]
Proof:

Given $X_1, X_2, \ldots, X_n$ are independent samples from $N(\mu, \sigma)$.

Then $\frac{X_1 - \mu}{\sigma}, \frac{X_2 - \mu}{\sigma}, \ldots, \frac{X_n - \mu}{\sigma} \sim N(0, 1)$

\[
\left(\frac{X_i - \mu}{\sigma}\right)^2, \left(\frac{X_2 - \mu}{\sigma}\right)^2, \ldots, \left(\frac{X_n - \mu}{\sigma}\right)^2
\]

each follow $\chi^2(1)$

Since $X_1, X_2, \ldots, X_n$ are independent, by additive property of chi-square distribution, we get,

\[
\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n).
\]

Problem 1: If $X_1, X_2, \ldots, X_n$ are $n$ random samples taken from $N(\mu, \sigma)$, find the distribution of sample variance $S^2$.

Solution:

For the random samples $X_1, X_2, \ldots, X_n$ taken from $N(\mu, \sigma)$, we have

\[
Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n).
\]

Note that $Y$ can be written as

\[
Y = \sum_{i=1}^{n} \left(\frac{x_i - \bar{X} + \bar{X} - \mu}{\sigma}\right)^2,
\]

where $\bar{X}$ is the sample mean

\[
= \sum_{i=1}^{n} \left(\frac{x_i - \bar{X} + \bar{X} - \mu}{\sigma}\right)^2
\]

\[
= \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2 + (\bar{X} - \mu)^2 + 2(x_i - \bar{X})(\bar{X} - \mu)}{\sigma^2}
\]

\[
= \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{\sigma^2} + \sum_{i=1}^{n} \frac{(\bar{X} - \mu)^2}{\sigma^2} + \sum_{i=1}^{n} \frac{2(x_i - \bar{X})(\bar{X} - \mu)}{\sigma^2}
\]

\[
= \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2} + 2\sum_{i=1}^{n} \frac{(x_i - \bar{X})(\bar{X} - \mu)}{\sigma^2}
\]

\[
= \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2}
\]

\[
= \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2
\]

\[
= \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{\sigma^2} + \frac{nS^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2
\]
Observe that \( \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \text{N}(0, 1) \) and

\[
\left( \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right)^2 \quad \text{follow chi-square distribution with one degree of freedom. Since } \bar{X} \text{ and } S^2 \text{ are independently distributed, then by additive property of chi-square distribution }
\]

\[\frac{nS^2}{\sigma^2} \sim \chi^2(n-1).\]

Let \( U = \frac{nS^2}{\sigma^2} \sim \chi^2(n-1) \quad \text{(1)} \)

Then,

\[
f(u) = \begin{cases} 
\frac{1}{n^{n-1}} e^{-u/n} u^{n-1}, & 0 < u < \infty \\
0, & \text{otherwise}
\end{cases}
\]

Equation (1) can be written as \( S^2 = \frac{u \sigma^2}{n} \), therefore \( f(S^2) = f(u) \bigg| \frac{du}{dS^2} \bigg| \)

\( S^2 = \frac{u \sigma^2}{n} \), then, \( f(S^2) = f(u) \) in terms of \( S^2 \times \frac{du}{dS^2} \)

\[
= \left( \frac{1}{n^{n-1}} \right) e^{-\frac{u \sigma^2}{2n}} \left( \frac{nS^2}{\sigma^2} \right)^{n-1/2} \times \frac{du}{dS^2}
\]

\[
= \left( \frac{1}{n^{n-1}} \right) e^{-\frac{u \sigma^2}{2n}} \left( \frac{nS^2}{\sigma^2} \right)^{n-1/2} \times \frac{n}{\sigma^2}.
\]

Hence,

\[
f(S^2) = \left( \frac{n}{2 \sigma^2} \right)^{n-1/2} e^{-\frac{nS^2}{2 \sigma^2}} (S^2)^{n-1/2}, \quad 0 < S^2 < \infty
\]

**Problem 2:** If \( X \sim \chi^2(n) \), find the mode of \( X \).

**Solution:**

We have \( f(x) = \left( \frac{1}{2} \right)^{\frac{n}{2}} e^{-\frac{x}{2}} x^{n-1} \)
Mode is the point \( x \) where \( f(x) \) attains its maximum. That is, the point \( x \), where 

\[
 f'(x) = 0 \quad \text{and} \quad f''(x) < 0.
\]

\[
 f'(x) = \left( \frac{\theta}{2} \right)^2 \left[ e^{-\frac{x}{\theta}} \left( \frac{x}{\theta} \right)^{n-2} - e^{-\frac{x}{\theta}} \frac{n}{2} \frac{n}{2} \right]
\]

\[
 f'(x) = 0 \quad \Rightarrow \quad \left[ e^{-\frac{x}{\theta}} \left( \frac{x}{\theta} \right)^{n-2} - e^{-\frac{x}{\theta}} \frac{n}{2} \frac{n}{2} \right] = 0
\]

\[
 \Rightarrow \quad e^{-\frac{x}{\theta}} \left( \frac{x}{\theta} \right)^{n-2} = e^{-\frac{x}{\theta}} \frac{n}{2} \frac{n}{2}
\]

\[
 \Rightarrow \quad \left( \frac{n}{2} \right)^{n-2} = x^{-1} \quad \Rightarrow \quad x^{-1} = \frac{1}{n-2}
\]

\[
 \Rightarrow \quad x = n-2
\]

For the value of \( x = n-2 \), \( f''(x) < 0 \). Hence the mode is at \( x = n-2 \).

**Problem 3:** If \( X \) is distributed as \( f(x) = \frac{1}{\theta}, \ 0 < x < \theta \), show that \( -2 \log_e \left( \frac{x}{\theta} \right) \) follows chi-square distribution with 2 degrees of freedom.

**Solution:**

Let \( Y = -2 \log_e \left( \frac{x}{\theta} \right) \). This can be written as \( x = \theta e^{-\frac{Y}{2}} \).

Note that \( \left| \frac{dx}{dy} \right| = \frac{\theta}{2} e^{-\frac{Y}{2}} \)

\[
 \therefore \quad f(y) = f(x) \text{ in terms of } y \cdot \left| \frac{dx}{dy} \right| = \frac{\theta}{2} e^{-\frac{Y}{2}}
\]

\[
 = \frac{\theta}{2} e^{-\frac{Y}{2}}
\]

\[
 = \frac{1}{2} \frac{\theta}{2} e^{-\frac{Y}{2}}
\]

\[
 = \left( \frac{1}{2} \right)^{2-1} \frac{\theta}{2} e^{-\frac{Y}{2}} \frac{2}{2-1}
\]

\[
 f(y) = \left( \frac{1}{2} \right)^{2-1} \frac{\theta}{2} e^{-\frac{Y}{2}} \frac{2}{2-1}
\]

This implies that \( Y \sim X_2(2) \).
Problem 4: For large $n$, show that chi-square distribution approximately normally distributed.

Solution:

Consider a random variable $X$ following chi-square distribution with $n$ degrees of freedom. Then $M_X(t) = (1 - 2t)^{-n/2}$. More over,

$$E(X) = n \quad \text{and} \quad V(X) = 2n.$$

Consider $Z = \frac{X - n}{\sqrt{2n}}$, then

$$M_Z(t) = M_{X-n}(t) = e^{-\frac{n}{\sqrt{2n}}t} M_X\left(\frac{t}{\sqrt{2n}}\right)$$

$$= e^{-\frac{n}{\sqrt{2n}}t}\left(1 - 2\frac{t}{\sqrt{2n}}\right)^{-n/2}. $$

Therefore,

$$\log M_Z(t) = \log \left(e^{-\frac{n}{\sqrt{2n}}t}\left(1 - 2\frac{t}{\sqrt{2n}}\right)^{-n/2}\right)$$

$$= -\frac{n}{\sqrt{2n}}t - \frac{n}{2}\log\left(1 - 2\frac{t}{\sqrt{2n}}\right)$$

$$= -\frac{n}{\sqrt{2n}}t + \frac{n}{2}\left(2\frac{t}{\sqrt{2n}} + \frac{1}{2} \left(2\frac{t}{\sqrt{2n}}\right)^2 + \ldots\right)$$

$$= -\frac{n}{\sqrt{2n}}t + \frac{n}{\sqrt{2n}}t + \frac{t^2}{2} + \ldots + \text{(many terms involving $n^2$ and its higher power in denominator)}$$

As $n$ become very large, $\log M_Z(t) \to \frac{t^2}{2}$.

That is $M_Z(t) \to e^{\frac{t^2}{2}}$. This is the m.g.f. of a standard normal random variable. Then by uniqueness theorem of m.g.f., $X \sim N(n, \sqrt{2n})$ for large $n$. 

Problem 5: For a random sample of size 16 from \( N(\mu, \sigma) \) population, the sample variance is 16.

Find \( a \) and \( b \) such that \( P( a < \sigma^2 < b ) = 0.60 \)

Solution:

\[
P( a < \sigma^2 < b ) = 0.60 \text{ implies } P( \frac{1}{a} > \frac{1}{\sigma^2} > \frac{1}{b} ) = 0.60
\]

\[
\Rightarrow P( \frac{nS^2}{a} > \frac{nS^2}{\sigma^2} > \frac{nS^2}{b} ) = 0.60, \text{ where } \frac{nS^2}{\sigma^2} \sim \chi^2_{(n-1)}.
\]

Putting \( S^2 = 16 \) and \( n = 16 \) in the above equation, we get

\[
P( \frac{16 \times 16}{a} > \chi^2_{(16-1)} > \frac{16 \times 16}{b} ) = 0.60
\]

That is,

\[
P( \frac{16 \times 16}{b} < \chi^2_{(15)} < \frac{16 \times 16}{a} ) = 0.60 \quad \text{--------- (1)}
\]

From the table of chi-square distribution,

\[
P(\chi^2_{(15)} > 10.307 ) = 0.80 \text{ and } P(\chi^2_{(15)} > 19.311) = 0.20.
\]

Therefore,

\[
P(10.307 < \chi^2_{(15)} < 19.311) = 0.60 \quad \text{--------- (2)}
\]

Comparing (1) and (2), we get \( \frac{16 \times 16}{b} = 10.307 \) and \( \frac{16 \times 16}{a} = 19.311 \)

Hence \( b = \frac{16 \times 16}{10.307} = 24.84 \), and \( a = \frac{16 \times 16}{19.311} = 13.26 \)

\[
\text{i.e., } P( 13.26 < \sigma^2 < 24.84 ) = 0.60.
\]

1.3 Student’s t-distribution

This is the probability distribution which was introduced by W.S Gosset and known in his pen name ‘student’. A continuous random variable \( t \) with density function

\[
f(t) = \frac{\left(\frac{n+1}{n}\right)^{\frac{n+1}{2}}}{\sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < t < \infty
\]

is said to follow student’s t-distribution with \( n \) degrees of freedom.
Note that if \( X \sim N(0,1) \) and \( Y \sim \chi^2(n) \), then \( t = \frac{X}{\sqrt{Y/n}} \) follows t-distribution with \( n \) degrees of freedom, where \( X \) and \( Y \) are independent.

**Examples of statistics following student’s t-distribution:**

1. Let \( \bar{X} \) be the mean of \( n \) random samples taken from \( N(\mu, \sigma) \). Let \( S^2 \) be the sample variance. Then we have \( \bar{X} \sim N(\mu, \sigma/\sqrt{n}) \) and \( \frac{nS^2}{\sigma^2} \sim \chi^2(n-1) \).

   Note that \( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \).

   Therefore, \( t = \frac{nS^2}{\sigma^2} \sim t_{(n-1)} \).

   This implies that \( t = \frac{(\bar{X} - \mu)\sqrt{n-1}}{S} \sim t_{(n-1)} \).

2. Let \( \bar{X}_1 \) and \( \bar{X}_2 \) be the means, \( S_1 \) and \( S_2 \) be the standard deviations of samples of sizes \( n_1 \) and \( n_2 \) taken independently from two normal populations with same mean \( \mu \) and standard deviation \( \sigma \). Then,

   \[
   t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{(n_1 + n_2 - 2)}
   \]

   Proof:

   We have \( \bar{X}_1 \sim N(\mu, \sigma/\sqrt{n_1}) \) and \( \bar{X}_2 \sim N(\mu, \sigma/\sqrt{n_2}) \).

   Then, \( \bar{X}_1 - \bar{X}_2 \sim N\left(0, \sigma^2 \frac{1}{n_1} + \frac{1}{n_2}\right) \).

   This implies that \( \frac{\bar{X}_1 - \bar{X}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1) \).
Also we have, \( \frac{n_1S_1^2}{\sigma^2} \sim \chi^2(n_1-1) \) and \( \frac{n_2S_2^2}{\sigma^2} \sim \chi^2(n_2-1) \).

Using additive (reproductive) property of chi-square distribution, we get

\[
\frac{n_1S_1^2}{\sigma^2} + \frac{n_2S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2) \quad (\because \frac{n_1S_1^2}{\sigma^2} \text{ and } \frac{n_2S_2^2}{\sigma^2} \text{ are ind.})
\]

So,

\[
t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sim t_{(n_1 + n_2 - 2)}
\]

That is,

\[
t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{n_1S_1^2 + n_2S_2^2}{n_1 + n_2 - 2}} \sim t_{(n_1 + n_2 - 2)}
\]

That is,

\[
t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{n_1S_1^2 + n_2S_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{(n_1 + n_2 - 2)}
\]

**Tables of t-distribution:**

Note that t-distribution is symmetric about zero and bell shaped. Tables of t-distribution gives the values of \( t_\alpha \) for various degrees of freedom and for various value of \( \alpha \), such that \( P(|t| > t_\alpha) = \alpha \).

**Problem 1:** If \( t \sim t_{(n)} \), find the mode of \( t \).

**Solution:** We have, \( f(t) = \frac{1}{\sqrt{2\pi n}} \left( 1 + \frac{t^2}{n} \right)^{-\frac{n+1}{2}} \), \( -\infty < t < \infty \).
Mode is the point \( t \) where \( f(t) \) attains its maximum. That is the point \( t \), where \( f'(t) = 0 \) and \( f''(t) < 0 \).

\[
f'(t) = -\frac{n+1}{2\sqrt{n\pi}} \frac{(n+1)}{2} \left( 1 + \frac{t^2}{n} \right)^{-\left(\frac{n+1}{2}\right)-1} \frac{2t}{n}. \]

\[
f'(t) = 0 \quad \Rightarrow \quad 2t \left( 1 + \frac{t^2}{n} \right)^{-\left(\frac{n+3}{2}\right)} = 0
\]

\[
\Rightarrow \quad t = 0
\]

\[
f''(t) = -\frac{n+1}{2\sqrt{n\pi}} \frac{(n+1)}{2} \left[ -\left( \frac{n+3}{2} \right) \left( 1 + \frac{t^2}{n} \right)^{-\left(\frac{n+3}{2}\right)-1} t + \left( 1 + \frac{t^2}{n} \right)^{-\left(\frac{n+3}{2}\right)} \right].
\]

At \( t = 0 \), \( f''(x) = -\frac{n+1}{2\sqrt{n\pi}} \frac{(n+1)}{2} < 0 \). Hence Mode of \( t \) is at \( t = 0 \).

**Problem 2:** If \( t \sim t_{(n)} \), then as \( n \to \infty \), prove that \( t \sim N(0,1) \).

**Solution:**

Given \( t \sim t_{(n)} \), we have \( f(t) = \frac{n+1}{2\sqrt{n\pi}} \left( 1 + \frac{t^2}{n} \right)^{-\left(\frac{n+1}{2}\right)} \), \(-\infty < t < \infty \)

Using the following results we can prove as \( n \to \infty \), \( t \sim N(0,1) \)

(i) as \( n \) becomes very large, \( \frac{n+k}{n} \approx n^k \)

(ii) \( \lim_{n \to \infty} \left( 1 + \frac{\lambda}{n} \right)^n = e^\lambda \)

Note that as \( n \to \infty \)
$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

which is the probability density function of standard normal distribution.

**Problem 3:** If \( t \sim t_{(5)} \), find a such that, \( P(-a < t < a) = 0.98 \)

**Solution:**

The graph of \( t \sim t_{(5)} \) is symmetric about zero. To find \( a \) such that the area under the \( t \) curve between \(-a\) and \( a\) is 0.98.

ie., to find \( a \) such that \( P(t < a) = 0.98 \) or \( P(t > a) = 0.02 \)

From table of t-distribution, for 5 d.f. we get, \( a = 3.365 \).

**Problem 4:** Prove that the ratio of two independent standard normal random variables is a student’s \( t \) random variable with 1 degree of freedom.

**Solution:**

Given \( X_1 \sim N(0,1) \) , \( X_2 \sim N(0,1) \) and they are independent.

Then, \( X_2^2 \sim \chi^2_{(1)} \). This implies that

$$t = \frac{X_1}{\sqrt{X_2^2/1}} \sim t_{(1)}$$

That is, \( t = \frac{X_1}{X_2} \sim t_{(1)} \)

That is, the ratio of two independent standard normal variables follows \( t_{(1)} \).

**Problem 5:** If \( X_1 \) and \( X_2 \) are two independent standard normal variables, find the distribution of \( t = \frac{\sqrt{2}X_1}{\sqrt{X_1^2 + X_2^2}} \).
Solution:

Given $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$ and they are independent.

So, $X_1^2 \sim \chi^2_1$ and $X_2^2 \sim \chi^2_1$.

By additive (reproductive) property of chi-square distribution,

$$X_1^2 + X_2^2 \sim \chi^2_2.$$ 

Then, 

$$t = \frac{X_1}{\sqrt{\frac{X_1^2 + X_2^2}{2}}} \sim t_{(2)}.$$ 

That is, 

$$t = \frac{\sqrt{2}X_1}{\sqrt{X_1^2 + X_2^2}}$$ follow t-distribution with 2 degrees of freedom.

Problem 6: Find the maximum difference that we can expect with probability 0.95 between the means of samples of sizes 10 and 12 from a normal population, if their standard deviations are found to be 2 and 3 respectively.

Solution:

Let $\bar{x}_1$ and $\bar{x}_2$ be the means of the samples of sizes $n_1 = 10$ and $n_2 = 12$ taken randomly from two normal populations. Assume samples are taken independently. The sample variances $S_1^2 = 4$ and $S_2^2 = 9$ respectively.

To find the value of $k$ such that,

$$P\left( |\bar{x}_1 - \bar{x}_2| \leq k \right) = 0.95.$$ 

For samples taken from two normal populations independently, we know that,

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{(n_1 + n_2 - 2)}$$ 

Then to find $k$ such that,

$$P\left( \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \right| \leq k \right) = 0.95$$ 

$$\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$
That is, 
\[ P \left( \left| t_{(10+12-2)} \right| \leq \frac{k}{\sqrt{10 \times 4 + 12 \times 9 \left( \frac{1}{10} + \frac{1}{12} \right)}} \right) = 0.95. \]

That is, 
\[ P \left( \left| t_{(20)} \right| \leq \frac{k}{1.165} \right) = 0.95 \quad \text{(1)} \]

The table of t- distribution for 20 d.f., 
\[ P \left( t_{(20)} \leq 2.0860 \right) = 0.95 \quad \text{---- (2)} \]

Comparing (1) and (2), we get, 
\[ \frac{k}{1.165} = 2.086. \text{ Hence} \]
\[ k = 2.086 \times 1.165 = 2.431 \]

That is the maximum difference that can expect with 95% probability is 2.431.

**1.4. Snedecor’s F-distribution**

A continuous random variable \( F \) with pdf
\[
f(F) = \frac{\left( \frac{n_1}{n_2} \right)^{\frac{n_1}{2}} \frac{1}{F^{\frac{n_2}{2}-1}}}{\beta \left( \frac{n_1}{2}, \frac{n_2}{2} \right)} \left( \frac{1 + \frac{n_1}{n_2} F}{2} \right)^{-\frac{n_1+n_2}{2}}, \quad 0 \leq F < \infty
\]
is said to follow F-distribution with \( (n_1, n_2) \) degrees of freedom.

If \( X_1 \) and \( X_2 \) are independent random variables following chi-square distribution with \( n_1 \) and \( n_2 \) degrees of freedom respectively, then,
\[
F = \frac{X_1}{X_2} \sim F(n_1, n_2)
\]

**Statistic following Snedecor’s F-distribution:**

1. Let independent samples of sizes \( n_1 \) and \( n_2 \) are taken from normal population with mean \( \mu \) and standard deviation \( \sigma \). Let \( S_1^2 \) and \( S_2^2 \) are the respective sample variance, then 
\[
F = \frac{n_1 S_1^2 (n_2 - 1)}{n_2 S_2^2 (n_1 - 1)} \sim F(n_1 - 1, n_2 - 1)
\]
Proof:

For the set of samples taken from normal population, we have,
\[ \frac{n_1 S_1^2}{\sigma^2} \sim \chi^2_{(n_1-1)} \quad \text{and} \quad \frac{n_2 S_2^2}{\sigma^2} \sim \chi^2_{(n_2-1)} \]

Then,
\[ F = \frac{n_1 S_1^2}{n_2 S_2^2} \sim F(n_1 - 1, n_2 - 1) \]

Hence
\[ F = \frac{n_1 S_1^2 (n_2 - 1)}{n_2 S_2^2 (n_1 - 1)} \sim F(n_1 - 1, n_2 - 1) \]

Tables of F-distribution:

Tables of F-distribution gives the values of \( F_\alpha \) for various values of \( n_1, n_2 \) and \( \alpha \), such that \( P(F_{n_1,n_2} > F_\alpha) = \alpha \).

Mode of F-distribution:

Mode is the point F where \( f(F) \) attains its maximum. That is the point F, where \( f'(F) = 0 \) and \( f''(F) < 0 \) or the F where \( \frac{\partial \log f(F)}{\partial F} = 0 \) and \( \frac{\partial^2 \log f(F)}{\partial F^2} < 0 \).

We have,
\[ f(F) = \frac{\left( \frac{n_1}{n_2} \right)^{\frac{n_1}{2}} \frac{n_1}{2}^{n_1 - 1}}{\beta \left( \frac{n_1}{2}, \frac{n_2}{2} \right) \left( 1 + \frac{n_1}{n_2} F \right)^{\frac{n_1 + n_2}{2}}} \]

Therefore,
\[ \log f(F) = \frac{n_1}{2} \log \left( \frac{n_1}{n_2} \right) + \left( \frac{n_1}{2} - 1 \right) \log F - \log \beta \left( \frac{n_1}{2}, \frac{n_2}{2} \right) - \frac{n_1 + n_2}{2} \log \left( 1 + \frac{n_1}{n_2} F \right) \]
\[ \frac{\partial \log f(F)}{\partial F} = \left( \frac{n_1}{2} - 1 \right) \frac{1}{F} - \frac{n_1 + n_2}{2} \frac{1}{\left( 1 + \frac{n_1}{n_2} F \right)} \]

\[ \Rightarrow \frac{\partial \log f(F)}{\partial F} = 0 \Rightarrow \left( \frac{n_1}{2} - 1 \right) \frac{1}{F} = \frac{(n_1 + n_2) n_1}{2} \frac{1}{(n_2 + n_1 F)} \]

\[ \Rightarrow F = \frac{n_2 (n_1 - 2)}{n_1 (n_2 + 2)} \] at this point it can be verified that \( \frac{\partial^2 \log f(F)}{\partial F^2} < 0 \).

Hence mode of \( F \sim F(n_1, n_2) \) is \( F = \frac{n_2 (n_1 - 2)}{n_1 (n_2 + 2)} \).

Remark: The mode \( \frac{n_2 (n_1 - 2)}{n_1 (n_2 + 2)} \) can be expressed as \( \frac{n_2}{(n_2 + 2)} \times \frac{(n_1 - 2)}{n_1} \). Since \( F > 0 \), the mode cannot be negative. Hence \( n_1 \) should not be less than 2. So the mode exists only when \( n_1 > 2 \). Again since \( \frac{n_2}{(n_2 + 2)} \) and \( \frac{(n_1 - 2)}{n_1} \) are less than 1, the mode is always less than unity.

Problem 1: Prove that the ratio of the squares of two independent standard normal random variables is an \( F \) random variable with \( (1, 1) \) degree of freedom.

Solution:

Let \( X_1 \sim N(0,1) \), \( X_2 \sim N(0,1) \) and they are independent.

So, \( X_1^2 \sim \chi^2_{(1)} \) and \( X_2^2 \sim \chi^2_{(1)} \)

then \( \frac{X_1^2 / 1}{X_2^2 / 1} \sim F(1,1) \)

\[ ie., \frac{X_1^2}{X_2^2} \sim F(1,1) \]
Problem 2: If \( X \) is a random variable following \( F \) distribution with \( (n_1, n_2) \) degrees of freedom.

Find the distribution of \( Y = \frac{1}{X} \).

Solution:

Given \( Y = \frac{1}{X} \) we have \( f(y) = f(x) \) in terms of \( y \).

Here

\[
f(x) = \frac{\left( \frac{n_1}{n_2} \right)^{\frac{n_1}{2}} \left( \frac{n_2}{x} \right)^{\frac{n_2}{2} - 1}}{\beta(n_1/2, n_2/2) \left( 1 + \frac{n_1}{n_2} x \right)^{\frac{n_1+n_2}{2}}} \quad 0 \leq x < \infty
\]

\( Y = \frac{1}{X} \), so

\[
X = \frac{1}{Y} \Rightarrow \frac{dx}{dy} = \frac{1}{y^2}
\]

\[
so, \quad f(y) = \frac{\left( \frac{n_1}{n_2} \right)^{\frac{n_1}{2}} \left( \frac{1}{y} \right)^{\frac{n_1}{2} - 1}}{\beta(n_1/2, n_2/2) \left( 1 + \frac{n_1}{n_2} y \right)^{\frac{n_1+n_2}{2}}} \times \frac{1}{y^2}
\]

\[
\Rightarrow f(y) = \frac{\left( \frac{n_1}{n_2} \right)^{\frac{n_1}{2}} \left( \frac{1}{y} \right)^{\frac{n_1}{2} - 1}}{\beta(n_1/2, n_2/2) \left( \frac{n_2 y + n_1}{n_2 y} \right)^{\frac{n_1+n_2}{2}}} \times \left[ \frac{1}{y} \right]^2
\]

\[
= \frac{\left( \frac{n_1}{n_2} \right)^{\frac{n_1}{2}} \left( \frac{1}{y} \right)^{\frac{n_1}{2} + 1}}{\beta(n_1/2, n_2/2) \left( \frac{y + n_1}{n_2} \right)^{\frac{n_1+n_2}{2}}} \times \left[ \frac{1}{y} \right]^{\frac{n_1+n_2}{2}}
\]

\[
= \frac{\left( \frac{n_1}{n_2} \right)^{\frac{n_1}{2}} \left( \frac{1}{y} \right)^{\frac{n_1}{2} + 1 - \frac{n_1}{2} - \frac{n_2}{2}}}{\beta(n_1/2, n_2/2) \left( \frac{n_2 y + n_1}{n_2} \right)^{\frac{n_1+n_2}{2}}} \times \left[ \frac{1}{y} \right]^{\frac{n_1+n_2}{2}}
\]

\[
= \frac{\left( \frac{n_1}{n_2} \right)^{\frac{n_1}{2}} \left( \frac{1}{y} \right)^{\frac{n_1}{2} + 1 - \frac{n_1}{2} - \frac{n_2}{2}}}{\beta(n_1/2, n_2/2) \left( \frac{y + n_1}{n_1} \right)^{\frac{n_1+n_2}{2}}} \times \left[ \frac{1}{y} \right]^{\frac{n_1+n_2}{2}}
\]
\[
\left( \frac{n_1}{n_2} \right)^{\frac{n_1}{2}} \frac{1}{y^{\left( \frac{n_2}{2} \right)}} \beta(\frac{n_1}{2}, \frac{n_2}{2}) \left( \frac{n_2}{n_1} y \right)^{\frac{n_1+n_2}{2}}
\]

\[f(y) = \frac{\left( \frac{n_2}{n_1} \right)^{\frac{n_2}{2}} y^{\left( \frac{n_2}{2} - 1 \right)}}{\beta(\frac{n_1}{2}, \frac{n_2}{2}) \left( 1 + \frac{n_2}{n_1} y \right)^{\frac{n_1+n_2}{2}}} \]

\[\Rightarrow Y \sim F(n_2, n_1)\]

**Problem 3:** If \( X \) following F distribution with \( (n_1, n_2) \) degrees of freedom \( Y \) follow F distribution with \( (n_2, n_1) \) degrees of freedom. Prove that \( P(X \geq c) = P(Y \leq \frac{1}{c}) \).

**Solution:**

\[P(X \geq c) = P\left(\frac{1}{X} \leq \frac{1}{c}\right)\]

But, given, \( X \sim F(n_1, n_2) \), then, \( \frac{1}{X} \sim F(n_2, n_1) \)

Also \( Y \) is a variable following \( F(n_2, n_1) \).

Hence, \[P(X \geq c) = P(Y \leq \frac{1}{c})\]

**Problem 4:** If \( X \) following F distribution with \( (n, n) \) degrees of freedom. If \( \alpha, \beta \ (\alpha < \beta) \) are such that \( P(X < \alpha) = P(X > \beta) \). Show that \( \alpha, \beta = 1 \)

**Solution:**

Since \( X \sim F(n, n) \), \( \frac{1}{X} \) also \( \sim F(n, n) \) ----(1)

\[P(X < \alpha) = P\left(\frac{1}{X} > \frac{1}{\alpha}\right)\]

Given \( P(X < \alpha) = P(X > \beta) \)

\[\Rightarrow P\left(\frac{1}{X} > \frac{1}{\alpha}\right) = P(X > \beta)\]
\[ P(X > \frac{1}{\alpha}) = P(X > \beta) \quad \text{(by (1))} \]

\[ \Rightarrow \frac{1}{\alpha} = \beta \quad \Rightarrow \alpha\beta = 1 \]

**Problem 5:** If \( t \) follows student’s \( t \)-distribution with \( n \) degrees of freedom, prove that \( t^2 \) follows \( F \) distribution with \( (1, n) \) degrees of freedom.

**Solution:**

Given \( t \sim t_{(n)} \), we have

\[ f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < t < \infty \]

Let \( Y = t^2 ; t = \sqrt{Y} \)

Therefore

\[ \left| \frac{dt}{dy} \right| = \frac{1}{2\sqrt{Y}} \]

Note that \( f(Y) = f(t) \) in terms of \( y \).

Then,

\[ f(Y) = 2 \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}} \left(1 + \frac{Y}{n}\right)^{-\frac{n+1}{2}} \frac{1}{2\sqrt{Y}}, \quad 0 < Y < \infty \]

That is,

\[ f(Y) = 2 \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}} \left(1 + \frac{Y}{n}\right)^{-\frac{n+1}{2}} \frac{1}{2\sqrt{Y}} \]

\[ = 2 \cdot \frac{1}{2} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}} \left(1 + \frac{Y^2}{n}\right)^{-\frac{n+1}{2}} \frac{1}{2\sqrt{Y}} \]

\[ \Rightarrow f(Y) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}} \left(1 + \frac{Y}{n}\right)^{-\frac{n+1}{2}} \frac{1}{2\sqrt{Y}} \]

\[ = \frac{1}{\sqrt{n\beta}} \left(\frac{1}{2}\right)^{\frac{1}{2}} \left(1 + \frac{Y}{n}\right)^{-\frac{n+1}{2}} \quad \therefore \beta(m,n) = \frac{m+n}{|m|n} \]
\[ f(Y) = \frac{\left( \frac{1}{n} \right)^{\frac{1}{2} y^{\frac{1}{2}-1}} \beta \left( \frac{1}{2}, \frac{n}{2} \left(1 + \frac{1}{n} Y \right) \frac{1}{2} \right)^{\frac{n+1}{2}}}{0 < Y < \infty, \quad \text{this is the density} \]

function of a random variable following F-distribution with \((I, n)\) degrees of freedom. Hence \( Y = t^2 \) follows \( F(1, n) \).

**Problem 7:** If \( X \) following \( F \) distribution with \((n_1, n_2)\) degrees of freedom, prove that as \( n_2 \to \infty \), \( Y = n_1 X \) follows chi-square distribution with \( n_1 \) degrees of freedom.

**Solution:**

Given \( Y = n_1 X \),

\[
\frac{dx}{dy} = \frac{1}{n_1};
\]

\[ f(Y) = f(x) \text{ in terms of } y, \quad \frac{dx}{dy} \]

\[ X \sim F(n_1, n_2); \quad f(x) = \frac{\left( \frac{n_1}{n_2} \right)^{\frac{n_2}{2} x^{\frac{n_1}{2}-1}}}{\beta\left( \frac{n_1}{2}, \frac{n_2}{2} \left(1 + \frac{n_1}{n_2} x \right)^{\frac{n_1 n_2}{2}} \right), \quad 0 \leq x < \infty} \]

\[
f(y) = \frac{\left( \frac{n_1}{n_2} \right)^{\frac{n_1}{2} y^{\frac{n_2}{2}-1}}}{\beta\left( \frac{n_1}{2}, \frac{n_2}{2} \left(1 + \frac{n_1}{n_2} y \right) \frac{n_1 n_2}{2} \right) \frac{1}{n_1}} \times \frac{1}{n_1}
\]

\[
= \frac{\left( \frac{n_1 + n_2}{2} \right)^{\frac{n_1}{2} y^{\frac{n_2}{2}-1}}}{\beta\left( \frac{n_1}{2}, \frac{n_2}{2} \left(1 + \frac{y}{n_2} \right) \frac{n_1 n_2}{2} \right) \frac{1}{n_1}} \times \frac{1}{n_1}
\]
As \( n_2 \to \infty \)
\[
\frac{n_1 + n_2}{2} \left( \frac{n_2}{2} \right) \left( \frac{n_2}{2} \right) = \frac{n_2^{\frac{1}{2}}}{2} = \frac{1}{2} \left( \frac{n + k}{n} \right) \to n^k \text{ as } n \to \infty
\]

Note that
\[
\left( 1 + \frac{y}{n_2} \right)^{\frac{n_1 + n_2}{2}} = \left[ \left( 1 + \frac{y}{n_2} \right)^{n_2} \right]^{\frac{1}{2}} \times \left( \frac{n_2}{n_2} \right)^{\frac{n_2}{2}}
\]

Also note that
\[
\lim_{n_2 \to \infty} \left( 1 + \frac{y}{n_2} \right)^{\frac{n_1 + n_2}{2}} = \lim_{n_2 \to \infty} \left[ \left( 1 + \frac{y}{n_2} \right)^{n_2} \right]^{\frac{1}{2}} \times \lim_{n_2 \to \infty} \left( 1 + \frac{y}{n_2} \right)^{\frac{n_2}{2}}
\]
\[
= e^{\frac{\gamma}{2}} \times 1 \quad \therefore \lim_{n_2 \to \infty} \left( 1 + \frac{y}{n_2} \right)^{\frac{n_2}{2}} = e^{\gamma}
\]

Hence, as \( n_2 \to \infty \),
\[
f(y) = \frac{\left( \frac{y}{n_1} \right)^{\frac{n_1}{2} - 1} \left( \frac{1}{n_1} \right)^{\frac{n_1}{2}} \cdot e^{-\frac{\gamma}{2} \times 1}}{\left( \frac{2}{n_1} \right)^\frac{m_1}{2}}
\]
\[
= \frac{\left( \frac{1}{n_1} \right)^{\frac{n_1}{2} - 1} \left( \frac{y}{n_1} \right)^{\frac{n_1}{2} - 1} \cdot e^{-\frac{\gamma}{2} \times 1}}{\left( \frac{2}{n_1} \right)^\frac{m_1}{2}}
\]

Hence,
\[
f(y) = \frac{1}{e^{\frac{\gamma}{2} \cdot \frac{n_1}{2} - 1}} \cdot e^{\frac{\gamma}{2} \cdot \frac{n_1}{2} - 1}, \quad 0 < y < \infty
\]

That is \( Y \) follow chi-square distribution with \( n_1 \) degrees of freedom.
EXERCISES

1. Explain what is meant by sampling distribution. State the relationship between normal and chi-square distribution.

2. Define chi-square distribution with \( n \) degrees of freedom. Derive its mean and variance.

3. State and prove the reproductive property of chi-square distribution.

4. Show that for Students t-distribution with \( n \) degrees of freedom, the mean deviation is given by \( \sqrt{\frac{n-1}{2}} \frac{n}{\pi n^2} \).


6. State the inter-relationship of t, chi-square and F distributions. A random variable \( X \) has F-distribution with \((n, m)\) degrees of freedom. Find the distribution of \( Y = \frac{1}{X} \).

7. Derive Student’s t-distribution and establish its relation with F-distribution.

8. If \( F \) has F-distribution with \((n, m)\) degrees of freedom, prove that as \( n \to \infty \), \( nF \) tends to be distributed as chi-square with \( n \) degrees of freedom.

9. If \( X \) and \( Y \) are independent standard normal variables, find the distribution of \( Z = \frac{X^2}{Y^2} \) and write down its p.d.f.

10. \( X_1, X_2, \) and \( X_3 \) are independent \( N(0,1) \) variables. Find the distribution of

\[
(i) \quad X_1^2 + X_2^2 \\
(ii) \quad \frac{X_2}{X_1} \\
(iii) \quad \frac{X_1^2 + X_2^2}{2X_3^2}
\]

****
CHAPTER 2

THEORY OF ESTIMATION- POINT ESTIMATION

2.1. Statistical Inference

Making inferences about the unknown aspects of the population using the samples drawn from the population is known as statistical inference. The unknown aspects may be the form of the probability distribution of the population or values of the parameters (i.e., function of population values) involved, or both.

Two important subdivision of statistical inference are.

(i) Estimation

(ii) Testing of hypothesis.

Estimation of parameters:

The theory of estimation was founded by Prof. R.A. Fisher, who is known as the father of modern Statistics. Estimation deals with function of sample values, the value of which may be taken as the values of the unknown parameters (known as point estimation) as well as with the determination of the intervals which will contain the unknown parameters with a specified probability (known as interval estimation), based on the samples taken from the population.

Testing of hypothesis deals with the method of deciding whether to accept or reject the hypothesis regarding the unknown aspects of the population, based on the samples taken from the population.

2.2. Point estimation

In point estimation a number is suggested as a value of the unknown parameter, using the values of the sample observations taken randomly from the population. The function of sample values, suggested as a good approximation for the required parameter is known as an estimator, and a particular value of the estimator is known as the estimate. For eg., to estimate population mean, sample mean is taken as an estimator and the value of sample mean of a particular sample is an estimate of population mean.

2.3. Desirable Properties of Good Estimator

An estimator of a parameter is said to be a good estimator if it satisfied some desirable properties. They are

(i) Unbiasedness    (ii) Consistency    (ii) Efficiency    (iv) Sufficiency
(i) Unbiasedness:

Let \( x_1, x_2, ..., x_n \) are random samples taken from a population with unknown parameter \( \theta \). The statistic \( t_n = t(x_1, x_2, ..., x_n) \) is said to be an unbiased estimator of \( \theta \), if \( E(t_n) = \theta \). \( t_n \) is an unbiased estimator of a function of \( \theta \), say \( f(\theta) \), if \( E(t_n) = f(\theta) \).

**Problem 1:** A random sample \( x_1, x_2, ..., x_n \) is taken from a population with mean \( \mu \). Show that the sample mean \( \bar{x} \) is an unbiased estimator of \( \mu \).

**Solution:**

Since the samples are taken from a population with mean \( \mu \),

\[
E(x_1) = E(x_2) = ... = E(x_n) = \mu
\]

we have \( \bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} \)

\[
E(\bar{x}) = E\left(\frac{x_1 + x_2 + ... + x_n}{n}\right) = \frac{1}{n} E(x_1 + x_2 + ... + x_n)
\]

\[
= \frac{1}{n} (\mu + \mu + ... + \mu) = \frac{1}{n} (n\mu)
\]

\[
\Rightarrow E(\bar{x}) = \mu
\]

Hence \( \bar{x} \) is an unbiased estimator of \( \mu \)

**Remark:** Unbiased estimator for a parameter need not be unique. For eg. in the above case consider the first two observations \( x_1 \) and \( x_2 \) only. Then

\[
E\left(\frac{x_1 + x_2}{2}\right) = \frac{1}{2} E(x_1 + x_2) = \frac{1}{2} E(\mu + \mu) = \mu.
\]

That means \( \frac{x_1 + x_2}{2} \) is also an unbiased estimator of \( \mu \). In similar way we can find many unbiased estimators for \( \mu \).

**Problem 2:** A random sample \( x_1, x_2, ..., x_n \) is taken from a normal population with mean \( \mu \) and standard deviation 1. Show that \( t = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \) is an unbiased estimator of \( \mu^2 + 1 \).

**Solution:**

\[
E(t) = E\left(\frac{1}{n} \sum_{i=1}^{n} x_i^2\right)
\]

\[
= \frac{1}{n} \left[ E(x_1^2) + E(x_2^2) + ... + E(x_n^2) \right]
\]

Given the population variance as 1, and population mean as \( \mu \),

\[
\Rightarrow E(x_i) = \mu \text{ and } E(x_i^2) - [E(x_i)]^2 = 1 \text{ for all } i
\]
\[ 1 + [E(x_i)]^2 = E(x_i^2) \text{ for all } i \]

\[ E(x_i^2) = 1 + [\mu]^2 \text{ for all } i \]

Hence, \[ E(t) = \frac{1}{n} \left[ (1 + \mu^2) (1 + \mu^2) + \ldots + (1 + \mu^2) \right] \]

\[ = \frac{1}{n} \times n (1 + \mu^2) \]

\[ \Rightarrow E(t) = (1 + \mu^2), \text{ i.e., } t = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \text{ is an unbiased estimator of } \mu^2 + 1. \]

**Problem 3:** For the random sample \( x_1, x_2, \ldots, x_n \) taken from \( N(\mu, \sigma) \), show that the sample variance is a biased estimator of the population variance.

**Solution:**

Here to show that \( E(S^2) \neq \sigma^2 \), where \( S^2 \) is the sample variance of the random samples \( x_1, x_2, \ldots, x_n \) taken from \( N(\mu, \sigma) \).

Note that \[ E(S^2) = \int_0^\infty S^2 f(S^2) dS^2 \]

\[ = \int_0^\infty S^2 \left( \frac{n}{2\sigma^2} \right)^{\frac{n-1}{2}} e^{-\frac{nS^2}{2\sigma^2}(S^2)} \frac{1}{2} dS^2 \]

\[ = \frac{n}{2\sigma^2} \int_0^\infty e^{-\frac{nS^2}{2\sigma^2}(S^2)} \frac{1}{2} dS^2 \]

\[ = \frac{n}{2\sigma^2} \int_0^\infty e^{-\frac{nS^2}{2\sigma^2}(S^2)} \frac{1}{2} dS^2 \]

\[ = \frac{n}{2\sigma^2} \int_0^\infty e^{-\frac{nS^2}{2\sigma^2}(S^2)} \frac{1}{2} dS^2 \]
\[ \left( \frac{n}{2 \sigma^2} \right)^{n-1} \times \left( \frac{n}{2 \sigma^2} \right)^{n-1} \times \left( \frac{n}{2 \sigma^2} \right)^{n-1} = \left( \frac{n}{2 \sigma^2} \right)^{n-1} \times \left( \frac{n}{2 \sigma^2} \right)^{n-1} \times \left( \frac{n}{2 \sigma^2} \right)^{n-1} \]

\[ = \left( \frac{n}{2 \sigma^2} \right)^{n-1} \times \left( \frac{n}{2 \sigma^2} \right)^{n-1} \times \left( \frac{n}{2 \sigma^2} \right)^{n-1} \]

\[ = \left( \frac{n}{2 \sigma^2} \right)^{n-1} \times \left( \frac{n}{2 \sigma^2} \right)^{n-1} \times \left( \frac{n}{2 \sigma^2} \right)^{n-1} \]

\[ \Rightarrow E(S^2) = \left( \frac{n-1}{n} \right) \times \sigma^2 \]

Hence \( S^2 \) is not an unbiased estimator of \( \sigma^2 \).

That is \( S^2 \) is a biased estimator of \( \sigma^2 \).

Here, \( E(S^2) = \left( \frac{n-1}{n} \right) \times \sigma^2 \)

\[ \Rightarrow E \left( \frac{nS^2}{n-1} \right) = \sigma^2 \]

That is, \( \frac{nS^2}{n-1} \) is an unbiased estimator of \( \sigma^2 \).

**Problem 4:** For the random sample \( x_1, x_2, \ldots, x_n \) taken from Poisson population with parameter \( \lambda \), obtain an unbiased estimate of \( e^{-\lambda} \).

**Solution:**

Consider a statistic \( t \) defined as follows,

\[ t = 1, \text{ if the first observation of the sample if zero} \]
\[ = 0, \text{ otherwise.} \]

\[ P(\text{the first observation of the sample if zero}) = \frac{e^{-\lambda} \lambda^0}{0!} = e^{-\lambda} \]

Hence, \( E(t) = 1 \times e^{-\lambda} + 0(1 - e^{-\lambda}) = e^{-\lambda} \)

This implies the statistic \( t \) is an unbiased estimator of \( e^{-\lambda} \).
Problem 5: For the random sample \( x_1, x_2, \ldots, x_n \) taken from \( B(1, p) \), show that \( \frac{T(T-1)}{n(n-1)} \) is an unbiased estimator \( p^2 \), where \( T = \sum_{i=1}^{n} x_i \).

Solution:

Here \( x_1, x_2, \ldots, x_n \) are from \( B(1, p) \). Then by the additive property,
\[
T = (x_1 + x_2 + \ldots + x_n) \text{ follows } B(n, p)
\]

Note that \( E(T) = np \), \( V(T) = npq \)

\[
E\left[ \frac{T(T-1)}{n(n-1)} \right] = \frac{1}{n(n-1)} E[T(T-1)] = \frac{1}{n(n-1)} E\left[ T^2 - T \right]
\]

\[
E(T^2) = V(T) + [E(T)]^2
\]

\[
= npq + n^2 p^2
\]

\[
\Rightarrow E\left[ \frac{T(T-1)}{n(n-1)} \right] = \frac{1}{n(n-1)} E\left[ npq + n^2 p^2 - np \right]
\]

\[
= \frac{1}{n(n-1)} E\left[ np(1-p) + n^2 p^2 - np \right]
\]

\[
= \frac{1}{n(n-1)} E\left[ n^2 p^2 - np^2 \right]
\]

\[
\Rightarrow E\left[ \frac{T(T-1)}{n(n-1)} \right] = p^2 , \text{ ie., } \frac{T(T-1)}{n(n-1)} \text{ is an unbiased estimator of } p^2 .
\]

( ii ) Consistency:

Let \( x_1, x_2, \ldots, x_n \) are random samples taken from a population with unknown parameter \( \theta \).

The statistic \( t_n = t(x_1, x_2, \ldots, x_n) \) is said to be a consistent estimator of \( \theta \), if \( P \{ |t_n - \theta| < \varepsilon \} \to 1 \) as \( n \to \infty \) or \( t_n \) is a consistent estimator of a function of \( \theta \), say \( f(\theta) \), if \( P \{ |t_n - f(\theta)| < \varepsilon \} \to 1 \) as \( n \to \infty \).

In other words \( t_n \) is a consistent estimator of \( \theta \), if \( t_n \) converges to \( \theta \) in probability, and is denoted as \( t_n \stackrel{p}{\longrightarrow} \theta \).

Sufficient conditions for consistency:

Let \( \{t_n\} \) sequence of estimators of \( \theta \), and if,

(i) \( E(t_n) = \theta \) or \( \theta \), as \( n \to \infty \) and (ii) \( V(t_n) \to 0 \), as \( n \to \infty \)
Then \( t_n \) is a consistent estimator of \( \theta \).

**Proof:**

Consider the statistic \( t_n \), then by Tchebycheff’s inequality,

\[
P\left( |t_n - \theta| < t.SD(t_n) \right) > 1 - \frac{1}{t^2}
\]

Let \( t.SD(t_n) = c \), then \( t = \frac{c}{SD(t_n)} \)

\[
\Rightarrow P\left( |t_n - \theta| < c \right) > 1 - \frac{1}{\left[ \frac{c}{SD(t_n)} \right]^2}
\]

as \( n \to \infty \), if \( E(t_n) = \theta \) or \( \theta \) and \( V(t_n) \to 0 \); then \( t.SD(t_n) = c \) becomes a small number and, \( \frac{c}{SD(t_n)} \to \infty \),

\[
\Rightarrow P\left( |t_n - \theta| < c \right) \to 1 \quad \text{ie.,} \quad t_n \overset{p}{\to} \theta
\]

Hence under the given conditions, \( t_n \) is a consistent estimator of \( \theta \).

**Problem 1:** For the random sample \( x_1, x_2, \ldots, x_n \) taken from \( N(\mu, \sigma) \), show that sample mean \( \bar{x} \) is consistent estimator of population mean \( \mu \).

**Solution:**

\[
\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \Rightarrow E(\bar{x}) = \mu \quad \text{and} \quad V(\bar{x}) = \frac{\sigma^2}{n}
\]

as \( n \to \infty \), \( E(\bar{x}) = \mu \) and \( V(\bar{x}) \to 0 \).

Hence, \( \bar{x} \) is consistent estimator of population mean \( \mu \).

**Problem 2:** For the random sample \( x_1, x_2, \ldots, x_n \) taken from Poisson population with parameter \( \lambda \), show that \( \frac{n \bar{x}}{n + 1} \) is consistent estimator \( \lambda \).

**Solution:**

Here \( x_1, x_2, \ldots, x_n \) are taken from Poisson population with parameter \( \lambda \), so \( E(X_i) = \lambda \) and \( V(X_i) = \lambda \) for all \( i \), then

\[
\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \Rightarrow E(\bar{x}) = \lambda \quad \text{and} \quad V(\bar{x}) = \frac{\lambda}{n}
\]

Hence, as \( n \to \infty \),
$$E\left(\frac{n\bar{x}}{n+1}\right) = \frac{n}{n+1}E(\bar{x})$$

$$= \left(\frac{1}{1+\frac{1}{n}}\right)E(\bar{x}) \rightarrow E(\bar{x}) = \lambda \, \text{, and}$$

$$V\left(\frac{n\bar{x}}{n+1}\right) = \frac{n^2}{(n+1)^2}V(\bar{x}) = \frac{1}{\left(1+\frac{1}{n}\right)^2} \lambda \rightarrow 0$$

Here \(\frac{n\bar{x}}{n+1}\) satisfies the sufficient conditions to be satisfied by consistent estimator and hence it is a consistent estimator of \(\lambda\).

**Problem 3:** For the random sample \(x_1, x_2, \ldots, x_n\) taken from \(B(1, p)\), show that \(T(1-T)\) is a consistent estimator of \(p(1-p)\), where \(T = \frac{1}{n} \sum_{i=1}^{n} x_i\).

**Solution:**

Note that \(x_1, x_2, \ldots, x_n\) are from \(B(1, p)\). Then by the additive property,

\[
X = \left(x_1 + x_2 + \ldots + x_n\right) \text{ follows } B(n, p)
\]

\[T = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \Rightarrow \quad E(T) = E\left(\frac{X}{n}\right) = p \, , \, V(T) = \frac{1}{n^2}V(X) = \frac{pq}{n}\]

\[
E(T^2) = V(T) + \left[E(T)\right]^2
\]

\[\Rightarrow E(T^2) = \frac{pq}{n} + p^2 \quad \text{-------- (1)}\]

\[\Rightarrow E[T(1-T)] = E(T) - E(T^2) = p - \left[\frac{pq}{n} + p^2\right]
\]

\[= \frac{np - p(1-p) - np^2}{n} = \frac{np(1-p) - p(1-p)}{n}
\]

\[= \frac{n-1}{n} p(1-p)
\]

as \(n \rightarrow \infty\), \(E[T(1-T)] \rightarrow p(1-p) \quad \text{-------- (2)}\)

\[
V[T(1-T)] = E\left([T(1-T)]^2\right) - \left(E[T(1-T)]\right)^2
\]
\[
E[T^4] = E \left[ \frac{X^4}{n} \right] = \frac{1}{n^4} E(X^4)
\]
\[
= \frac{1}{n^4} \left[ n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np \right]
\]
as \( n \to \infty \), \( E[T^4] \to p^4 \)

Similarly, \( E[T^3] = \frac{1}{n^3} \left[ n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np \right] \)
as \( n \to \infty \), \( E[T^3] \to p^3 \) and \( E[T^2] \to p^2 \) (by (1))

Then, as \( n \to \infty \), from (3), we get,
\[
V[T(1-T)] \to p^2 + p^4 - 2p^3 - \left[ p - p^2 \right]^2
\]

ie., \( V[T(1-T)] \to 0; \) as \( n \to \infty \) \hspace{2cm} (4)

Hence, by (2) and (4), \( T(1-T) \) is a consistent estimator of \( p(1-p) \)

**Problem 4:** For the random sample \( x_1, x_2, ..., x_n \) taken from \( N(\mu, \sigma) \), show that the sample variance is a consistent estimator of the population variance.

**Solution:**

Let \( S^2 \) is the sample variance of the random samples \( x_1, x_2, ..., x_n \) taken from \( N(\mu, \sigma) \).

Then we have.
\[
f(S^2) = \left( \frac{n}{2\sigma^2} \right)^{\frac{n-1}{2}} \frac{1}{\sqrt{\Gamma(\frac{n-1}{2})}} e^{-\frac{nS^2}{2\sigma^2}(S^2)^{\frac{n-1}{2}}} \left[ 0 < S^2 < \infty \right]
\]

It is already found in a problem of last section,
\[
E(S^2) = \left( \frac{n-1}{n} \right) \times \sigma^2
\]

Then, as \( n \to \infty \), \( E(S^2) \to \sigma^2 \) \hspace{2cm} (1)

\[
E[S^2]^2 = \int_0^\infty S^2 f(S^2) dS^2 = \int_0^\infty \left( \frac{n}{2\sigma^2} \right)^{\frac{n-1}{2}} \frac{1}{\sqrt{\Gamma(\frac{n-1}{2})}} e^{-\frac{nS^2}{2\sigma^2}(S^2)^{\frac{n-1}{2}}} dS^2
\]
\[
\frac{n}{2\sigma^2} \frac{n-1}{2} \int_0^{\frac{n}{2\sigma^2}} e^{-\frac{n}{2\sigma^2} (S^2)^{\frac{n+1}{2}}} dS^2
\]

\[
= \frac{n}{2\sigma^2} \frac{n-1}{2} \int_0^{\frac{n}{2\sigma^2}} e^{-\frac{n}{2\sigma^2} (S^2)^{\frac{n+3}{2}}} dS^2
\]

\[
= \frac{n}{2\sigma^2} \frac{n-1}{2} \times \frac{\frac{n+3}{2}}{2} \left( \frac{n}{2\sigma^2} \right)^{\frac{n+3}{2}}
\]

\[
= \frac{n^2-1}{n^2} \times \sigma^4
\]

Hence, \( V(S^2) = E \left[ S^2 \right]^2 - \left( E \left[ S^2 \right] \right)^2 = \frac{n^2-1}{n^2} \times \sigma^4 - \left( \frac{n-1}{n} \times \sigma^2 \right)^2 \)

as \( n \to \infty \), \( V(S^2) \to \sigma^4 - \sigma^4 = 0 \)

------- (2)

From (1) and (2), it can infer that \( S^2 \) is a consistent estimator of the population variance \( \sigma^2 \)

**Problem 5:** Let \( t \) be a consistent estimator of \( \theta \), and let \( \psi(\theta) \) be a continuous function of \( \theta \). Then prove that \( \psi(t) \) is a consistent estimator of \( \psi(\theta) \).

**Solution:**

Since \( t \) is consistent for \( \theta \), \( P( | t - \theta | < \epsilon ) \to 1 \) as \( n \) becomes large.

If \( \psi \) is a continuous function, we have for \( \epsilon \) such that, \( | t - \theta | < \epsilon \),

\[
| \psi(t) - \psi(\theta) | < \epsilon_1
\]

\( \Rightarrow P( | \psi(t) - \psi(\theta) | < \epsilon_1 ) \to 1 \)
(iii) Efficiency:

As we already seen, to estimate a particular population parameter there may exist more than one unbiased estimators. Based on the variance of these unbiased estimators their efficiency is defined. Consider $t_1$ and $t_2$ are two unbiased estimators of the parameter $\theta$. The estimator $t_1$ is said to be more efficient than $t_2$, if $\text{Var}(t_1) < \text{Var}(t_2)$. Let $t$ be the most efficient estimator for the parameter $\theta$, then efficiency of any other unbiased estimator $t_i$ of $\theta$ is defined as $E(t_i) = \frac{\text{var}(t_i)}{\text{var}(t)}$. The efficiency of most efficient estimator is 1 and any other unbiased estimator is less than 1.

The relative efficiency of $t_i$ with respect to $t_2$ is denoted by $E$ and is defined as $E = \frac{\text{var}(t_2)}{\text{var}(t_i)}$.

**Problem 1:** For the random sample $x_1, x_2, ..., x_n$ taken from $N(\mu, \sigma)$, test whether the following statistics are unbiased estimators of $\mu$. Which one is more efficient?

(i) $\frac{x_1 + x_2}{2}$  
(ii) $\frac{x_1 + 2x_2}{3}$  
(iii) $\frac{x_1 + x_2 + x_3 + x_4}{4}$

**Solution:**

(i) $E\left(\frac{x_1 + x_2}{2}\right) = \frac{1}{2}E(x_1 + x_2) = \frac{2\mu}{2} = \mu$, so $\frac{x_1 + x_2}{2}$ is an unbiased estimator of $\mu$.

(ii) $E\left(\frac{x_1 + 2x_2}{3}\right) = \frac{1}{3}E(x_1 + 2x_2) = \frac{1}{3}(\mu + 2\mu) = \mu$, so $\frac{x_1 + 2x_2}{3}$ is an unbiased estimator of $\mu$.

(iii) $E\left(\frac{x_1 + x_2 + x_3 + x_4}{4}\right) = \frac{1}{4}E(x_1 + x_2 + x_3 + x_4) = \frac{4\mu}{4} = \mu$, so $\frac{x_1 + x_2 + x_3 + x_4}{4}$ also is an unbiased estimator of $\mu$.

$$V\left(\frac{x_1 + x_2}{2}\right) = \frac{1}{4}V(x_1 + x_2) = \frac{2\sigma^2}{4} = \frac{\sigma^2}{2} \quad (\because x_i's \text{ are random samples})$$

$$V\left(\frac{x_1 + 2x_2}{3}\right) = \frac{1}{9}V(x_1 + 2x_2) = \frac{1}{9}(\sigma^2 + 4\sigma^2) = \frac{5\sigma^2}{9}$$

$$V\left(\frac{x_1 + x_2 + x_3 + x_4}{4}\right) = \frac{1}{16}V(x_1 + x_2 + x_3 + x_4) = \frac{4\sigma^2}{16} = \frac{\sigma^2}{4}$$

Among these $V\left(\frac{x_1 + x_2 + x_3 + x_4}{4}\right) < V\left(\frac{x_1 + x_2}{2}\right) < V\left(\frac{x_1 + 2x_2}{3}\right)$

Hence $\frac{x_1 + x_2 + x_3 + x_4}{4}$ is more efficient.
(iv) Sufficiency:

In many problems of statistical inference, a function of the sample observations contains as much information about the unknown parameter as do all observed values. To estimate probability of head \( p \) when a coin is tossed, let the coin is tossed \( n \) times and let \( x_i = 1 \), if the \( i^{th} \) toss is a success and \( =0 \), otherwise. Then, \( t = \sum_i x_i \) - the total number of heads out of \( n \) tosses is enough to estimate \( p \). It seems unnecessary to know which toss resulted in a head. That is \( t = \sum_i x_i \) is sufficient to estimate the parameter \( p \). The result of \( n \) tosses \( x_1, x_2, \ldots, x_n \) contains no other information about \( p \) than that contains in \( t \). Hence the conditional probability of \( x_1, x_2, \ldots, x_n \) given \( t = \sum_i x_i \) is independent of \( p \).

That is \( P(x_1, x_2, \ldots, x_n / \sum_i x_i = t) = \frac{p^{\sum_i x_i} (1-p)^{n-\sum_i x_i}}{C \sum_i x_i p^{\sum_i x_i} (1-p)^{n-\sum_i x_i}} = \frac{1}{C_i} \).

Hence a statistic \( t \) is said to be sufficient for the parameter \( \theta \), if it contains all information about the parameter contained in the sample, or

If the conditional distribution of any other statistic given \( t = r \), is independent of \( \theta \).

Fisher-Neymaan Factorization Theorem (Condition for Sufficiency):

Let \( x_1, x_2, \ldots, x_n \) be a random sample from a population with pmf/pdf \( f(x, \theta) \) then the joint pmf/pdf of the sample (usually called the likelihood of the sample) is \( L(x_1, x_2, \ldots, x_n, \theta) = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta) \), the statistic \( t \) is a sufficient estimator of \( \theta \), if and only if it is possible to write

\[
L(x_1, x_2, \ldots, x_n, \theta) = L(t, \theta) \times L_2(x_1, x_2, \ldots, x_n)
\]

where \( L_2(t, \theta) \) is function of \( t \) and \( \theta \) alone and \( L_2(x_1, x_2, \ldots, x_n) \) is a function independent of \( \theta \).

Proof:

If \( t \) is a sufficient estimator of \( \theta \), then the conditional distribution of \( x_1, x_2, \ldots, x_n \) given \( t = r \) is independent of \( \theta \). That is,
\[
P(x_1, x_2, \ldots, x_n / t = r) = h(x_1, x_2, \ldots, x_n), \text{ which is independent of } \theta \quad \text{------ (1)}
\]

But \( P(x_1, x_2, \ldots, x_n / t = r) = \frac{P(x_1, x_2, \ldots, x_n)}{P(t = r)} = \frac{L(x_1, x_2, \ldots, x_n, \theta)}{P(t, \theta)} \)

Then by (1), for sufficient estimator \( t \),
\[
L(x_1, x_2, \ldots, x_n, \theta) = h(x_1, x_2, \ldots, x_n) \times P(t, \theta)
\]
Problem 1: For a Poisson distribution with parameter $\lambda$, show that sample mean $\bar{x}$ is the sufficient estimator of $\lambda$.

Solution:

Let $x_1, x_2, ..., x_n$ be the random sample taken from $P(\lambda)$

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

$$L(x_1, x_2, ..., x_n) = f(x_1, \lambda) f(x_2, \lambda) ... f(x_n, \lambda)$$

$$= e^{-\lambda} \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \frac{\lambda^{x_2}}{x_2!} ... e^{-\lambda} \frac{\lambda^{x_n}}{x_n!}$$

$$= \frac{e^{-n\lambda} \lambda^{\sum x_i}}{x_1! x_2! ... x_n!}$$

$$= e^{-n\lambda} \frac{\lambda^{\sum x_i}}{x_1! x_2! ... x_n!}$$

$$= e^{-n\lambda} \frac{\lambda^{n\bar{x}}}{x_1! x_2! ... x_n!}$$

$$= L_4(\bar{x}, \lambda) \times L_2(x_1, x_2, ..., x_n)$$

where $L_4(\bar{x}, \lambda) = e^{-n\lambda} \lambda^{n\bar{x}}$ and $L_2(x_1, x_2, ..., x_n) = \frac{1}{x_1! x_2! ... x_n!}$

Hence, by factorization theorem, $\bar{x}$ is a sufficient estimator of $\lambda$.

Problem 2: Let $x_1, x_2, ..., x_n$ be the random sample taken from a population with p.d.f. $f(x, \theta) = \theta x^{\theta-1}, 0 < x < 1, \theta > 0$. Find a sufficient estimator for $\theta$.

Solution:

Likelihood function $L(x_1, x_2, ..., x_n, \theta) = f(x_1, \theta) f(x_2, \theta) ... f(x_n, \theta)$

$$= \theta x_1^{\theta-1} \theta x_2^{\theta-1} ... \theta x_n^{\theta-1}$$

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

$$= L_4 \left( \prod_{i=1}^{n} x_i, \theta \right) \times L_2(x_1, x_2, ..., x_n)$$

where $L_4 \left( \prod_{i=1}^{n} x_i, \theta \right) = \theta^n \left( \prod_{i=1}^{n} x_i \right)^{\theta}$ and $L_2(x_1, x_2, ..., x_n) = \frac{1}{\prod_{i=1}^{n} x_i}$, then by
factorization theorem \( t = \prod_{i=1}^{n} x_i \) can be considered as a sufficient estimator of \( \theta \).

**Problem 3:** Obtain a sufficient estimator for \( p \), using samples \( x_1, x_2, \ldots, x_n \) taken from \( B(n, p) \).

**Solution:**

The likelihood function of \( x_1, x_2, \ldots, x_n \),

\[
L(x_1, x_2, \ldots, x_n, \theta) = f(x_1, p) f(x_2, p) \cdots f(x_n, p)
\]

\[
= \prod_{i=1}^{n} C_{x_i} p^x_i (1 - p)^{n-x_i} = \prod_{i=1}^{n} C_{x_i} p^{x_i} (1 - p)^{n-x_i}
\]

\[
= \sum_{i=1}^{n} (1 - p)^{x_i} \prod_{i=1}^{n} C_{x_i} = \sum_{i=1}^{n} C_{x_i} p^{x_i} \prod_{i=1}^{n} C_{x_i}
\]

\[
L(\overline{x}, p) = \prod_{i=1}^{n} C_{x_i} p^{x_i} \prod_{i=1}^{n} C_{x_i}
\]

where \( L(\overline{x}, p) = p^{n\overline{x}} (1 - p)^{n-n\overline{x}} \), and \( L_2(x_1, x_2, \ldots, x_n) = C_{x_1} C_{x_2} \cdots C_{x_n} \), then by factorization theorem \( \overline{x} \) is a sufficient estimator of \( p \).

**Problem 4:** If \( t \) is sufficient estimator for \( \theta \), prove that \( \frac{\partial \log L}{\partial \theta} \) is a function of \( t \) and \( \theta \) only.

**Solution:**

If \( t \) is a sufficient estimator of \( \theta \), then the likelihood function

\[
L(x_1, x_2, \ldots, x_n, \theta) = L_1(t, \theta) \times L_2(x_1, x_2, \ldots, x_n)
\]

\[
\therefore \log L(x_1, x_2, \ldots, x_n, \theta) = \log L_1(t, \theta) + \log L_2(x_1, x_2, \ldots, x_n)
\]

\[
\Rightarrow \frac{\partial}{\partial \theta} \log L(x_1, x_2, \ldots, x_n, \theta) = \frac{\partial}{\partial \theta} \log L_1(t, \theta) + 0
\]

Since, \( L_1(t, \theta) \) is a function of \( t \) and \( \theta \) only, \( \frac{\partial}{\partial \theta} \log L \) is also a function of \( t \) and \( \theta \) only.

**Problem 5:** If \( t \) is sufficient estimator for \( \theta \), then prove that any 1-1 function of \( t \) is also sufficient for \( \theta \).

**Solution:**

Let \( h = g(t) \), assume \( h \) is a 1-1 function of \( t \), then \( t = g^{-1}(h) \)

Since \( t \) is sufficient for \( \theta \),

\[
L(x_1, x_2, \ldots, x_n, \theta) = L_1(t, \theta) \times L_2(x_1, x_2, \ldots, x_n)
\]

\[
\Rightarrow L(x_1, x_2, \ldots, x_n, \theta) = L_1(g^{-1}(h), \theta) \times L_2(x_1, x_2, \ldots, x_n)
\]

Hence \( L(x_1, x_2, \ldots, x_n, \theta) \) is the product of a function of \( h \) and \( \theta \) only and a function independent of \( \theta \). Then by factorization theorem \( h \) is also sufficient for \( \theta \).
2.4. Method of Estimation

(i) Maximum Likelihood Estimator:

Let \( x_1, x_2, \ldots, x_n \) be the sample taken from the population with p.m.f/p.d.f \( f(x, \theta_1, \theta_2, \ldots, \theta_k) \), where \( \theta_1, \theta_2, \ldots, \theta_k \) are the parameters involved. The likelihood function of the sample \( L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_k) = f(x_1, \theta_1, \theta_2, \ldots, \theta_k) f(x_2, \theta_1, \theta_2, \ldots, \theta_k) \cdots f(x_n, \theta_1, \theta_2, \ldots, \theta_k) \).

The method of maximum likelihood suggests, the best estimators for estimating the parameters \( \theta_1, \theta_2, \ldots, \theta_k \) are the estimators which maximizes the likelihood function. Such estimators are known as Maximum Likelihood Estimators (M.L.E) of \( \theta_1, \theta_2, \ldots, \theta_k \).

The Principle of M.L.E says that the best estimators of the parameters based on a sample obtained are, those values of the parameters which make the probability of getting that sample a maximum.

Using the method of differential calculus, the function of sample values for a parameter which maximizing the likelihood function- called MLE of that parameter, can be obtained. Let \( L(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_k) \) be the likelihood function corresponds to the sample \( x_1, x_2, \ldots, x_n \). The value of \( \theta_1 \), as a function of \( x_1, x_2, \ldots, x_n \), maximizing the likelihood function can be obtained from \( \frac{\partial L}{\partial \theta_1} = 0 \), and if for that value of \( \theta_1 \), \( \frac{\partial^2 L}{\partial \theta_1^2} < 0 \). But since we know the value of \( \theta_1 \), which maximizing the likelihood function also maximizes \( \log L \), such value of \( \theta_1 \) can also be obtained by using \( \frac{\partial \log L}{\partial \theta_1} = 0 \) if for that value of \( \theta_1 \), \( \frac{\partial^2 \log L}{\partial \theta_1^2} \) is less than zero.

Maximum Likelihood Estimators possess some desirable properties of a good estimator.

(i) MLE’s are asymptotically unbiased.

(ii) MLE’s are consistent.

(iii) MLE’s are most efficient.

(iv) MLE’s are sufficient, if sufficient statistics exist.

(v) MLE’s are asymptotically normally distributed.

**Problem 1:** Find the M.L.E. of \( \mu \) and \( \sigma \), using the random sample \( x_1, x_2, \ldots, x_n \) taken from the normal population \( N(\mu, \sigma) \)

**Solution:**

Given the random sample \( x_1, x_2, \ldots, x_n \) from \( N(\mu, \sigma) \)

The likelihood function,
\[ L(x_1, x_2, \ldots, x_k, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \times \cdots \times \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \]

\[ = \left( \frac{1}{\sqrt{2\pi\sigma}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2} \]

\[ \log L(x_1, x_2, \ldots, x_k, \mu, \sigma) = n \log \left( \frac{1}{\sqrt{2\pi\sigma}} \right) - \frac{n}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \]

\[ \frac{\partial \log L}{\partial \mu} = 0 \Rightarrow \frac{n}{\sqrt{2\pi\sigma}} \times \frac{1}{\sqrt{2\pi\sigma}} \times \frac{-1}{\sigma^2} \times \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} (-2) = 0 \]

\[ \Rightarrow \sum_{i=1}^{n} x_i - \mu = 0 \Rightarrow \mu = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \]

\[ \frac{\partial^2 \log L}{\partial \mu^2} = \frac{1}{\sigma^2} (-n) < 0 \]

Hence, \( \bar{x} \) is the MLE of \( \mu \)

To obtain the MLE of \( \sigma \),

\[ \frac{\partial \log L}{\partial \sigma} = 0 \Rightarrow n\sqrt{2\pi\sigma} \times \frac{1}{\sqrt{2\pi\sigma}} \times \frac{-1}{\sigma^2} \times \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} (-2) = 0 \]

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

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

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

\[ \frac{\partial^2 \log L}{\partial \sigma^2} = \frac{n}{\sigma^2} \times 3 \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \]

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

\[ \frac{\partial^2 \log L}{\partial \sigma^2} = \frac{n}{\sum_{i=1}^{n} (x_i - \mu)^2} - 3 \frac{n^2}{\sum_{i=1}^{n} (x_i - \mu)^2} \]

\[ = \frac{-2n^2}{\sum_{i=1}^{n} (x_i - \mu)^2} < 0 \]
Then the MLE of $\sigma^2$ is $\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$. Since MLE of $\mu$ is $\bar{x}$, MLE of $\sigma^2$ is considered as $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$, which is the sample variance.

**Problem 2:** Find the MLE of $\lambda$, based on random samples taken from Poisson population with parameter $\lambda$.

**Solution:**

Let $x_1, x_2, ..., x_n$ are the random sample taken from $P(\lambda)$, then

$$L(x_1, x_2, ..., x_n, \lambda) = f(x_1, \lambda) f(x_2, \lambda) ... f(x_n, \lambda)$$

$$= \frac{e^{-\lambda} \lambda^{x_1}}{x_1!} \frac{e^{-\lambda} \lambda^{x_2}}{x_2!} ... \frac{e^{-\lambda} \lambda^{x_n}}{x_n!} = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{x_1! x_2! ... x_n!}$$

$$\log L = - n\lambda + \sum x_i \log \lambda - \log(x_1! x_2! ... x_n!)$$

$$\frac{\partial \log L}{\partial \lambda} = 0 \Rightarrow - n + \sum x_i \frac{1}{\lambda} - (0) = 0$$

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

$$\frac{\partial^2 \log L}{\partial \lambda^2} = - \sum x_i \frac{1}{\lambda^2}$$

At $\lambda = \bar{x}$, $\frac{\partial^2 \log L}{\partial \lambda^2} = - \sum x_i (\bar{x})^{-2} < 0$ (samples $x_i$'s from Poisson population are $\geq 0$ for $P(\lambda)$)

Hence, $\bar{x}$ is the MLE of $\lambda$.

**Problem 3:** Obtain the MLE of $\theta$ for the following distribution $f(x) = \frac{1}{2} e^{-|x-\theta|}$, $-\infty < x < \infty$.

**Solution:**

Let $x_1, x_2, ..., x_n$ be the sample from the given population, then,

$$L(x_1, x_2, ..., x_n, \theta) = f(x_1, \theta) f(x_2, \theta) ... f(x_n, \theta)$$

$$= \frac{1}{2} e^{-|x_1-\theta|} \frac{1}{2} e^{-|x_2-\theta|} ... \frac{1}{2} e^{-|x_n-\theta|}$$

$$= \left(\frac{1}{2}\right)^n e^{-\sum |x_i-\theta|}$$
Therefore, \[ \log L = -n \log 2 - \sum_{i} |x_i - \theta| \]

Here \( \log L \) is maximum when \( \sum_{i} |x_i - \theta| \) is minimum.

This happens when \( \theta \) is the median of the random sample \( x_1, x_2, ..., x_n \). So MLE of \( \theta \) is the median of \( x_1, x_2, ..., x_n \).

**Problem 4:** Obtain the MLE of \( \alpha \) and \( \beta \) using the random samples \( x_1, x_2, ..., x_n \) taken from the population with pdf \( f(x) = \frac{1}{\beta} e^{-\frac{(x-\alpha)}{\beta}}, \ x \geq \alpha, \ \beta > 0 \).

**Solution:**

The likelihood function \( L(x_1, x_2, ..., x_n, \alpha, \beta) \) can be written as

\[
L(x_1, x_2, ..., x_n, \alpha, \beta) = \frac{1}{\beta} e^{-\frac{(x_1-\alpha)}{\beta}} \cdot \frac{1}{\beta} e^{-\frac{(x_2-\alpha)}{\beta}} \cdot ... \cdot \frac{1}{\beta} e^{-\frac{(x_n-\alpha)}{\beta}}
\]

\[
= \left( \frac{1}{\beta} \right)^n e^{-\sum_{i} \frac{(x_i-\alpha)}{\beta}}
\]

\[
\therefore \quad \log L = -n \log \beta - \frac{\sum (x_i-\alpha)}{\beta}
\]

\[
\therefore \quad \frac{\partial \log L}{\partial \alpha} = 0 \quad \Rightarrow \quad 0 + \frac{n}{\beta} = 0 \quad \therefore \quad (1)
\]

\[
\therefore \quad \frac{\partial \log L}{\partial \beta} = 0 \quad \Rightarrow \quad -\frac{n}{\beta} + \frac{1}{\beta^2} \sum_{i} (x_i-\alpha) = 0
\]

\[
\Rightarrow \quad \frac{\sum_{i} (x_i-\alpha)}{\beta} = n
\]

\[
\Rightarrow \quad \beta = \frac{1}{n} \sum_{i} (x_i-\text{Min} \ x_i) \quad \therefore \quad (2)
\]

Equation (1) cannot imply the MLE of \( \alpha \). But we know \( \log L \) is maximized when \( \sum_{i} (x_i-\alpha) \) a minimum is. This happens when \( \alpha \) is a maximum. But \( \alpha \) cannot be greater than \( \text{Min} \ x_i \). Hence \( \text{Min} \ x_i \) is the MLE of \( \alpha \). Then by (2), the value of \( \beta \) can be written as \( \frac{1}{n} \sum_{i} (x_i-\text{Min} \ x_i) \) and it can be verified that at this value of \( \beta \),
\[ \frac{\partial^2 \log L}{\partial \beta^2} < 0 . \] Hence the MLE of \( \beta \) is \( \frac{1}{n} \sum (x_i - \text{Min} x_i) \).

**Problem 5:** Obtain the MLE of \( a \) and \( b \) using the random samples \( x_1, x_2, \ldots, x_n \) taken from a rectangular population over the interval \( (a - b \ , \ a + b) \).

**Solution:**

Here random samples \( x_1, x_2, \ldots, x_n \) taken from a rectangular population over the interval \( (a - b \ , \ a + b) \). Hence \( f(x) \) is given by

\[ f(x) = \frac{1}{(a+b)-(a-b)} = \frac{1}{2b}, \quad a-b < x < a+b \]

In this case the likelihood function \( L(x_1, x_2, \ldots, x_n, a, b) = \left[ \frac{1}{(a+b)-(a-b)} \right]^n \)

The method of differential calculus cannot be applied here. The likelihood function \( L \) is maximum when \( (a+b)-(a-b) \) is minimum. This happens when \( (a+b) \) is taking its minimum and \( (a-b) \) is taking its maximum possible value.

But \( (a+b) \) cannot be less than the largest value of \( x_1, x_2, \ldots, x_n \) and \( (a-b) \) cannot be greater than the smallest value of \( x_1, x_2, \ldots, x_n \).

Hence MLE of \( (a+b) = \text{Max} (x_1, x_2, \ldots, x_n) \) and

\[ \text{MLE of } (a-b) = \text{Min} (x_1, x_2, \ldots, x_n) \]

Then, the MLE of \( a = \frac{\text{Max}(x_1, x_2, \ldots, x_n) + \text{Min}(x_1, x_2, \ldots, x_n)}{2} \) and

the MLE of \( b = \frac{\text{Max}(x_1, x_2, \ldots, x_n) - \text{Min}(x_1, x_2, \ldots, x_n)}{2} \)

**Problem 6:** Obtain the MLE of the parameter \( \theta \) using the random samples \( x_1, x_2, \ldots, x_n \) taken from a population with pdf, \( f(x) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta \).

The likelihood function \( L(x_1, x_2, \ldots, x_n, \theta) = \left[ \frac{1}{\theta} \right]^n = \frac{1}{\theta^n} \)

The likelihood function getting its maximum value when \( \theta^n \) is minimum, i.e., \( \theta \) is minimum. But for the given pdf, \( 0 \leq x \leq \theta \), so, \( \theta \) cannot be less than the highest observation of the sample. Hence the minimum possible value of \( \theta \) is the highest value of the sample. That is the MLE of \( \theta = \text{Max} (x_1, x_2, \ldots, x_n) \)

**Problem 7:** If a sufficient statistics \( T \) exists for \( \theta \), then prove that MLE of \( \theta \) is a function of \( T \).

**Solution:**
If $T$ is sufficient for estimating $\theta$, then,

\[ L(x_1, x_2, \ldots, x_n, \theta) = L_1(T, \theta) \times L_2(x_1, x_2, \ldots, x_n) \]

\[
\therefore \frac{\partial \log L}{\partial \theta} = \frac{\partial}{\partial \theta} \log L_1(T, \theta) + 0 \quad \text{and} \quad \frac{\partial \log L}{\partial \theta} = 0 \Rightarrow \frac{\partial}{\partial \theta} L_1(T, \theta) = 0
\]

This implies $\theta$ is a function of $T$.

**Problem 8:** Given that the frequency function \( f(x, \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \), where $X$ can assume only nonnegative integral values and given the following observed values, 4,5,7,21,24,12,15,7,9,14. Find the M.L.E of $\lambda$.

**Solution:**

Here 10 observations are taken from the given Poisson population. If let $x_1, x_2, \ldots, x_{10}$ are the random sample taken from $P(\lambda)$, then

\[ L(x_1, x_2, \ldots, x_{10}, \lambda) = f(x_1, \lambda). f(x_2, \lambda) \ldots f(x_{10}, \lambda) \]

\[ = \frac{e^{-\lambda} \lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda} \lambda^{x_2}}{x_2!} \cdots \frac{e^{-\lambda} \lambda^{x_{10}}}{x_{10}!} = \frac{e^{-10\lambda} \lambda^{\sum x_i}}{x_1! x_2! \cdots x_{10}!} \]

\[ \log L = -10\lambda + \sum_i x_i \log \lambda - \log(x_1! x_2! \cdots x_{10}!) \]

\[ \frac{\partial \log L}{\partial \lambda} = 0 \Rightarrow -10 + \sum_i x_i \frac{1}{\lambda} -(0) = 0 \]

\[ \Rightarrow \lambda = \frac{1}{10} \sum_i x_i = \bar{x} \]

\[ \frac{\partial^2 \log L}{\partial \lambda^2} = - \sum_i x_i \frac{1}{\lambda^2} \]

at $\lambda = \bar{x}$ ; \[ \frac{\partial^2 \log L}{\partial \lambda^2} = - \sum_i x_i \frac{1}{(\bar{x})^2} < 0 \quad (\because \text{samples } x_i \text{'s from Poisson population are } \geq 0 \text{ for } P(\lambda)) \]

Hence, \[ \frac{1}{10} \sum_i x_i \text{ is the MLE of } \lambda. \]

\[ \frac{1}{10} \sum_i x_i = \frac{4 + 5 + 7 + 21 + 24 + 12 + 15 + 7 + 9 + 14}{10} = 11.8. \]
(ii) Method of moments:

Let \( f(x, \theta_1, \theta_2, \ldots, \theta_n) \) be the pdf of the population and \( x_1, x_2, \ldots, x_n \) be the random sample taken from it. In the method of moment, we find the first \( k \) moments of the population and equate them to the corresponding moments of the sample. The values of \( \theta_1, \theta_2, \ldots, \theta_n \), obtained as a function of \( x_1, x_2, \ldots, x_n \), by solving the equations are considered as the moment estimators of \( \theta_1, \theta_2, \ldots, \theta_n \).

**Problem 1:** For a normal population \( N(\mu, \sigma) \), find the estimators of \( \mu \) and \( \sigma \) by the method of moments.

**Solution:**

Let \( x_1, x_2, \ldots, x_n \) be the random sample taken from \( N(\mu, \sigma) \),

First raw moment of the population \( E(X) = \mu \).  
First raw moment of the sample is \( \bar{x} \).

Equating first moment of the sample and the population we get \( \bar{x} \) as the estimator of \( \mu \).

Second raw moment of the population \( E(X^2) = (E(X))^2 + V(X) = \mu^2 + \sigma^2 \)

Second raw moment of the sample is \( \frac{1}{n} \sum_{i=1}^{n} x_i^2 \)

Equating these two, we get \( \mu^2 + \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \)

\[ \Rightarrow \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \mu^2; \quad \text{But } \bar{x} \text{ is the estimator of } \mu. \]

Hence moment estimator of \( \sigma^2 \) is \( \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \), which is the sample variance.

**Problem 2:** \( X \) is a random variable with probability masses as shown below.

\[
\begin{align*}
X &: \quad 0 \quad 1 \quad 2 \\
f(x) &: (1-\theta - \theta^2) \quad \theta \quad \theta^2; \quad 0 < \theta < 1,
\end{align*}
\]

Find the moment estimate of \( \theta \), if in a 25 observations there were 10 ones and 4 twos.

**Solution:**

Out of 25 samples taken from the population, it is recorded 10 ones, 4 twos and the remaining 9 zeroes.

Then first moment of the sample is \( \frac{9 \times 0 + 10 \times 1 + 4 \times 2}{25} = \frac{18}{25} \).

First moment of the population,
Equating these we get, $\theta + 2\theta^2 = \frac{18}{25}$, i.e., $500^2 + 25\theta - 18 = 0$

Solving this quadratic expression, we get $\theta = 0.295$.

**Problem 3:** Obtain the moment estimate of $\theta$, if the probability masses are $X: 1, 2, 3, 4$; $f(x): \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{1+\theta}{4}, \frac{1+\theta}{4}$; $0 < \theta < 1$, and the observed frequencies are 1, 5, 7 and 7 respectively.

**Solution:**

Out of 20 samples taken from the population, it is recorded 1 one, 5 twos, 7 threes and the remaining 7 fours.

Then first moment of the sample is $\frac{1\times1 + 5\times2 + 7\times3 + 7\times4}{20} = \frac{60}{20} = 3$.

First moment of the population,

$E(X) = 1\times\left(\frac{1-\theta}{4}\right) + 2\times\left(\frac{1-\theta}{4}\right) + 3\times\left(\frac{1+\theta}{4}\right) + 4\times\left(\frac{1+\theta}{4}\right)$

$= \frac{10}{4} + \frac{4\theta}{4}$

Equating these we get, $\frac{10}{4} + \frac{4\theta}{4} = 3$,

solving this quadratic expression, we get $\theta = 0.5$.

**EXERCISES**

1. $X_1, X_2, X_3$ are random samples from population with mean $\mu$ and standard deviation $\sigma$. $T_1, T_2, T_3$ are defined as $T_1 = X_1 + X_2 - X_3$; $T_2 = 2X_1 + 3X_3 - 4X_2$; and $T_3 = \frac{1}{9}(\lambda X_1 + X_2 + X_3)$. Are (i) $T_1$ and $T_2$ are unbiased estimator? (ii) Find $\lambda$, such that $T_3$ is an unbiased estimator of $\mu$ (iii) Which is the most efficient estimator?

2. Define consistent estimator. Obtain the sufficient conditions for consistency.
3. For Poisson distribution with parameter \( \lambda \), show that \( \frac{1}{x} \) is a consistent estimator of \( \frac{1}{\lambda} \).

4. Let \( x_1, x_2, ..., x_n \) are random sample taken from \( N(\mu, \sigma) \). Find sufficient estimators of \( \mu \) and \( \sigma \).

5. Define sufficient statistic. Find a sufficient statistic when \( f(x, \theta) = \frac{1}{\theta}, \ 0 < x < \theta \)

6. Find the MLE of \( p \), based on sample taken from Binomial distribution with parameters \( N \) and \( p \).

7. Obtain the MLE of \( \theta \) in \( f(x, \theta) = (1+\theta)x^\theta, \ 0 < x < 1 \) based on random sample \( x_1, x_2, ..., x_n \) taken from the population. Also verify whether the MLE is a sufficient estimator of \( \theta \).

8. Find the MLE of \( \theta \), where the random samples \( x_1, x_2, ..., x_n \) are taken from the population with pdf \( f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \ 0 < x < \infty \)

9. An urn contains white and black balls in unknown proportions, the total number of balls being 12. Four balls are drawn at random, of which 3 are found to be white and 1 black. Find the maximum likelihood estimate of the number of white balls in the urn.

10. Explain the method of maximum likelihood estimation. Find M.L.E. of \( \theta \), when

\[
f(x, \theta) = 1, \quad \text{if} \quad \theta - \frac{1}{2} < x < \theta + \frac{1}{2} \\
= 0, \quad \text{otherwise}
\]

11. Find the moment estimator of \( \theta \), if \( f(x, \theta) = \frac{1}{\theta e^{\theta}} x^{\theta-1} e^{-\frac{x}{\theta}}, \ 0 < x < \infty ; \ p \) is known based on the random samples \( x_1, x_2, ..., x_n \) are taken from the population.

12. Explain the method of moments. Using this, obtain the estimators of the parameters \( a \) and \( b \) of a uniform distribution over the interval \([a,b]\)

13. Find an estimator of \( \lambda \), based on random samples taken from Poisson population with parameter \( \lambda \) by the method of moments.

14. Obtain the moment estimate of \( \theta \), if the probability masses are

\[
X: \quad 1 \quad 2 \quad 3 \quad 4 \\
f(x): \quad \frac{1-\theta}{4} \quad \frac{1-\theta}{4} \quad \frac{1+\theta}{4} \quad \frac{1+\theta}{4} ; \quad 0 < \theta < 1, \quad \text{and the observed frequencies are 1,5,7 and 7 respectively.}
\]

***********************

STATISTICAL INference
3.1. Interval Estimation

In point estimation we are finding an estimator to estimate the unknown parameter under the expectation that the true value of the parameter is very close to the value of the estimator suggested. We consider the value of the estimator as the value of the parameter. For eg., in case of $N(\mu, \sigma)$, $\bar{x}$ is suggested as an estimator of $\mu$. But when we are dealing with interval estimation, we are estimating an interval where the value of the unknown parameter lying with a pre-assigned probability.

That is in case of interval estimation for a parameter $\theta$, it is estimating two statistics $t_1$ and $t_2$ ($t_1 < t_2$) such that the probability that the interval $(t_1, t_2)$ contains the true value of the unknown parameter $\theta$ with a specified probability $(1-\alpha); 0<\alpha<1$. Then $(t_1, t_2)$ is termed as $100(1-\alpha)%$ confidence interval for the parameter $\theta$, and $(1-\alpha)$ is the confidence coefficient. It is to be noted that there may be many confidence interval for a particular parameter $\theta$ with same confidence coefficient. Shortness, stability etc., are some desirable property to identify a good interval.

3.2. Confidence interval for the mean of a normal population with confidence coefficient $(1-\alpha)$:

**Case I: When $\sigma$ is known**

Let $x_1, x_2, ..., x_n$ be the sample taken from $N(\mu, \sigma)$ and let the sample mean be $\bar{x}$. We use $\bar{x}$ - the point estimator of $\mu$ for its interval estimation. The mean $\bar{x}$ follows $N(\mu, \frac{\sigma}{\sqrt{n}})$,

or $t = \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma} \sim N(0,1)$.

From standard normal table it can observe the value $t_{\alpha/2}$ such that,
\[ P(|t| < \frac{t_{\alpha}}{\sqrt{2}}) = 1 - \alpha \]

\[ \Rightarrow P\left(\frac{|\bar{x} - \mu|\sqrt{n}}{\sigma} < \frac{t_{\alpha}}{\sqrt{2}}\right) = 1 - \alpha \]

\[ \Rightarrow P\left(-\frac{t_{\alpha}}{\sqrt{2}} < \frac{|\bar{x} - \mu|\sqrt{n}}{\sigma} < \frac{t_{\alpha}}{\sqrt{2}}\right) = 1 - \alpha \]

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

Multiplying by -1 \[ \Rightarrow P\left(\bar{x} + \frac{\sigma}{\sqrt{n}} \frac{t_{\alpha}}{\sqrt{2}} > \mu > \bar{x} - \frac{\sigma}{\sqrt{n}} \frac{t_{\alpha}}{\sqrt{2}}\right) = 1 - \alpha \]

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

Hence the confidence interval for \( \mu \) with confidence coefficient \( (1-\alpha) \) is,

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

**Case II: When \( \sigma \) is unknown**

Let \( x_1, x_2, \ldots, x_n \) be the sample taken from \( N(\mu, \sigma) \) and let the sample mean be \( \bar{x} \). It is already derived for random samples taken from \( N(\mu, \sigma) \),

\[ t = \frac{(\bar{x} - \mu)\sqrt{n-1}}{s} \sim t_{n-1}. \]

From the table of t-distribution, it can be observed a number \( t_{\alpha/2} \) such that,

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

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

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

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

Hence the confidence interval for \( \mu \) with confidence coefficient \( (1 - \alpha) \) is,

\[ \left[ \frac{x - t_{\alpha/2} \sqrt{\frac{s}{n-1}}}{s}, \frac{x + t_{\alpha/2} \sqrt{\frac{s}{n-1}}}{s} \right]. \]

**Problem 1:** Estimate a 95% confidence interval for \( \mu \), based on 10 random samples 17,21,20,18,19,22,20,21,16,19 taken from \( N(\mu, 3) \)

**Solution:**

Here \( \sigma \) is known. Hence the 100(1 - \( \alpha \))% confidence interval for \( \mu \) is,

\[ \left[ \frac{x - t_{\alpha/2} \sigma}{\sqrt{n}}, \frac{x + t_{\alpha/2} \sigma}{\sqrt{n}} \right]. \]

We have, \( \bar{x} = \frac{17 + 21 + 20 + 18 + 19 + 22 + 20 + 21 + 16 + 19}{10} = 19.3 \)
\[ \sigma = 3; \text{ Given } (1 - \alpha) = 0.95, \]

Hence, from standard normal table we get, \( t_{\frac{\alpha}{2}} = 1.96 \),

so that, \[ P(\left| t \right| < t_{\frac{\alpha}{2}}) = 0.95. \]

Hence the confidence interval is,

\[ \left[ 19.3 - 1.96 \frac{3}{\sqrt{10}}, \ 19.3 + 1.96 \frac{3}{\sqrt{10}} \right] = [17.44, \ 21.16] \]

**Problem 2:** Find the least sample size required if the length of 95% confidence interval for the mean of a normal population with standard deviation 4 should be less than 5.

**Solution:**

Let \( n \) random samples are taken from the population \( N(\mu, 4) \). The confidence interval for the mean \( \mu \) is,

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

Given the confidence coefficient is 95 %. Hence from the standard normal table \( t_{\frac{\alpha}{2}} = 1.96 \).

To find the minimum number of samples such that, the length of the interval of \( \mu \),

\[ \left( \bar{x} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) - \left( \bar{x} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = t_{\frac{\alpha}{2}} \frac{2\sigma}{\sqrt{n}} < 5 \]
\[ \Rightarrow 1.96 \frac{2 \times 4}{\sqrt{n}} < 5 \]
\[ \Rightarrow 1.96 \frac{2 \times 4}{5} < \sqrt{n} \]
\[ \Rightarrow n > \left(1.96 \frac{2 \times 4}{5}\right)^2 \Rightarrow n > 9.83 \]

Hence the least number of samples required is 10

**Problem 3:** A sample of size 17 taken from \( N(\mu, \sigma) \). Mean of the sample is 12 and the sample variance is 4. Using the data, find a 90% confidence interval for \( \mu \).

**Solution:**

Here the value of \( \sigma \) is unknown. Then the confidence interval for \( \mu \) is,

\[
\left[ \bar{x} - t_{\alpha} \frac{s}{\sqrt{n-1}}, \bar{x} + t_{\alpha} \frac{s}{\sqrt{n-1}} \right]
\]

Since the confidence coefficient is 90%, from the table of t distribution for 16 d.f., we get \( t_{\alpha/2} = 1.746 \), so as \( P\left(\left|t_{16}\right| < t_{\alpha/2}\right) = 0.90 \).

Also given \( \bar{x} = 12 \) and \( s = \sqrt{4} = 2 \)

Then the 90% confidence interval for \( \mu \) is,

\[
\left[ 12 - 1.746 \frac{2}{\sqrt{16}}, 12 + 1.746 \frac{2}{\sqrt{16}} \right] = [11.127, 12.873]
\]

**Problem 4:** For a \( N(\mu, 3) \) population, construct a 95% confidence interval for \( 3\mu + 5 \), on the basis of the random sample of size 25. The sample mean was found to be 30.

**Solution:**

Consider \( Y = 3X + 5 \); then \( Y \) follows \( N(3\mu + 5, \sqrt{3^2 \times \sigma^2}) = N(3\mu + 5, \sqrt{9}) \)
Sample mean corresponding to \( Y \), \( \bar{y} \sim N(3\mu + 5, \frac{9}{\sqrt{n}}) \)

Here \( n=25 \) and, \( \bar{y} = 3 \times 30 + 5 = 95 \) \((\because \bar{x} = 30)\)

\[
\frac{u = \frac{\bar{y} - (3\mu + 5)}{9}}{\sqrt{25}} \sim N(0,1)
\]

Given \((1-\alpha) = 0.95\), hence from standard normal table , \( t_{a/2} = 1.96 \), so as

\[
P(|t| < t_{a/2}) = 0.95. \text{ Hence, } P\left(\frac{|\bar{y} - (3\mu + 5)|}{\frac{9}{\sqrt{25}}} < 1.96\right) = 0.95
\]

This implies 95% confidence interval for \( 3\mu + 5 \) as

\[
\left[ \bar{y} - 1.96 \times \frac{9}{\sqrt{25}}, \bar{y} + 1.96 \times \frac{9}{\sqrt{25}} \right]
\]

Since \( \bar{y} = 95 \), the interval is

\[
\left[ 95 - 1.96 \times \frac{9}{\sqrt{25}}, 95 + 1.96 \times \frac{9}{\sqrt{25}} \right] = [91.472, 98.528]
\]

**Problem 5:** Show that the length of the confidence interval for the mean of a normal population with known variance can be made however small we please by increasing the sample size.

**Solution:**

The confidence interval for the mean of a normal population when \( \sigma^2 \) is known is given by

\[
\left[ \bar{x} - t_{a/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + t_{a/2} \frac{\sigma}{\sqrt{n}} \right].
\]

The length of the interval = \( \left( \bar{x} + t_{a/2} \frac{\sigma}{\sqrt{n}} \right) - \left( \bar{x} - t_{a/2} \frac{\sigma}{\sqrt{n}} \right) \)

\[= 2 \times t_{a/2} \frac{\sigma}{\sqrt{n}}, \]
Here as the number of sample \( n \) increases the length of the interval decreases and which can be adjusted to any given number for a given significance level by selecting suitable number of samples.

### 3.3. Confidence interval for the difference of means of two normal populations having known common variance \( \sigma^2 \):

Let \( n_1 \) and \( n_2 \) are the number of samples independently drawn from to normal populations \( N(\mu_1, \sigma) \) and \( N(\mu_2, \sigma) \) respectively. Let \( \bar{x}_1 \) and \( \bar{x}_2 \) be the means and \( S_1 \) and \( S_2 \) be the standard deviations of the samples drawn from the first and second population respectively.

\[
\bar{x}_1 \sim N\left( \mu_1, \frac{\sigma}{\sqrt{n_1}} \right) \quad \text{and} \quad \bar{x}_2 \sim N\left( \mu_2, \frac{\sigma}{\sqrt{n_2}} \right)
\]

Then, \( \bar{x}_1 - \bar{x}_2 \sim N\left( \mu_1 - \mu_2, \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \right) \)

\[
\Rightarrow t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)
\]

From standard normal table it can observe the value \( t_{\frac{\alpha}{2}} \) such that,

\[
P(|t| < t_{\frac{\alpha}{2}}) = 1 - \alpha
\]

\[
\Rightarrow P\left( \left| \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| < t_{\frac{\alpha}{2}} \right) = 1 - \alpha
\]

\[
\Rightarrow P\left( -t_{\frac{\alpha}{2}} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < (\mu_1 - \mu_2) < t_{\frac{\alpha}{2}} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) = 1 - \alpha
\]

\[
\Rightarrow P\left( \frac{\bar{x}_1 - \bar{x}_2}{t_{\frac{\alpha}{2}} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > (\mu_1 - \mu_2) > \frac{\bar{x}_1 - \bar{x}_2}{-t_{\frac{\alpha}{2}} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right) = 1 - \alpha
\]
Hence the 100(1 - \( \alpha \)) % confidence interval for \((\mu_1 - \mu_2)\) is,

\[
\left[ \bar{x}_1 - \bar{x}_2 - t_{\alpha} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_1}}, \ \bar{x}_1 - \bar{x}_2 + t_{\alpha} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_1}} \right]
\]

**Problem 1:** The average mark scored by 32 boys in an examination is 72 with a standard deviation of 8, while that scored by 32 girls is 70 with a standard deviation of 6. Construct a 99 % confidence interval for the difference of means. (Assume S.D’s are equal)

**Solution:**

The confidence interval for difference in mean,

\[
\left[ (\bar{x}_1 - \bar{x}_2) - t_{\alpha} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_1}}, \ (\bar{x}_1 - \bar{x}_2) + t_{\alpha} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_1}} \right]
\]

The common value for variance \(\sigma = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}}\) (since we have large samples)

\[
= \sqrt{\frac{(32)8^2 + (32)6^2}{32 + 32}} = 7.07.
\]

\(\bar{x}_1 = 72; \ \sigma_1 = 8; \ n_1 = 32\) and \(\bar{x}_2 = 70; \ \sigma_2 = 6; \ n_2 = 32\)

Confidence coefficient is 99%. From standard normal table \(t_{\alpha/2} = 2.57\)

Hence the confidence interval is,

\[
\left[ (72 - 70) - 2.57 \times 7.07 \sqrt{\frac{1}{32} + \frac{1}{32}}, \ (72 - 70) + 2.57 \times 7.07 \sqrt{\frac{1}{32} + \frac{1}{32}} \right]
\]

\[
= [2 - 4.543, 2 + 4.543]
\]

\[
= [-2.543, 6.543].
\]
3.4. Confidence interval for the variance of a normal population:

Let \( x_1, x_2, \ldots, x_n \) be the sample taken from \( N(\mu, \sigma) \) with sample variance \( S^2 \). Then,

\[ \chi^2 = \frac{nS^2}{\sigma^2} \]

follow chi-square distribution with \((n-1)\) d.f.

From the table of \( \chi^2 \)-distribution, identify the numbers \( \chi^2_{\frac{\alpha}{2}} \) and \( \chi^2_{1-\frac{\alpha}{2}} \) so as,

\[ P\left( \chi^2_{(n-1)} > \chi^2_{\frac{\alpha}{2}} \right) = \frac{\alpha}{2} \quad \text{and} \quad P\left( \chi^2_{(n-1)} > \chi^2_{1-\frac{\alpha}{2}} \right) = 1-\frac{\alpha}{2} \]

respectively.

\[ \Rightarrow P\left( \chi^2_{1-\frac{\alpha}{2}} < \chi^2 < \chi^2_{\frac{\alpha}{2}} \right) = 1-\alpha \]

ie., \[ P\left( \chi^2_{1-\frac{\alpha}{2}} < \frac{nS^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}} \right) = 1-\alpha \quad \Rightarrow \quad P\left( \frac{\chi^2_{1-\frac{\alpha}{2}}}{nS^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{\frac{\alpha}{2}}}{nS^2} \right) = 1-\alpha \]

\[ \Rightarrow P\left( \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2}}} > \sigma^2 > \frac{nS^2}{\chi^2_{\frac{\alpha}{2}}} \right) = 1-\alpha \]

\[ \Rightarrow P\left( \frac{nS^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2}}} \right) = 1-\alpha \]

Hence the confidence interval for \( \sigma^2 \) with confidence coefficient \((1-\alpha)\) is,
Problem 1: A sample of size 12 taken from $N(\mu, \sigma)$. Mean of the sample is 10 and the sample variance is 9. Find a 90% confidence interval for $\sigma^2$.

Solution:

Given $n = 12$, $S^2 = 9$

Confidence interval for $\sigma^2$ is given by

$$\left( \frac{nS^2}{\chi^2_{\frac{\alpha}{2}}} \cdot \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2}}} \right)$$

For 90% confidence interval, ie., for $\alpha = 0.10$, from table of chi-square distribution for 11 d.f,

$$P(\chi^2_{11} > 4.58) = 1 - \frac{0.10}{2} = 0.95 \quad \text{and} \quad P(\chi^2_{11} > 19.68) = \frac{0.10}{2} = 0.05$$

$$\Rightarrow \chi^2_{\frac{\alpha}{2}} = 4.58 \quad \text{and} \quad \chi^2_{1-\frac{\alpha}{2}} = 19.68$$

Hence the 90% confidence interval for $\sigma^2$ is,

$$\left( \frac{12 \times 9}{19.68} \cdot \frac{12 \times 9}{4.58} \right)$$

$$= [ 5.49, 23.58 ]$$

Problem 2: An optical firm purchases glass for making lenses. Assume that the refractive index of 20 pieces of glass have variance of $1.20 \times 10^{-4}$. Construct a 95% confidence interval for the population variance.

Solution:

Confidence interval for $\sigma^2$ with confidence coefficient $(1 - \alpha)$ is,

$$\left( \frac{nS^2}{\chi^2_{\frac{\alpha}{2}}} \cdot \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2}}} \right)$$
Here 20 samples are drawn and the sample variance is $1.20 \times 10^{-4}$. For confidence coefficient 95%, and for $(20 - 1) = 19$ d.f. , chi-square table implies,

$$\chi^2_{\frac{\alpha}{2}} = 32.8523 \text{ and } \chi^2_{1-\frac{\alpha}{2}} = 8.9066.$$  

Hence the 95% confidence interval for $\sigma^2$ is,

$$\left( \frac{20 \times 1.20 \times 10^{-4}}{32.8523} \times \frac{20 \times 1.20 \times 10^{-4}}{8.9066} \right)$$  

$$= \left[ 0.7305 \times 10^{-4}, 2.6946 \times 10^{-4} \right].$$

**Problem 3:** Construct a 95% confidence interval for the variance $\sigma^2$ of the normal population with unknown mean using the following sample:

4.5, 10.2, 10.5, 9.8, 13.0, 19.2, 15.5, 13.3, 10.8 and 16.4

**Solution:**

Given $n = 10$ samples from the normal distribution.

$(1-\alpha)\%$ Confidence interval for $\sigma^2$ is given by

$$\left( \frac{nS^2}{\chi^2_{\frac{\alpha}{2}}} , \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2}}} \right)$$

For $(1-\alpha)\% = 95\%$, $\alpha = 0.05$; from table of chi-square distribution for $n-1 = 9$, d.f,

$$P(\chi^2_g > 2.7004) = 1 - \frac{0.05}{2} = 0.975 \text{ and } P(\chi^2_g > 19.0228) = \frac{0.05}{2} = 0.025$$

$$\Rightarrow \chi^2_{\frac{\alpha}{2}} = 2.7004 \text{ and } \chi^2_{1-\frac{\alpha}{2}} = 19.0228$$

To find the sample variance of the population, using the given 10 samples

The calculations follow:
\[ x = \frac{123.2}{10} = 12.32 \]

\[ s = \sqrt{\frac{1}{10} \sum x_i^2 - (x \bar{x})^2} = \sqrt{\frac{1}{10} (1670.96) - (12.32)^2} = 3.9 \]

Hence the 90% confidence interval for \( \sigma^2 \) is,
\[ \left( \frac{10 \times (3.9)^2}{19.0228}, \frac{10 \times (3.9)^2}{2.7004} \right) \]
\[ = [7.996, 56.325]. \]
3.5. Confidence interval for large samples:

I. \(100(1-\alpha)\%\) Confidence interval for the proportion of a binomial population:

Consider a binomial population with parameters \(N\) and \(p\). Assume \(N\) is known. Repeat the Bernoulli trial for \(n\) number of times. Then the binomial variable \(X\), follow \(B(n,p)\). When \(n\) becomes very large, \(X\) follows normal distribution \(N(np,\sqrt{npq})\).

Then, \(t = \frac{X-np}{\sqrt{npq}} \sim N(0,1)\)

\[
\Rightarrow t = \frac{X - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)
\]

As an approximation,

\[
\Rightarrow t = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1) ; \quad q = 1 - p
\]

(where, \(\bar{p} = \frac{X}{n}\) is the sample proportion)

Hence from standard normal table, obtain \(t_{\frac{\alpha}{2}}\) such that,

\[
P(|t| < t_{\frac{\alpha}{2}}) = 1 - \alpha
\]

\[
\Rightarrow P\left(|\frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}}| < t_{\frac{\alpha}{2}}\right) = 1 - \alpha
\]

\[
\Rightarrow P\left(-t_{\frac{\alpha}{2}} < \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}} < t_{\frac{\alpha}{2}}\right) = 1 - \alpha
\]
\[ P(-\bar{p} - t_{\alpha/2} \sqrt{\frac{\bar{p} \bar{q}}{n}} < -p < -\bar{p} + t_{\alpha/2} \sqrt{\frac{\bar{p} \bar{q}}{n}}) = 1 - \alpha \]

\[ P(\bar{p} + t_{\alpha/2} \sqrt{\frac{\bar{p} \bar{q}}{n}} > p > \bar{p} - t_{\alpha/2} \sqrt{\frac{\bar{p} \bar{q}}{n}}) = 1 - \alpha \]

\[ \Rightarrow P(\bar{p} - t_{\alpha/2} \sqrt{\frac{\bar{p} \bar{q}}{n}} < p < \bar{p} + t_{\alpha/2} \sqrt{\frac{\bar{p} \bar{q}}{n}}) = 1 - \alpha \]

Hence the 100(1−α)% confidence interval for \( p \) for large \( n \) is,

\[ \left( \bar{p} - t_{\alpha/2} \sqrt{\frac{\bar{p} \bar{q}}{n}}, \bar{p} + t_{\alpha/2} \sqrt{\frac{\bar{p} \bar{q}}{n}} \right) \]

where, \( \bar{q} = 1 - \bar{p} \)

**Problem 1:** Random samples of 120 workers of a factory 40 are dissatisfied with their working conditions. Form a 95% confidence interval for the proportion of dissatisfied workers of the factory.

**Solution:**

Given the sample proportion of dissatisfied workers \( \bar{p} = \frac{40}{120} = \frac{1}{3} \). Also given the confidence coefficient = 95%

The confidence interval for proportion of dissatisfied workers is given by,

\[ \left( \bar{p} - t_{\alpha/2} \sqrt{\frac{\bar{p} \bar{q}}{n}}, \bar{p} + t_{\alpha/2} \sqrt{\frac{\bar{p} \bar{q}}{n}} \right) \]

For 95% confidence coefficient, from table of standard normal distribution, \( t_{\alpha/2} = 1.96 \)

Hence the confidence interval is,

\[ \left( \frac{1}{3} - 1.96 \sqrt{\frac{\frac{1}{3} \times \frac{2}{3}}{120}}, \frac{1}{3} + 1.96 \sqrt{\frac{\frac{1}{3} \times \frac{2}{3}}{120}} \right) = [0.214, 0.4526] \]

**Problem 2:** Each computer chip produced by a certain manufacturer is either acceptable or unacceptable. A large batch of such chips produced and it is supposed that each chip in this batch will be independently acceptable with some unknown probability \( p \). To obtain a 99% confidence
interval for \( p \), which is to be of length approximately 0.05, a sample of size 30 is initially taken. If 24 of the 30 chips are deemed acceptable, find the approximate sample size.

\((n=1492)\)

**Solution:**

The confidence interval for proportion of mortality rate is given by,

\[
\left( \bar{p} - t_{\alpha} \sqrt{\frac{\bar{p} \cdot (1 - \bar{p})}{n}} , \bar{p} + t_{\alpha} \sqrt{\frac{\bar{p} \cdot (1 - \bar{p})}{n}} \right)
\]

The given confidence coefficient is 99%. From standard normal table \( t_{\alpha} = 2.57 \)

From the 30 samples taken, \( \bar{p} = \frac{24}{30} = \frac{4}{5} \).

The length of the confidence interval = \( 2t_{\alpha} \sqrt{\frac{\bar{p} \cdot (1 - \bar{p})}{n}} \).

To find the approximate sample size so as length of 99% confidence interval

\[
2t_{\alpha} \sqrt{\frac{\bar{p} \cdot (1 - \bar{p})}{n}} = 0.05 .
\]

That is to get \( n \), such that

\[
2 \times 2.57 \sqrt{\frac{\frac{4}{5} \left(1 - \frac{4}{5}\right)}{n}} = 0.05
\]

\[
\Rightarrow n = \left( \frac{2 \times 2.57}{0.05} \right)^2 \times \frac{4}{5} \left(1 - \frac{4}{5}\right) = 1690.85 \approx 1691
\]

2. **100(1–\( \alpha \))% Confidence interval for \( \lambda \) in Poisson population:**

Consider \( n \) random samples from Poisson population and assume \( n \) is large. As \( n \) become very large, the Poisson random variable \( X \) follows normal distribution \( N(\lambda, \sqrt{\lambda}) \).

The sample mean \( \bar{x} \sim N(\lambda, \frac{\lambda}{n}) \) for large \( n \).

\[
\Rightarrow t = \frac{\bar{x} - \lambda}{\sqrt{\frac{\lambda}{n}}} \sim N(0,1), \text{ approximately (} \because \text{ } E(\bar{x})=\lambda \)
From standard normal table, obtain $t_{\alpha/2}$ such that,

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

$$\Rightarrow P(-t_{\alpha/2} \sqrt{\frac{\bar{x}}{n}} < \xi - \lambda < t_{\alpha/2} \sqrt{\frac{\bar{x}}{n}}) = 1 - \alpha$$

$$\Rightarrow P(-\bar{x} - t_{\alpha/2} \sqrt{\frac{\bar{x}}{n}} < -\lambda < -\bar{x} + t_{\alpha/2} \sqrt{\frac{\bar{x}}{n}}) = 1 - \alpha$$

$$\Rightarrow P(\bar{x} - t_{\alpha/2} \sqrt{\frac{\bar{x}}{n}} < \lambda < \bar{x} + t_{\alpha/2} \sqrt{\frac{\bar{x}}{n}}) = 1 - \alpha$$

Hence the confidence interval with confidence coefficient $(1 - \alpha)$ is,

$$\left(\bar{x} - t_{\alpha/2} \sqrt{\frac{\bar{x}}{n}}, \bar{x} + t_{\alpha/2} \sqrt{\frac{\bar{x}}{n}}\right)$$
EXERCISES

1. The mean of a sample of size 30 drawn from a normal with mean $\mu$ and variance 36 was recorded as 45.6. Find a 95% confidence interval for $\mu$.

2. Construct a $100(1-\alpha)$% confidence interval for the difference of means of two normal populations having common variance.

3. Obtain $100(1-\alpha)$% confidence interval for variance $\sigma^2$ of a normal population $N(\mu, \sigma)$, where the mean $\mu$ is known.

4. Derive 95% confidence interval for the mean of a normal population when (i) variance is known (ii) variance is unknown.

5. Obtain 95% confidence interval for the parameter of Poisson distribution on the basis of a random sample of size $n$.

6. A random sample of size 20 is drawn from a normal population $N(\mu, \sigma)$. Sample mean and the sample variance are respectively 22 and 16. Find a 90% confidence interval for $\mu$.

7. The mean of a sample of size 24 drawn from $N(\mu, \sigma)$ is 25.5. The sample variance is 14. Construct a 95% confidence interval for $\sigma^2$.

8. Using a random sample of size 40 drawn from a Poisson population, construct a 90% confidence interval for the parameter $\lambda$.

9. In a random sample of 40 articles 40 are found to be defective. Obtain 95% confidence interval for the true proportion of defectives in the population of articles.

10. Obtain a large sample $100(1-\alpha)$% confidence interval for the parameter $\theta$, in random sampling from the population with pdf $f(x) = \theta e^{-\theta x}$, $x > 0, \theta > 0$.

*********************
4.1. Statistical Hypothesis:

Statistical inference mainly deals with estimation and testing of hypothesis. As we already seen, the area of estimation consists of making an estimate of an appropriate value of the unknown parameter or an interval for the unknown value of the parameter with a specified probability.

A statistical hypothesis is some statement or assertion about the population parameters or about the form of probability distribution or the population. Theory on testing of hypothesis was initiated by J.Neymaan and E.S.Pearson.

Simple and composite hypothesis:

A statistical hypothesis which completely specifies the population, i.e., it specifies the values of all parameters involved in the probability distribution of the population, then it is called simple hypothesis, otherwise it is called composite.

For eg., assume \( x_1, x_2, \ldots, x_n \) are \( n \) random samples taken from a normal population \( N(\mu, \sigma) \). Then the hypothesis \( H : \mu = \mu_1, \sigma = \sigma_1 \) is a simple hypothesis. But the hypotheses (i) \( H : \mu = \mu_0 \); (ii) \( H : \sigma = \sigma_1 \); (iii) \( H : \mu = \mu_0, \sigma > \sigma_1 \) (iv) \( H : \mu \geq \mu_0, \sigma = \sigma_1 \) (v) \( H : \mu \geq \mu_0, \sigma > \sigma_1 \) etc., are not specifying on the exact values of all the parameters involved, hence they are considered as composite hypothesis.

4.2. Testing of Hypothesis:

Testing of hypothesis is a decision making whether to accept or reject the proposed hypothesis about the population based on the random samples drawn from the population.
Consider the situation where a light bulb manufacturing firm introducing a new type of light bulb. They are claiming that the new product is superior to the existing standard type in terms of the life length. Suppose to test the claim of the firm. Assume it is known that the average life length of the bulb of standard type as 500 hrs. Here we propose a hypothesis regarding the average life length \( \mu \), of the new type of light bulb as \( H : \mu = 500 \). This hypothesis is to be tested, against the alternatives, (i) \( \mu \) is greater than 500 or (ii) \( \mu \) is less than 500 or (iii) either \( \mu > 500 \) or \( \mu < 500 \); ie., \( \mu \neq 500 \).

In a statistical testing of hypothesis, the hypothesis is to be tested is termed as **null hypothesis**. The null hypothesis is denoted by \( H_0 \). Here the null hypothesis is \( H_0 : \mu = 500 \).

Another hypothesis in our mind which we will accept or reject according as we reject or accept \( H_0 \) is termed as **alternative hypothesis**, denoted by \( H_1 \). In the above illustration, \( H_1 \) can be (i) \( H_1 : \mu > 500 \) or (ii) \( H_1 : \mu < 500 \) or (iii) \( H_1 : \mu \neq 500 \).

In this situation we are taking a random sample of \( n \) bulbs of new type and find the average life length of the sample item. Based on the sample mean (which is a good estimator of population mean) we decide whether to accept or reject the null hypothesis. Roughly speaking, if the alternate hypothesis considered is \( H_1 : \mu > 500 \), and the sample mean is much higher than 500, the hypothesis \( H_0 : \mu = 500 \) is rejected. If \( H_1 : \mu < 500 \), and the sample mean is much lesser than 500, \( H_0 \) is rejected and if \( H_1 : \mu \neq 500 \), and the sample mean is reasonably distant from 500, \( H_0 \) is rejected.

Here we are making decision based on the sample mean, because we had to make a decision on population mean and sample mean is a good statistic to say something about the population mean. As this, in any statistical test we have to find an appropriate statistic to make decision based on its value. The value of statistic can be calculated by the value of the samples selected. Such a statistic used in testing of hypothesis is termed as **test statistic**.

In a statistical test, according to the alternative hypothesis selected, we divide the range of variation of the test statistic in to two. One is acceptance region and the other is
rejection region. If the value of the test statistic is in the rejection region, \( H_0 \) is rejected. The rejection region is known as \textit{critical region}.

4.3. Errors in Testing of Hypothesis:

In a hypothesis testing procedure, since decision is making is based on the sample drawn from the population, there exists possibilities for two types of errors. They are termed as Type-I error and Type-II error.

\textit{Type-I error}: It is an error due to rejecting the null hypothesis \( H_0 \), when \( H_0 \) is true.

\textit{Type-II error}: It is an error due to accepting the null hypothesis \( H_0 \), when \( H_0 \) is false.

The possible errors can be tabulated as follows:

<table>
<thead>
<tr>
<th>Action taken Based on the sample</th>
<th>( H_0 ) is true</th>
<th>( H_0 ) is false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject ( H_0 )</td>
<td>TYPE-I ERROR</td>
<td>NO ERROR</td>
</tr>
<tr>
<td>Accept ( H_0 )</td>
<td>NO ERROR</td>
<td>TYPE-II ERROR</td>
</tr>
</tbody>
</table>

Using better statistical criteria, the possible errors in testing procedure can be minimized.

4.4. Steps in Testing of Hypothesis:

(i) Define the population and formulate the hypothesis.

(ii) Choose an appropriate test statistic.

(iii) Divide the range of variation of the test statistics into two regions, acceptance region and rejection region, considering some probabilistic restrictions.

(iv) Take a random sample from the given population and calculated the value of the test statistic and decide whether to accept or reject the hypothesis.
In a testing procedure, the null hypothesis is rejected when the value of the test statistic falls in the pre-decided rejection region. Since the value of the test statistic is decided by the sample drawn, there is a chance for the value of the test statistic to fall in the rejection region even though the null hypothesis is true (type-I-error). The probability of the value of the test statistic to fall in the rejection region, even though the null hypothesis is true is known as **significance level or size of the test**, denoted by $\alpha$.

i.e.,  
**Significance level**  
$\alpha = P(\text{Rej. } H_0 / H_0 \text{ is true}) = P(\text{Type-I-error})$

The probability of the value of the test statistic to fall in the rejection region, when the alternative hypothesis is true is known as **power of the test**, denoted by $\beta$.

i.e.,  
**Power of the test**  
$\beta = P(\text{Rej. } H_0 / H_1 \text{ is true}) = 1 - P(\text{Acc. } H_0 / H_1 \text{ is true})$

$\Rightarrow \beta = 1 - P(\text{Acc. } H_0 / H_0 \text{ is false}) = 1 - P(\text{Type-II error})$

**Problem 1**: To test $H_0 : \theta = 1$ against the alternative $H_1 : \theta = 2$, based on a random sample of size one taken from the population with pdf

$$f(x,\theta) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta.$$  
Find the size and power of the tests if the critical regions are (i) $x > 0.5$ (ii) $1 < x < 1.5$

**Solution**:  
Let $x$ be the random sample drawn from the population.

Given $H_0 : \theta = 1$ and $H_1 : \theta = 2$

(i) When the test is with critical region $x > 0.5$,

Size of the test $\alpha = P(\text{Rej } H_0 / H_0 \text{ is true})$

$= P(x > 0.5 / \theta = 1)$

$= \int_{0.5}^{\theta} \frac{1}{\theta} d\theta / \theta = 1$
= \int_{0.5}^{1} dx

= [x]_{0.5}^{1} = 1 - 0.5 = 0.5

Power of the test \((1 - \beta) = P(\text{Rej.} H_0 / H_1 \text{ is true})\)

= P(x > 0.5/\theta = 2)

= \int_{0.5}^{2} \frac{1}{\theta} dx/\theta = 2 = \int_{0.5}^{2} \frac{1}{2} dx

= \frac{1}{2} [x^2]_{0.5}^{2} = 0.75

(ii) When the test is with critical region \(1 < x < 1.5,\)

Size of the test \(\alpha = P(\text{Rej} H_0/ H_0 \text{ is true})\)

= P (1 < x < 1.5 / \theta = 1)

= \int_{1}^{1.5} \frac{1}{\theta} dx/\theta = 1, \text{for the given population} \ 0 \leq x \leq \theta

. Hence when \(\theta = 1,\) the critical region not exists.

\[\Rightarrow \alpha = 0.\]

Power of the test \((1 - \beta) = P(\text{Rej.} H_0 / H_1 \text{ is true})\)

= P (1 < x < 1.5 / \theta = 2)

= \int_{1}^{1.5} \frac{1}{\theta} dx/\theta = 2 = \int_{1}^{2} \frac{1}{2} dx = 0.25

**Problem 2:** In a coin tossing experiment, let \(p\) be the probability of getting a head. The coin is tossed 10 times to test the hypothesis \(H_0: p = 0.5\) against the alternative \(H_1: p = 0.7\). Reject \(H_0\), if 6 or more tosses out of 10 result in head. Find significance level and power of the test.
Solution:

To test, $H_0 : p = 0.5$ against $H_1 : p = 0.7$. 10 tosses of the coin is considered and let $X$ denote total number of heads obtained. Then $X$ follows binomial distribution $B(10,p)$. The critical region is $X \geq 6$.

Significance level $\alpha = P \left( \text{Type-I error} \right) = P \left( \text{Rej } H_0 / H_0 \text{ true} \right) = P \left( X \geq 6 / p = 0.5 \right) = \sum_{x=6}^{10} C_{10} p^x q^{10-x} / p = 0.5 = \sum_{x=6}^{10} C_{10} 0.5^x (1 - 0.5)^{10-x} = \frac{386}{2^{10}}$

Power of the test $(1 - \beta) = P \left( \text{Rej } H_0 / H_1 \text{ true} \right) = \sum_{x=6}^{10} C_{10} p^x q^{10-x} / p = 0.7 = \sum_{x=6}^{10} C_{10} 0.7^x (0.3)^{10-x} = 0.8495$

Problem 3: Based on a single observation $x$, taken from an exponential population $f(x) = \theta e^{-\theta x}$, $x \geq 0, \theta > 0$, to test $H_0 : \theta = 2$ against $H_1 : \theta = 1$. If the critical region suggested is, $x \geq 1$, find the probability of type-I and type-II errors. Also find the power function of the test, if $H_1$ suggested is $H_1 : \theta = r$, where $r \leq 2$.

Solution:

To test, $H_0 : \theta = 2$ against $H_1 : \theta = 1$. Critical region is $x \geq 1$.

$\theta$ is the parameter of exponential population then, $f(x) = \theta e^{-\theta x}$, $x \geq 0, \theta > 0$.

$P(\text{Type-I Error}) = P \left( \text{Rej } H_0 / H_0 \text{ true} \right) = P \left( x \geq 1 / \theta = 2 \right) = \int_{1}^{\infty} \theta e^{-\theta x} dx / \theta = 2$
\[ = \int_{1}^{\infty} 2e^{-2x} \, dx = 2 \left[ \frac{e^{-2x}}{-2} \right]_{1}^{\infty} = e^{-2} \]

Power of the test \((1 - \beta) = P (\text{Rej } H_0 / H_1 \text{ is true})\)

\[ = P \left( x \geq 1/\theta = 1 \right) \]

\[ = \int_{1}^{\infty} \theta e^{-\theta x} \, dx / \theta = 1 \]

\[ = \int_{1}^{\infty} e^{-x} \, dx = \left[ \frac{e^{-x}}{-1} \right]_{1}^{\infty} = e^{-1} \]

\(P\) (Type II error) \(\beta = 1 - \) Power of the test \(= 1 - e^{-1}\)

When the alternative hypothesis is \(H_1: \theta = r, \text{ where } r \leq 2\),

Power of the test \((1 - \beta) = P \left( x \geq 1/\theta = r \right), r \leq 2\)

\[ = \int_{1}^{\infty} \theta e^{-\theta x} \, dx / \theta = r \]

\[ = \int_{1}^{\infty} r e^{-rx} \, dx = r \left[ \frac{e^{-rx}}{-r} \right]_{1}^{\infty} = e^{-r}, r \leq 2 \]

\[ \Rightarrow \quad \text{Power function } (1 - \beta) (r) = e^{-r}, r \leq 2 \]

Hence for different values of \(\theta = r, r \leq 2\) in alternative hypothesis the power of the given critical region can be calculated.

**Problem 4:** In a city the milk consumption of families, \(x\), is assumed following the distribution

with, p.d.f. \(f(x) = \frac{1}{\theta} e^{-x/\theta}, x \geq 0, \theta > 0\). To test \(H_0 : \theta = 5\) against \(H_1 : \theta = 10\). \(H_0\) is rejected, if a family selected at random consumes 15 units or more. Obtain the size and power of the test.

**Solution:**

To test \(H_0 : \theta = 5\) against \(H_1 : \theta = 10\), reject \(H_0\), if \(x \geq 15\).

Size of the test \(\alpha = P \left( \text{Rej } H_0 / H_0 \text{ is true} \right) = P \left( x \geq 15 / \theta = 5 \right)\)
\[ \frac{1}{\theta} \int_{15}^{\infty} e^{-\frac{x}{\theta}} \, dx = 5 \]

\[ = \frac{1}{5} \int_{15}^{\infty} e^{-\frac{x}{5}} \, dx = - \left[ e^{-\frac{x}{5}} \right]_{15}^{\infty} = e^{-3} \]

Power of the test \((1 - \beta) = P (x \geq 15/\theta = 10)\)

\[ = \int_{15}^{\infty} \frac{1}{\theta} e^{-\frac{x}{\theta}} \, dx / \theta = 10 \]

\[ = \frac{1}{10} \int_{15}^{\infty} e^{-\frac{x}{10}} \, dx = - \left[ e^{-\frac{x}{10}} \right]_{15}^{\infty} \]

\[ = e^{-\frac{15}{10}} = e^{-\frac{3}{2}}. \]

**Problem 5:** \(x_1, x_2, \ldots, x_9\) are 9 random samples drawn from a normal population \(N(\mu, 5)\). To test \(H_0 : \mu = 5\) against \(H_1 : \mu = 8\). The critical region suggested is \(\bar{x} \geq 7\), where \(\bar{x}\) the sample mean. Find significance level and power of the test.

**Solution:**

To test \(H_0 : \mu = 5\) against \(H_1 : \mu = 8\). Critical region is \(\bar{x} \geq 7\).

Significance level \(\alpha = P\) (Type I error) \(= P\) (Rej \(H_0 / H_0\) is true) \(= P\) (\(\bar{x} \geq 7 / \mu = 5\))

Since \(x_1, x_2, \ldots, x_9\) are random sample from normal population, the sample mean \(\bar{x}\) is distributed as \(N(\mu, \frac{5}{\sqrt{9}})\).

Hence \(U = \frac{(\bar{x} - \mu)\sqrt{9}}{5} \sim N(0, 1)\)

\[ P(\bar{x} \geq 7 / \mu = 5) = P \left( \frac{(\bar{x} - \mu)\sqrt{9}}{5} \geq \frac{(7 - \mu)\sqrt{9}}{5} \right) / \mu = 5 \]
\[ P(U \geq 1.2) = 0.1151 \text{ (from std. normal table)} \]

Power of the test \[ = P(\text{Rej } H_0/ H_1 \text{ is true}) \]

\[ = P(\bar{x} \geq 7/ \mu = 8) \]

\[ = P\left(\frac{(\bar{x} - \mu)\sqrt{\theta}}{\sqrt{5}} \geq \frac{(7 - \mu)\sqrt{\theta}}{\sqrt{5}} / \mu = 8\right) \]

\[ = P(U \geq -0.6) = 0.7257 \text{ (from std. normal table)} \]

**Problem 6:** \( x_1, x_2 \) are 2 random samples drawn from a population with p.d.f. \( f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, x \geq 0, \theta > 0 \). To test \( H_0: \theta = 2 \) against \( H_1: \theta = 4 \). Reject the hypothesis if \( x_1 + x_2 \geq 9.5 \). Obtain significance level and power of the test

**Solution:**

To test \( H_0: \theta = 2 \) against \( H_1: \theta = 4 \). Critical region is \( x_1 + x_2 \geq 9.5 \).

Significance level \( \alpha = P(\text{Rej } H_0/ H_0 \text{ is true}) = P\left(x_1 + x_2 \geq 9.5 / \theta = 2\right) \)

Here the given critical region is a plane in the first quadrant as given in the graph, where \( x_1 + x_2 \geq 9.5, x_1 \geq 0, x_2 \geq 0 \).

The joint density of \( x_1, x_2 \) is given by

\[ f(x_1, x_2) = \frac{1}{\theta} e^{-\frac{x_1}{\theta}} \times \frac{1}{\theta} e^{-\frac{x_2}{\theta}}, x_1 \geq 0, x_2 \geq 0 \text{ (} \therefore x_1, x_2 \text{ are ind.)} \]
\[ f(x_1, x_2; \theta) = \frac{1}{\theta^2} e^{-\left(\frac{x_1 + x_2}{\theta}\right)} \quad \text{for } x_1, x_2 \geq 0, \theta > 0 \]

\[ P(x_1 + x_2 \geq 9.5 / \theta = 2) = 1 - P(x_1 + x_2 < 9.5 / \theta = 2) \]

Note that,

\[
P(x_1 + x_2 < 9.5 / \theta = 2) = \int_{x_1=0}^{9.5} \int_{x_2=0}^{9.5-x_1} \frac{1}{\theta^2} e^{-\left(\frac{x_1 + x_2}{\theta}\right)} \, dx_2 \, dx_1 / \theta = 2
\]

\[
= \int_{x_1=0}^{9.5} \int_{x_2=0}^{9.5-x_1} \frac{1}{16} e^{-\left(\frac{x_1 + x_2}{4}\right)} \, dx_2 \, dx_1
\]

\[
= \frac{1}{16} \int_{x_1=0}^{9.5} \int_{x_2=0}^{9.5-x_1} e^{-\left(\frac{x_1 + x_2}{4}\right)} \, dx_2 \, dx_1
\]

\[
= \frac{1}{4} \int_{x_1=0}^{9.5} e^{-\frac{x_1}{2}} \left[ -2 \left( e^{-\frac{(9.5-x_1)}{2}} - e^0 \right) \right] \, dx_1
\]

\[
= -\frac{1}{2} \left[ x_1 e^{-\frac{9.5}{2}} + 2 e^{-\frac{x_1}{2}} \right]_{x_1=0}^{9.5}
\]

\[
= -\frac{1}{2} \left[ 9.5 e^{-\frac{9.5}{2}} + 2 e^{-\frac{9.5}{2}} - 2 \right]_{0}^{9.5}
\]

\[
= 1 - 4.75 e^{-4.75} - e^{-4.75}
\]

Hence, significance level \( \alpha = 1 - (1 - 4.75 e^{-4.75} - e^{-4.75}) \)

\[
= 5.75 e^{-4.75} = (0.05 \text{ approx.})
\]

Power of the test \( = P(\text{Rej } H_0 / H_1 \text{ is true}) \)
\[ P(x_1 + x_2 \geq 9.5 / \theta = 4) = 1 - P(x_1 + x_2 < 9.5 / \theta = 4) \]

\[ P(x_1 + x_2 < 9.5 / \theta = 4) = \int_{x_1=0}^{9.5} \int_{x_2=0}^{9.5-x_1} \frac{1}{\theta} e^{-\frac{(x_1+x_2)}{\theta}} \, dx_2 \, dx_1 / \theta = 4 \]

\[ = \int_{x_1=0}^{9.5} \int_{x_2=0}^{9.5-x_1} \frac{1}{16} e^{-\frac{(x_1+x_2)}{4}} \, dx_2 \, dx_1 \]

\[ = \frac{1}{16} \int_{0}^{9.5} \int_{0}^{9.5-x_1} e^{-\frac{(x_1+x_2)}{4}} \, dx_2 \, dx_1 \]

\[ = -\frac{1}{4} \left[ 9.5 e^{-\frac{9.5}{4}} + 4 e^{-\frac{9.5}{4}} - 4 \right]_{9.5}^{0} \]

\[ = 1 - \frac{9.5}{4} e^{-\frac{9.5}{4}} - e^{-\frac{9.5}{4}} \]

Hence, power of the test \( = 1 - \left( 1 - \frac{9.5}{4} e^{-\frac{9.5}{4}} - e^{-\frac{9.5}{4}} \right) \)

\[ = \frac{13.5}{4} e^{-\frac{9.5}{4}} = 0.31 \text{ (approx.)} \]

**Problem 7:** Let \( X \) have a pdf of the form \( f(x) = \theta x^{\theta-1}, \ 0 \leq x \leq 1, \theta > 0 \). To test the hypothesis \( H_0 : \theta = 1 \) against \( H_1 : \theta = 2 \), using a random sample \( x_1, x_2 \) of size 2 and define the critical region as \( C = \{(x_1, x_2) : \frac{3}{4x_1} \leq x_2 \} \). Obtain significance level and power of the test

**Solution:**

To test \( H_0 : \theta = 1 \) against \( H_1 : \theta = 2 \), given critical region is \( \frac{3}{4x_1} \leq x_2 \), or \( x_1x_2 \geq \frac{3}{4} \)

for two independent observations \( x_1 \) and \( x_2 \).

\[ f(x) = \theta x^{\theta-1}, \ 0 \leq x \leq 1, \theta > 0 \], since \( x_1 \) and \( x_2 \) are ind. we get,
School of Distance Education

\[ f(x_1, x_2) = \theta^2 (x_1x_2)^{\theta - 1}, \quad 0 \leq x_1, x_2 \leq 1, \quad \theta > 0. \]

Significance level \( \alpha = P(\text{Rej } H_0 / H_1 \text{ is true}) \)

\[ = P(x_1, x_2 \geq \frac{3}{4} / \theta = 1) \]

The given critical region is as shaded in the following graph,

Hence, \( P(x_1, x_2 \geq \frac{3}{4} / \theta = 1) = \int_{x_1 = \frac{3}{4}}^{1} \int_{x_2 = \frac{3}{4} x_1}^{1} \theta^2 (x_1x_2)^{\theta - 1} dx_2 dx_1 / \theta = 1 \)

\[ = \int_{x_1 = \frac{3}{4}}^{1} \int_{x_2 = \frac{3}{4} x_1}^{1} 1 dx_2 dx_1 = \int_{x_1 = \frac{3}{4}}^{1} \left(1 - \frac{3}{4x_1}\right) dx_1 \]

\[ \Rightarrow \alpha = \left(x_1 - \frac{3}{4} \log x_1\right)_{\frac{3}{4}}^{1} = \left[1 - \frac{3}{4} + \frac{3}{4} \log \frac{3}{4}\right] \]

\[ = \frac{1}{4} + \frac{3}{4} \log \frac{3}{4} \]

Power of the test \( = P(\text{Rej } H_0 / H_1 \text{ is true}) \)

\[ = P(x_1, x_2 \geq \frac{3}{4} / \theta = 2) \]
\[
\int_{x_1=\frac{3}{4}}^{1} \int_{x_2=\frac{3}{4}}^{x_1} 2 \left( x_1 x_2 \right)^{\theta-1} dx_2 dx_1 / \theta = 2
\]

\[
= \int_{x_1=\frac{3}{4}}^{1} \int_{x_2=\frac{3}{4}}^{x_1} 4x_1 x_2 dx_2 dx_1
\]

\[
= 4 \int_{x_1=\frac{3}{4}}^{1} x_1 \left( \frac{1}{2} - \frac{9}{32x_1^2} \right) dx_1
\]

\[
= 4 \left[ \frac{x_1^2}{4} - \frac{9}{32} \log x_1 \right]_{\frac{3}{4}}^{1}
\]

\[
\Rightarrow \text{power } (1-\beta) = 4 \left[ \frac{1}{4} - \frac{9}{64} + \frac{9}{32} \log \frac{3}{4} \right]
\]

\[
= \frac{7}{16} + \frac{9}{8} \log \frac{3}{4}
\]

4.5. Most Powerful Test:

So far we considered some testing of hypothesis problem with given critical region. The power and significance level corresponding to a given critical region is calculated. Now a question arising is, can we find a critical region with maximum power and zero significance level? But it is not possible. When we are making the probability of type I error minimum, the probability of type II error increases, thereby power decreases, and vice versa.

In this dilemma, an approach due to Neymaan and Pearson helps us to achieve a most powerful test. By their approach, we fix an affordable level of type I error (say 5% or 1% etc) then consider critical regions with the fixed significance level and among them
choose a critical region with maximum power. The test using such a critical region is known as most powerful test. And the critical region is most powerful critical region (or Best Critical Region).

Neyman-Pearson Theorem suggests the method to find a most powerful critical region to test a simple hypothesis against a simple alternative.

4.6. Neymann-Pearson Theorem:

Let \((x_1, x_2, \ldots, x_n)\) be a random sample from a continuous population with pdf \(f(x, \theta)\). Let \(S\) be a subset of the sample space such that, for a constant \(c\),

and for each \((x_1, x_2, \ldots, x_n)\) belongs to \(S\), \(\frac{L_0}{L_1} \leq c\)

for each \((x_1, x_2, \ldots, x_n)\) NOT belongs to \(S\) \(\frac{L_0}{L_1} \geq c\)

where \(L_0\) and \(L_1\) are the likelihood of the sample when \(H_0\) is true and \(H_1\) is true respectively. The constant, \(c\) is determined so as, \(P(x_1, x_2, \ldots, x_n \in S / H_0) = \alpha\).

Then \(S\) is the most powerful critical region with significance level \(\alpha\) to test the simple hypothesis \(H_0\) against a simple alternative \(H_1\).

Proof:

Consider \(H_0 : \theta = \theta_0\) and \(H_1 : \theta = \theta_1\). Let the likelihood of the sample when \(H_0\) is true is denoted as \(L_0 = f(x_1, \theta_0) f(x_2, \theta_0) \ldots f(x_n, \theta_0) = \prod_{i=1}^{n} f(x_i, \theta_0)\)

when \(H_0\) is true is denoted as, \(L_1 = f(x_1, \theta_1) f(x_2, \theta_1) \ldots f(x_n, \theta_1) = \prod_{i=1}^{n} f(x_i, \theta_1)\)

Neyman-Pearson Theorem, for a best critical region \(S\),

\[\frac{L_0}{L_1} \leq c, \text{ for each } (x_1, x_2, \ldots, x_n) \in S, \text{ and } \frac{L_0}{L_1} \geq c \text{ for each } (x_1, x_2, \ldots, x_n) \notin S\]

Consider another critical region \(S'\), for the given test of size \(\alpha\). \(\Omega\) be the sample space.
Given \( \alpha = P(x_1, x_2, \ldots, x_n \in S / H_0) = \int_S L_0 \, dx = P(x_1, x_2, \ldots, x_n \in S' / H_0) = \int_{S'} L_0 \, dx \)

But, \( \int_S L_0 \, dx = \int_{S \cap S} L_0 \, dx + \int_{S \cap S'} L_0 \, dx \) and \( \int_S L_0 \, dx = \int_{S} L_0 \, dx + \int_{S'} L_0 \, dx \)

\[ \Rightarrow \int_{S \cap S} L_0 \, dx = \int_{S \cap S'} L_0 \, dx \quad \text{(1)} \]

The power of the critical region,

\[ = \int_S L_1 \, dx = \int_{S \cap S} L_1 \, dx + \int_{S \cap S'} L_1 \, dx \]

\[ \geq \int_{S \cap S} L_0 \, dx + \int_{S \cap S'} L_1 \, dx \quad (\because \frac{L_0}{L_1} \leq c \text{ for } x_1, x_2, \ldots, x_n \in S) \]

By (1), \( \int_{S \cap S} L_0 \, dx \quad \Rightarrow \int_S L_1 \, dx \geq \int_{S \cap S} L_0 \, dx + \int_{S \cap S'} L_1 \, dx \)

\( \Rightarrow \int_{S \cap S} L_0 \, dx \quad \Rightarrow \int_{S} L_1 \, dx \geq \int_{S \cap S'} L_1 \, dx (\because \frac{L_0}{L_1} \geq c \text{ for } x_1, x_2, \ldots, x_n \notin S) \)

\[ \Rightarrow \int_{S} L_1 \, dx \geq \int_{S'} L_1 \, dx \]
ie., power of the critical region $S \geq Power$ of the critical region $S'$, Hence the critical region satisfying the conditions of Neyman-Pearson theorem is the best or most powerful critical region.

**Problem 1:** Use Neyman-Pearson Theorem to find a most powerful test with significance level $\alpha$ for testing the hypothesis $H_0: \mu = \mu_0$ against, $H_1: \mu = \mu_1$, $(\mu_1 > \mu_0)$ using a random sample $x_1, x_2, ..., x_n$ drawn from the population with pdf $f(x) = \frac{1}{\sqrt{18\pi}} e^{-\frac{1}{18}(x-\mu)^2}$, $-\infty < x < \infty$.

Solution:

Given $f(x) = \frac{1}{\sqrt{18\pi}} e^{-\frac{1}{18}(x-\mu)^2}$, $-\infty < x < \infty$. For the random samples $x_1, x_2, ..., x_n$,

the likelihood function $L = f(x_1, x_2, ..., x_n, \mu) = \left(\frac{1}{\sqrt{18\pi}}\right)^n e^{-\frac{n}{18}\sum(x_i-\mu)^2}$.

But $\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2$

$=\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}((\bar{x} - \mu)^2 = nS^2 + n(\bar{x} - \mu)^2$, where $\bar{x}$ is the sample mean and $S^2$, the sample variance.

Therefore, $f(x_1, x_2, ..., x_n, \mu) = L = \left(\frac{1}{\sqrt{18\pi}}\right)^n e^{-\frac{n}{18}[S^2 + (\bar{x} - \mu)^2]}$

By Neyman-Pearson theorem, for most powerful critical region $S$,

for $(x_1, x_2, ..., x_n)$ belongs to $S$, $\frac{L_{0}}{L_1} \leq c$

ie., $\frac{L_{0}}{L_1} = \left(\frac{1}{\sqrt{18\pi}}\right)^n e^{-\frac{n}{18}(S^2 + (\bar{x} - \mu_0)^2)} \leq c$
\[
\Rightarrow \quad \frac{e^{-\frac{n}{18}(\bar{x} - \mu_0)^2}}{e^{-\frac{n}{18}(\bar{x} - \mu_1)^2}} \leq c
\]

\[
\Rightarrow \quad e^{-\frac{n}{18}(\bar{x} - \mu_0)^2 - (\bar{x} - \mu_1)^2} \leq c
\]

Taking logarithm on both sides,

\[
\Rightarrow \quad -\frac{n}{18}((\bar{x} - \mu_0)^2 - (\bar{x} - \mu_1)^2) \leq \log c \quad (1)
\]

Since \( \mu_1 > \mu_0 \), dividing both sides of (1) by a negative quantity, \(-\frac{n}{18}(\mu_1 - \mu_0)\), we get,

\[
(2\bar{x} - \mu_0 - \mu_1) \geq \frac{-18 \log c}{n(\mu_1 - \mu_0)}
\]

\[
\Rightarrow \quad \text{for a most powerful critical region, } \quad \bar{x} \geq \frac{1}{2}\left(\frac{-18 \log c}{n(\mu_1 - \mu_0)} + \mu_0 + \mu_1\right)
\]

Let, \( \frac{1}{2}\left(\frac{-18 \log c}{n(\mu_1 - \mu_0)} + \mu_0 + \mu_1\right) = c_i \); Then for most powerful critical region,

\[
\bar{x} \geq c_i \quad \text{---------- (2)}
\]

The size of the test is considered as \( \alpha \). So probability of \( x_1, x_2, \ldots, x_n \) to fall in the critical region, ie., \( \bar{x} > c_i \), when \( H_0 \) is true should be \( \alpha \). Using this condition we can identify the value of the constant \( c_i \)

\[
i.e., \quad P(\bar{x} > c_i / \mu = \mu_0) = \alpha
\]

Since \( x_1, x_2, \ldots, x_n \) are random samples from \( N(\mu, 3) \), the sample mean \( \bar{x} \sim N(\mu, \frac{3}{\sqrt{n}}) \)

\[
\Rightarrow \quad t = \frac{(\bar{x} - \mu)\sqrt{n}}{3} \sim N(0,1)
\]

\[
\Rightarrow \quad P(t > \frac{c_i - \mu}{\frac{3}{\sqrt{n}} / \mu = \mu_0}) = \alpha
\]
From standard normal table a value $t_\alpha$ can be identified as shown such that,

$$P(t > t_\alpha) = \alpha$$

Comparing (3) and (4),

$$t_\alpha = \frac{(c_1 - \mu_0) \sqrt{n}}{3} \quad \Rightarrow \quad c_1 = \mu_0 + \frac{3t_\alpha}{\sqrt{n}}$$

Hence the most powerful critical region is,

$$\bar{x} \geq \mu_0 + \frac{3t_\alpha}{\sqrt{n}}$$

**Remark:** In the derived most powerful critical region, it can be observed that whatever be the value of $\mu$ in the alternative hypothesis, keeping the condition ($\mu_i > \mu_0$), the most powerful critical region is unchanged. That is the most powerful test using this critical region is uniform.

Hence such a critical region is called Uniformly Most Powerful Critical region or the test by using such a critical region is Uniformly Most Powerful Test (UMPT).

**Problem 2:** Use Neymaan-Pearson Theorem to find a most powerful test with significance level $\alpha$ for testing the hypothesis $H_0 : \sigma^2 = \sigma_0^2$ against, $H_1 : \sigma^2 = \sigma_1^2$ ($\sigma_1^2 > \sigma_0^2$) using a random sample $x_1, x_2, ..., x_n$ drawn from the population with pdf $f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{- \frac{(x-\mu)^2}{2\sigma^2}}$, $-\infty < x < \infty$, where $\mu$ is known.

**Solution:**

Note that $x_1, x_2, ..., x_n$ are random sample taken from $N(\mu, \sigma^2)$. Hence the
likelihood function, \[ L = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2} \]

By Neyman-Pearson theorem, for most powerful critical region \( S \),

for \( (x_1, x_2, \ldots, x_n) \) belongs to \( S \), \[ \frac{L_0}{L_1} \leq c \]

ie., \[ \frac{L_0}{L_1} = \frac{\left( \frac{1}{\sigma_0 \sqrt{2\pi}} \right)^n e^{-\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (x_i - \mu)^2}}{\left( \frac{1}{\sigma_1 \sqrt{2\pi}} \right)^n e^{-\frac{1}{2\sigma_1^2} \sum_{i=1}^{n} (x_i - \mu)^2}} \leq c \]

\[ \Rightarrow \frac{L_0}{L_1} = \frac{\sigma_1^n}{\sigma_0^n} e^{-\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{1}{2\sigma_1^2} \sum_{i=1}^{n} (x_i - \mu)^2} \leq c \]

\[ \Rightarrow \log \left( e^{-\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{1}{2\sigma_1^2} \sum_{i=1}^{n} (x_i - \mu)^2} \right) \leq \log \left( c \frac{\sigma_0^n}{\sigma_1^n} \right) \]

\[ \Rightarrow \left( \frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_0^2} \right) \sum_{i=1}^{n} (x_i - \mu)^2 \leq \log \left( c \frac{\sigma_0^n}{\sigma_1^n} \right) \]

Since \( \sigma_1^2 > \sigma_0^2 \),

\[ \left( \frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_0^2} \right) = \frac{\sigma_0^2 - \sigma_1^2}{2\sigma_0^2 \sigma_1^2} \]

is negative

\[ \Rightarrow \sum_{i=1}^{n} (x_i - \mu)^2 \geq \log \left( c \frac{\sigma_0^n}{\sigma_1^n} \right) = c_1 \] (say)

ie., the most powerful critical region is given by,

\[ \sum_{i=1}^{n} (x_i - \mu)^2 \geq c_1 \]
The size of the test is considered as \( \alpha \). So probability of \( x_1, x_2, ..., x_n \) to fall in the critical region, ie., \( \sum_{i=1}^{n} (x_i - \mu)^2 \geq c_1 \), when \( H_0 \) is true should be \( \alpha \).

Using this condition we can identify the value of the constant \( c_1 \)

\[
\begin{align*}
    \text{ie.} \quad P\left( \sum_{i=1}^{n} (x_i - \mu)^2 \geq \frac{c_1}{\sigma^2} = \sigma_0^2 \right) = \alpha \\
    \Rightarrow P\left( \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \geq \frac{c_1}{\sigma^2} = \sigma_0^2 \right) = \alpha \\
    \text{but we have} \quad \frac{(x_i - \mu)}{\sigma} \sim N(0,1) \quad \text{and} \quad \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \sim \chi^2_{(n)} \\
    \Rightarrow P\left( \sum_{i=1}^{n} \chi^2_{(n)} / H_0 \geq \frac{c_1}{\sigma_0^2} \right) = \alpha \quad ---- (1)
\end{align*}
\]

From the table of \( \chi^2_{(n)} \), the value of \( \chi^2_{\alpha:n} \), as shown below

Comparing (1) and (2), \( \frac{c_1}{\sigma_0^2} = \chi^2_{\alpha:n} \quad \Rightarrow \quad c_1 = \chi^2_{\alpha:n}\sigma_0^2 \)

Hence for most powerful critical region, \( \sum_{i=1}^{n} (x_i - \mu)^2 \geq \chi^2_{\alpha:n}\sigma_0^2 \)

Thus the most powerful critical region for testing \( H_0 : \sigma^2 = \sigma_0^2 \) against \( H_1 : \sigma^2 = \sigma^1_2 \quad (\sigma^1_2 > \sigma_0^2) \) with significance level \( \alpha \) is, \( \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \geq \chi^2_{\alpha:n} \)
**EXERCISES**

1. Let $p$ be the probability that a coin will fall head in a single toss in order to test 

   $H_0: p = \frac{1}{2}$ against $H_1: p = \frac{3}{4}$. The coin is tossed 5 times and $H_0$ is rejected if more than 3 heads obtained. Find the size and power of the test.

2. 10 random samples $x_1, x_2, \ldots, x_{10}$ are taken from $N(\mu, 5)$. To test $H_0: \mu = 0$ against $H_1: \mu = 2$. The critical region suggested is $x_1 + 2x_2 + 3x_3 + \ldots + 10x_{10} > 1$. Obtain the probability of type-I and type-II errors.

3. Let $p$ be the proportion of smokers in a city. To test $H_0: p = \frac{1}{2}$ against $H_1: p = \frac{3}{4}$. $H_0$ is rejected if 60 or more persons are found smokers in a sample of 100 persons. Compute significance level and power of the test.

4. A single value $x$ is drawn from a normal population $N(m, 5)$. The null hypothesis $H_0: \mu = 50$ is accepted if $x \leq 75$. Otherwise $H_1: \mu = 60$ is considered. Evaluate significance level and power of the test.

5. A sample of size 10 is taken from a normal population with $\sigma = 1$ to test $H_0: \mu = 5$ against $H_1: \mu = 6$. The critical region is $x \geq 5.52$. Find significance level and power of the test.

6. Obtain the best critical region for testing $H_0: \mu = \mu_0$ against $H_1: \mu > \mu_0$ in $N(\mu, \sigma)$ using a random sample of size $n$. Also find the power function.

7. In testing $H_0: \sigma = \sigma_0$ against $H_1: \sigma = \sigma_1 (\neq \sigma_0)$ for the distribution with pdf $f(x) = e^{-\frac{(x-\theta)^2}{\sigma}}$, $0 < x < \infty, \sigma > 0$. Show that the UMP test is of the form $\sum x_i \geq \text{constant}$ and $\sum x_i \leq \text{constant}$.

8. $x_1, x_2, \ldots, x_n$ be a random sample of size $n$ drawn from a population with pdf $f(x) = \theta x^{\theta-1}$. $0 < x < 1, \theta > 0$. Obtain most powerful test for testing $H_0: \theta = \theta_0$ against $H_1: \theta > \theta_0$. Examine whether the most powerful test is UMPT.

***************
CHAPTER 5

LARGE SAMPLE TESTS

5.1. Large Sample Tests:

To test a statistical hypothesis, we construct a test criterion based on an appropriate test statistic with known probability distribution. If the probability distribution of the statistic considered is not familiar, it is troublesome to perform the testing procedure. Fortunately it is known that when the sample size is large, by central limit theorem most of the statistics are normally or at least approximately normally distributed. That is if \( U \) is a statistic considered, then \( U \sim N[E(U),SD(U)] \) as \( n \) become very large, or
\[
    t = \frac{U - E(U)}{SD(U)} \sim N(0,1)
\]
for large \( n \). This important result may profitably used for the test construction.

The standard deviation of any statistic is called its standard error. While testing a hypothesis \( H_0: \theta = \theta_0 \), naturally taking a statistic \( U \) with \( E(U) = \theta \). Using standard error of \( U \), a test statistic \( t \) following standard normal distribution can be formed. Then for large \( n \), the probability that it will fall in any region in its range of variation can be found using standard normal table.

5.2. Testing mean of a population:

Consider the hypothesis regarding the mean \( \mu \) of a population. Let \( \sigma \) be the standard deviation of the population. Consider the hypothesis \( H_0: \mu = \mu_0 \). The alternative hypothesis may be (i) \( H_i: \mu > \mu_0 \) (ii) \( H_i: \mu < \mu_0 \) or (iii) \( H_i: \mu \neq \mu_0 \).

Case I: \( \sigma \) is known:

Let \( x_1, x_2, ..., x_n \) be the samples from the population with sample mean \( \bar{x} \) and sample variance \( s^2 \). In testing of the population mean \( \mu \) the sample mean \( \bar{x} \) (which is an unbiased estimator of population mean) is considered as the best test statistic. If
$H_0 : \mu = \mu_0$ is to be tested against $H_1 : \mu > \mu_0$, we reject $H_0$, if $\bar{x} > c$ a constant $c$. If $\alpha$ is the given significance level, to find the constant $c$ such that $P(\bar{x} > c / H_0) = \alpha$, that is

$$P(\bar{x} > c / \mu = \mu_0) = \alpha.$$  

Since the samples are taken from a population with mean $\mu$ and standard deviation $\sigma$, $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$ for large $n$. That is, $z = \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma} \sim N(0,1)$ for large $n$.

$$P(\bar{x} > c / \mu = \mu_0) = \alpha$$

$$\Rightarrow P\left(\frac{\bar{x} - \mu}{\sigma} > \frac{(c - \mu)\sqrt{n}}{\sigma} / \mu = \mu_0\right) = \alpha$$

$$\Rightarrow P\left(z > \frac{(c - \mu_0)\sqrt{n}}{\sigma}\right) = \alpha , z \sim N(0,1) \quad ---- (1)$$

From the table of standard normal distribution, one can get a $t_\alpha$, such that

$$P(z > t_\alpha) = \alpha.$$  Hence from (1), $\frac{(c - \mu_0)\sqrt{n}}{\sigma} = t_\alpha$

$$\Rightarrow c = t_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0$$

Then the test criterion is to reject $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$, when $\bar{x} > t_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0$

ie., reject $H_0$ when $t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} > t_\alpha$

If $H_0 : \mu = \mu_0$ is to be tested against $H_1 : \mu < \mu_0$, we reject $H_0$, if $\bar{x} < c$ a constant $c$. If $\alpha$ is the given significance level, to find the constant $c$ such that $P(\bar{x} < c / H_0) = \alpha$, that is

$$P(\bar{x} < c / \mu = \mu_0) = \alpha.$$  

$$\Rightarrow P\left(z < \frac{(c - \mu_0)\sqrt{n}}{\sigma}\right) = \alpha , z \sim N(0,1) \quad ---- (2)$$
Find a \(-t_\alpha,\) from standard normal table, such that \(P( z < -t_\alpha ) = \alpha .\) Hence from (2),

\[
\frac{(c-\mu_0)\sqrt{n}}{\sigma} = -t_\alpha .
\]

Then the test criterion is to reject \(H_0 : \mu = \mu_0\) against \(H_1 : \mu < \mu_0,\) when, \(\bar{x} < -t_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0\)

ie., reject \(H_0\) when \(t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} < -t_\alpha\)

If \(H_0 : \mu = \mu_0\) is to be tested against \(H_1 : \mu \neq \mu_0,\) we reject \(H_0,\) if \(|\bar{x} - \mu_0| > c\) a constant \(c.\) If \(\alpha\) is the significance level, to find the constant \(c\) such that \(P(|\bar{x} - \mu| > c / H_0) = \alpha ,\)

that is \(P(|\bar{x} - \mu_0| > c) = \alpha .\) That is,

\[
P( |z| > \frac{(c)\sqrt{n}}{\sigma} ) = \alpha , \, z \sim N(0,1) \quad \text{--------- (3)}
\]

Standard normal table gives value of \(t_\alpha,\) such that

\[
\Rightarrow \quad P( |z| > t_\frac{\alpha}{2} ) = \alpha , \, z \sim N(0,1)
\]

Hence, from (3)

\[
\frac{c\sqrt{n}}{\sigma} = t_\frac{\alpha}{2}
\]

\[
\Rightarrow \quad c = t_\frac{\alpha}{2} \frac{\sigma}{\sqrt{n}}
\]

Then the test criterion is to reject \(H_0 : \mu = \mu_0\) against \(H_1 : \mu \neq \mu_0,\) when, \(|\bar{x} - \mu_0| > t_\frac{\alpha}{2} \frac{\sigma}{\sqrt{n}}\)

ie., reject \(H_0\) when \(|t| = \left| \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} \right| > t_\frac{\alpha}{2} .\)
**Problem:** The average height of a sample of 400 college students is found to be 4.75 ft. The standard deviation of the population is believed to be 1.5. Does the data contradict the hypothesis that the mean height of students is 4.48 ft at 1% level of significance?

**Solution:**

Let \( \mu \) be the mean height of the students. To test whether the data consists of 400 samples contradict the belief that average height is 4.48

To test \( H_0 : \mu = 4.48 \) against \( H_1 : \mu \neq 4.48 \)

Test criterion is 

\[
 t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma}
\]

Here, 
\[
 t = \frac{(4.75 - 4.48)\sqrt{400}}{1.5} = 3.6
\]

The test procedure is to reject \( H_0 \) if \( |t| > t_{\alpha} \).

Given \( \alpha = 0.01 \), from std normal table, \( t_{\alpha} = 3 \)

Then, here \( |t| > t_{\alpha} \). So reject \( H_0 \) at 1% level. That is the data contradict the assumption that the average height is 4.48.

**Case II: \( \sigma \) is Unknown:**

We have sample variance \( S^2 \) as a consistent estimator of \( \sigma^2 \). As \( n \) becomes large \( E(S^2) = \sigma^2 \). Here we consider large sample, hence the value of \( S^2 \) is considered as the value of \( \sigma^2 \) and can perform the test as already described.

**Problem:** An insurance agent has claimed that the average age of policy-holders who insure through him is less than the average for all agents, which is 30.5 years. A random sample of policy-holders who had insured through him gave the following age distribution:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of persons</td>
<td>12</td>
<td>22</td>
<td>20</td>
<td>30</td>
<td>16</td>
</tr>
</tbody>
</table>

Test the claim at 5% level of significance.
Solution:

Let $\mu$ be the mean age of the policy-holders insured by the agent. A sample of 100 policy-holders insured by the agent is given.

To test $H_0 : \mu = 30.5$ against $H_1 : \mu < 30.5$.

The population standard deviation of the age of policy-holders is not given. Since it is a large sample, sample standard deviation is approximated by the population standard deviation.

<table>
<thead>
<tr>
<th>Age $x$</th>
<th>No. of persons $f$</th>
<th>Mid-$x$</th>
<th>$fx$</th>
<th>$fx^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 - 20</td>
<td>12</td>
<td>18</td>
<td>216</td>
<td>3888</td>
</tr>
<tr>
<td>21 - 25</td>
<td>22</td>
<td>23</td>
<td>506</td>
<td>11638</td>
</tr>
<tr>
<td>26 - 30</td>
<td>20</td>
<td>28</td>
<td>560</td>
<td>15680</td>
</tr>
<tr>
<td>31 - 35</td>
<td>30</td>
<td>33</td>
<td>990</td>
<td>32670</td>
</tr>
<tr>
<td>36 - 40</td>
<td>16</td>
<td>38</td>
<td>608</td>
<td>23104</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2880</td>
<td>86980</td>
</tr>
</tbody>
</table>

$$\bar{x} = \frac{1}{N} \sum_i f_i x_i = \frac{1}{100} (2880) = 28.8$$

$$S.D. = \sqrt{\frac{1}{N} \sum_i f_i x_i^2 - (\bar{x})^2}$$

$$= \sqrt{\frac{1}{100} 86980 - (28.8)^2}$$

$$= \sqrt{869.8 - 829.44} = \sqrt{40.36} = 6.35$$
Since the given sample is of large size, the sample standard deviation is approximated to the population standard deviation.

Corresponding to 5% level of significance, from standard normal table we get 
\( t_a = 1.65 \). The test criterion is to reject \( H_0 \), if, \( t < -t_a \), where 
\[
    t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s}
\]

Now
\[
    t = \frac{(28.8 - 30.5)\sqrt{100}}{6.35} = -2.68
\]

Here, the calculated value of \( t = -2.68 < -t_a = -1.65 \). Hence reject \( H_0: \mu = 30.5 \) at 5% level of significance. That is the agent’s claim is acceptable.

5.3. Testing the inequalities of means of two populations:

Let \( \mu_1, \mu_2 \) be the means and \( \sigma_1, \sigma_2 \) be the standard deviations of two populations and let to test 
\( H_0: \mu_1 = \mu_2 \).

If \( H_1: \mu_1 > \mu_2 \), reject \( H_0 \) if \( \bar{x}_1 - \bar{x}_2 > c \), where \( \bar{x}_1 \) and \( \bar{x}_2 \) are the sample means of \( n_1 \) samples from first population and \( n_2 \) samples from the second population taken independently. Let the significance level is \( \alpha \), then to consider \( c \) such that
\[
P(\bar{x}_1 - \bar{x}_2 > c / H_0) = \alpha
\]

If \( t = \bar{x}_1 - \bar{x}_2 \), \( z = \frac{t - E(t)}{SD(t)} \sim N(0,1) \) for large \( n \). Then,
\[
P \left( \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > \frac{c - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}/H_0 \right) = \alpha .
\]
This implies that
\[
P \left( z > \frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right) = \alpha , \text{ from standard normal table } P( z > t_\alpha ) = \alpha
\]
\[ \Rightarrow \frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = t_\alpha; \quad \text{so,} \quad \Rightarrow c = t_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \]

Hence reject \( H_0: \mu_1 = \mu_2 \) against \( H_1: \mu_1 > \mu_2 \), if \( \bar{x}_1 - \bar{x}_2 > t_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \)

The critical region is \( t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > t_\alpha \)

In similar way,

Reject \( H_0: \mu_1 = \mu_2 \) against \( H_1: \mu_1 < \mu_2 \), if, \( t < -t_\alpha \)

And,

Reject \( H_0: \mu_1 = \mu_2 \) against \( H_1: \mu_1 \neq \mu_2 \), if, \( |t| > \frac{t_\alpha}{\sqrt{2}} \)

Remark: If \( \sigma_1, \sigma_2 \) are unknown, since the sample size is large, the values of sample standard deviations \( S_1, S_2 \) are approximated to perform the testing.

If \( \sigma_1, \sigma_2 \) are unknown and in addition it is assumed that \( \sigma = \sigma_1 = \sigma_2 \), then the value of \( \sigma \) is approximated by \( \sqrt{\frac{n_1S_1^2 + n_2S_2^2}{n_1 + n_2}} \).

**Problem 1:** A sample of 400 men from South India has a mean height of 170 cms. and a standard deviation of 30 cms. while a sample of 200 men from North India has a mean height of 178 cms with a standard deviation of 32 cms. Do the data indicate that North Indians are on the average taller than the South Indians?

**Solution:**

Large sample of sizes 400 and 200 respectively are taken from two populations and their mean is found to be 170 and 178. The sample standard deviations are 30 and 32 respectively.
Let $\mu_1$ be the mean height of south Indians and $\mu_2$ be that of North Indians. To test $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 < \mu_2$.

The test statistic used is

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

Since the sample sizes are large, then if $\sigma_1^2$ and $\sigma_2^2$ are unknown, they are approximated by the sample variances.

Hence here,

$$t = \frac{170 - 178}{\sqrt{\frac{30^2}{400} + \frac{32^2}{200}}} = -2.94$$

Reject $H_0$, if $t < -t_\alpha$.

Consider a significance level of 5%. Then, from standard normal table $t_\alpha = 1.645$. Therefore here, $t = -2.94 < -t_\alpha = -1.645$. Hence reject $H_0$. That is the data indicates that North Indians are on average taller than South Indians.

**Problem 2:** The mean height of 50 male students who showed above average participation in college athletics was 68.2 inches with a standard deviation of 2.5 inches; while 50 male students who showed NO interest in such participation had a mean height of 67.5 inches with a standard deviation of 2.8 inches.

i. Test the hypotheses that male students who participate in college athletics are taller than other male students.

ii. By how much should the sample size of each of the two groups be increased in order that the observed difference of 0.7 inches in the mean heights significant at 5% level of significance. [Assume samples are of equal in size].

**Solution:**

i. Let $X_1$ and $X_2$ are the variables representing respectively, the heights of the students who showed and who not showed interest in athletics. Let $n_1$ be the size of sample with mean $\bar{x}_1$ and S.D. $s_1$, taken from the first set. And $n_2$ be that of the second set with sample mean $\bar{x}_2$ and S.D $s_2$. 

\[\text{STATISTICAL INFERENCE}\] 

Page 97
Given, \( n_1 = 50, \bar{x}_1 = 68.2 \) and \( s_1 = 2.5 \); \( n_2 = 50, \bar{x}_2 = 67.5 \) and \( s_2 = 2.8 \).

Let \( \mu_1 \) and \( \mu_2 \) be the population mean of the sets of students considered. To test the hypothesis \( H_0: \mu_1 = \mu_2 \) against \( H_1: \mu_1 > \mu_2 \).

The test statistic used is \( t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \). Since the sample sizes are large, then if \( \sigma_1^2 \) and \( \sigma_2^2 \) are unknown, they are approximated by the sample variances.

Hence here, \( t = \frac{68.2 - 67.5}{\sqrt{\frac{2.5^2}{50} + \frac{2.8^2}{50}}} = \frac{0.7}{\sqrt{0.2818}} = 1.319 \).

Reject \( H_0 \), if \( t > t_\alpha \)

For 5% significance level, from standard normal table, \( t_\alpha = 1.645 \).

But here \( t < t_\alpha \). Hence we accept \( H_0 \). That is, the data is not significant to believe the first set students are taller than that of the second set.

ii. Let \( n \) be the size of the samples taken from the two sets of students.

The difference of 0.7 units in the mean heights of the samples of sizes \( n \), will become significant if \( t > t_\alpha \), for that value of \( n \).

At 5% significance level, the data become significant if,

\[
t = \frac{68.2 - 67.5}{\sqrt{\frac{2.5^2}{n} + \frac{2.8^2}{n}}} > 1.645
\]

\[
\Rightarrow \frac{0.7}{\sqrt{\frac{14.09}{n}}} > 1.645
\]

\[
\Rightarrow n > \left( \frac{1.645}{0.7} \right)^2 \times 14.09 = 77.81.
\]
That is, when the numbers of samples from two sets are 78 each, the difference 0.7 in the sample mean will become significant. Hence the sample sizes should be increased by 28, in order that the observed difference of 0.7 inches in the mean heights to become significant at 5% significance level.

**Problem 3:** A sample of 200 students from college ‘A’ scored mean mark of 65 with standard deviation 8 for mathematics in a university examination. Another sample of 100 students from college ‘B’ scored a mean mark of 60 in the same paper with a standard deviation 6. Does the data indicate any significance difference between the colleges in terms of the performance in mathematics paper? Assume the S.D’s are same. (sig. level 5%).

**Solution:**

Let $\mu_1$ be the mean marks of the students of college ‘A’ and $\mu_2$ be that of college ‘B’. To test $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$.

Number of sample from first college, $n_1 = 200$, $\bar{x}_1 = 65$, $s_1 = 8$

Number of sample from college ‘B’, $n_2 = 100$, $\bar{x}_2 = 60$, $s_2 = 6$

The S.D’s of the populations are unknown and assumed same. Then the common standard deviations can be approximated by,

$$\sigma = \sqrt{\frac{n_1s_1^2 + n_2s_2^2}{n_1 + n_2}}$$

$$\Rightarrow \sigma = \sqrt{\frac{200 \times 64 + 100 \times 36}{200 + 100}} = 7.393$$

In this case, the test statistic is,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{n_1s_1^2 + n_2s_2^2}{n_1 + n_2}\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\Rightarrow t = \frac{65 - 60}{\sqrt{\left(\frac{200 \times 64 + 100 \times 36}{200 + 100}\right) \left(\frac{1}{200} + \frac{1}{100}\right)}} = 5.49$$

Reject $H_0$, if $|t| > t_{\alpha/2}$. At 5% level of significance, from standard normal table, $t_{\alpha/2} = 1.96$. Hence here, it can verify that $|t| > t_{\alpha/2}$. 

STATISTICAL INFERENCE
Then reject $H_0$. That is the two colleges are differing significantly.

**Problem 4:** Electric bulbs manufactured by X and Y companies gave the following result:

<table>
<thead>
<tr>
<th>Company</th>
<th>No. of bulbs used</th>
<th>Mean life in hours</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>100</td>
<td>1300</td>
<td>82</td>
</tr>
<tr>
<td>Y</td>
<td>100</td>
<td>1248</td>
<td>93</td>
</tr>
</tbody>
</table>

Using standard error of the difference between means, state whether there is any significant difference in the life of the two makes.

**Solution:**

With usual notations, given $n_1 = 100, \bar{x}_1 = 1300, s_1 = 82; n_2 = 100, \bar{x}_2 = 1248, s_2 = 93$.

Let $\mu_1$ and $\mu_2$ be the average life of bulbs by company X and company Y respectively.

To test, $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$.

Standard error of the difference between means $= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

The test criterion is to reject $| t | > t_{\alpha}$, where, $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.

Since the sample sizes are large use corresponding sample standard deviations instead of $\sigma_1$ and $\sigma_2$

Hence, $t = \frac{1300 - 1248}{\sqrt{\frac{82^2}{100} + \frac{93^2}{100}}} = 4.19$

For 5% significance level, from standard normal population $t_{\alpha} = 1.96$.

Here, $| t | > t_{\alpha}$. So reject $H_0$. There is significant difference in two makes.
5.4. Testing the population proportion:

Consider the human population of a district. The proportion of smokers of the population is to be evaluated. The entire population can be divided into smokers and non-smokers. Let \( X \) denotes the number of smokers out of \( N \) members of the population then, \( p = \frac{X}{N} \) is the population proportion regarding smoking habit.

Consider a particular characteristic of the population and assume to test \( H_0 : p = p_0 \) based on a sample of size \( n \) taken from the population. Consider the value of \( \frac{x}{n} \), which is the sample proportion regarding the particular characteristic. \( E(\frac{x}{n}) = p \).

Then reject \( H_0 \) against \( H_1 : p > p_0 \), if \( \frac{x}{n} > c \).

Assume the significance level is \( \alpha \). Then for the critical region, it is to find \( c \), such that \( P(\frac{x}{n} > c / H_0) = \alpha \).

\[
\Rightarrow P \left( \frac{x}{n} - E(\frac{x}{n}) > \frac{c - E(\frac{x}{n})}{SD(\frac{x}{n})} \right) / H_0 = \alpha , \text{ where, } z = \frac{\frac{x}{n} - E(\frac{x}{n})}{SD(\frac{x}{n})} \sim N(0,1) \text{ for large } n
\]

\[
E(\frac{x}{n}) = p, SD(\frac{x}{n}) = pq \Rightarrow P \left( z > \frac{c - p_0}{\sqrt{\frac{pq}{n}}} \right) = \alpha
\]

from standard normal table \( P( z > t_\alpha ) = \alpha \),

\[
\Rightarrow \frac{c - p_0}{\sqrt{\frac{pq}{n}}} = t_\alpha \Rightarrow c = p_0 + t_\alpha \sqrt{\frac{pq}{n}}
\]

Hence the critical region with size \( \alpha \) is, \( \frac{x}{n} > p_0 + t_\alpha \sqrt{\frac{pq}{n}} \)

That is reject \( H_0 : p = p_0 \) against \( H_1 : p > p_0 \), with significance level \( \alpha \), if
Similarly, Reject $H_0: p = p_0$ against $H_1: p < p_0$, with significance level $\alpha$, if

$$t < -t_\alpha,$$

and

Reject $H_0: p = p_0$ against $H_1: p \neq p_0$, with significance level $\alpha$, if

$$|t| > \frac{t_\alpha}{\sqrt{2}}.$$

**Problem:** A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Test the hypothesis that the percentage of bad apples is 20% at 5% level of significance. Also obtain a 95% confidence interval for the percentage of bad pineapples in the consignment.

**Solution:**

Let $p$ be the proportion of bad pineapples in the consignment.

It is to test $H_0: p = 0.20$ against $H_1: p \neq 0.20$

The test statistic is $t = \frac{x - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, reject $H_0$, if $|t| > \frac{t_\alpha}{\sqrt{2}}$.

$$t = \frac{\frac{65}{500} - 0.20}{\sqrt{\frac{0.20 \times 0.80}{500}}} = -3.91$$

Given, $n = 500$, $x = 65$. Then $t = \frac{65}{500} - 0.20 = -3.91$

At 5% level of significance, standard normal table gives, $t_\alpha = 1.96$.

But, here $|t| > \frac{t_\alpha}{\sqrt{2}}$. Thus reject $H_0: p = 0.20$. 

$$t = \frac{x - p_0}{n} > t_\alpha \sqrt{n}$$
We have \( \alpha = P \left( \left| t \right| > \frac{t_{\alpha/2}}{H_0} \right) \), i.e., \( 1 - \alpha = P \left( \left| t \right| < \frac{t_{\alpha/2}}{H_0} \right) \).

Hence for \( \alpha = 0.05 \), we get a confidence interval for the proportion of bad pineapples as

\[
\left( \frac{x}{n} - t_{\alpha/2} \sqrt{\frac{p_0 q_0}{n}}, \frac{x}{n} + t_{\alpha/2} \sqrt{\frac{p_0 q_0}{n}} \right).
\]

Since the population proportion of bad pineapples is not known approximate \( p_0 \) by \( \frac{x}{n} \).

So a 95% confidence interval for bad pineapples is,

\[
\left( \frac{65}{500} - 1.96 \sqrt{\frac{0.13 \times 0.87}{500}}, \frac{65}{500} + 1.96 \sqrt{\frac{0.13 \times 0.87}{500}} \right)
\]

\[
= (0.1005, 0.1595)
\]

Hence, 95% confidence limit for percentage of bad pineapples are \((10.05, 15.95)\).

5.5. Testing the equality of proportion of two populations:

Two test equality of proportion (\( H_0 : p_1 = p_2 \)) of two populations, consider independently taken \( n_1 \) samples from first population and \( n_2 \) samples from the second population. Let \( \frac{x_1}{n_1}, \frac{x_2}{n_2} \) are the sample proportions regarding the character considered, for the first and second population respectively.

If \( H_1 : p_1 > p_2 \), reject \( H_0 \) when \( \frac{x_1}{n_1} - \frac{x_2}{n_2} > c \) where \( E \left( \frac{x_1}{n_1} - \frac{x_2}{n_2} \right) = p_1 - p_2 \) and

\[
SD \left( \frac{x_1}{n_1} - \frac{x_2}{n_2} \right) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}.\]

To find with significance level \( \alpha \), \( P \left( \frac{x_1}{n_1} - \frac{x_2}{n_2} > c / H_0 \right) = \alpha \).
For large \( n_1 \) and \( n_2 \),

\[
z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2} - E(\frac{x_1}{n_1} - \frac{x_2}{n_2})}{SD(\frac{x_1}{n_1} - \frac{x_2}{n_2})} \sim N(0,1), \text{ then,}
\]

\[
P \left( \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2} - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} > \frac{c-(p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} / H_0 \right) = \alpha
\]

\[\Rightarrow \quad P \left( z > \frac{c-(0)}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \right) = \alpha \quad \text{under } H_0: p_1 = p_2 = p \text{ (say)}
\]

from standard normal table \( P( z > t_\alpha ) = \alpha \),

\[\Rightarrow \quad \frac{c-(0)}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} = t_\alpha \quad \Rightarrow \quad c = t_\alpha \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}
\]

Then, reject \( H_0 \), against \( H_1: p_1 > p_2 \), if \( t = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} > t_\alpha \)

If \( H_1: p_1 < p_2 \), reject \( H_0 \), if \( t < -t_\alpha \)

If \( H_1: p_1 \neq p_2 \), reject \( H_0 \), if \( | t | > t_\alpha \frac{2}{\alpha} \)

If \( p \) is unknown, estimate its value by \( p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \), where \( p_1 = \frac{x_1}{n_1} \) and \( p_2 = \frac{x_2}{n_2} \).

**Problem 1:** A random sample of 400 men and 600 women were asked whether they should like to have a flyover near their residence. 200 men and 325 women were in favor of the proposal. Test the hypothesis that proportions of men and women in favor of the proposal are same against that they are not, at 5% level of significance.
Solution:

Let \( p_1 \) be the proportion of women in favor of the proposal and \( p_2 \) is that of men. Then the problem is to test \( H_0 : p_1 = p_2 \) against \( H_1 : p_1 \neq p_2 \).

Given a sample of size \( n_1 = 600 \) from the women, and \( n_2 = 400 \) from the group of men. From the sample the proportion of women in favor of fly over \( p_1 = \frac{325}{600} \) and that of the second city is \( p_2 = \frac{200}{400} \). Assume \( p_1 = p_2 = p \), where

\[
p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2},
\]

This implies \( p = \frac{325 + 200}{600 + 400} = 0.525 \).

The test statistic used is \( t = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \)

\[
= \frac{\frac{325}{600} - \frac{200}{400}}{\sqrt{\frac{0.525 \times 0.475}{600} + \frac{0.525 \times 0.475}{400}}} = \frac{0.542 - 0.5}{0.0322} = 1.304
\]

For \( \alpha = 0.05 \), from std normal table \( t_{\frac{\alpha}{2}} = 1.96 \).

It can be verified that, \( |t| < t_{\frac{\alpha}{2}} \). Then accept \( H_0 \).

That is men and women do not differ significantly regarding the proposal for the fly over.

Problem 2: In a year there are 956 births in a town A, of which 52.5% were males, while when towns A and B are combined; this proportion in a total of 1,406 births was 0.496. Is there any significance difference in the proportion of male births in the two towns? (sig. level 1%)
Solution:

Let $p_1$ be the proportion male birth in the town A, and $p_2$ be the proportion of male births in town B. Then the problem is to test $H_0: p_1 = p_2$ against $H_1: p_1 \neq p_2$.

The number of samples from town A $n_1 = 956$. Given the sample proportion of male birth of town A = 0.525.

The combined sample proportion of male birth out of 1,406 births from town A and B is given as 0.496.

Hence the number of samples from town B, $n_2 = 1,406 - 956 = 450$.

From the given combined proportion, $p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = 0.496$, the sample proportion of the town B can be found as,

$$p_2 = \frac{0.496 \times 1,406 - 956 \times 0.525}{450} = 0.434.$$

The test statistic used is $t = \frac{\bar{x}_1 - \bar{x}_u}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}}$

$$= \frac{0.525 - 0.434}{\sqrt{\frac{0.496 \times 0.504}{956} + \frac{0.496 \times 0.504}{450}}} = 3.791$$

At $\alpha = 0.01$, from std normal table, $t_{a/2} = 3$. Therefore here $|t| > t_{a/2}$.

Hence reject $H_0$, that the proportion of male birth are equal for both the town.
Problem 3: A machine produced 20 defective parts in a batch of 400. After overhauling, it produced 10 defectives in a batch of 300. Has the machine improved? [at 5% level]

Solution:

Let \( p_1 \) be the proportion of defective production before overhauling [repair] and \( p_2 \) is that of after overhauling. The machine is said to be improved if the proportion of defective production is less. Then the problem is to test \( H_0: p_1 = p_2 \) against \( H_1: p_1 > p_2 \).

Given a sample of size \( n_1 = 400 \) and \( n_2 = 300 \). From the sample, the proportion of defective production before overhauling \( p_1 = \frac{20}{400} \) and that of after overhauling is \( p_2 = \frac{10}{300} \). Assume \( p_1 = p_2 = p \), where \( p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \).

This implies \( p = \frac{20 + 10}{400 + 300} = 0.0429 \)

The test statistic used is \( t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \)

\[
= \frac{\frac{20}{400} - \frac{10}{300}}{\sqrt{\frac{0.0429 \times 0.9571}{400} + \frac{0.0429 \times 0.9571}{300}}} = \frac{0.016666}{0.0154762} = 1.0769
\]

For \( \alpha = 0.05 \), from std normal table \( t_\alpha = 1.645 \).

It can be verified that, \( t < t_\alpha \). Then accept \( H_0 \).

That is the machine has not at all improved after overhauling.

5.6. Goodness of fit:

The chi-square test is one of the simplest and most commonly used non-parametric tests of significance by Karl Pearson. It is the most suitable test to compare the obtained
set of ‘observed’ frequencies in a given set with a set of theoretical frequencies within them.

Consider a set of \( n \) possible events, arranged in classes or cells. Let these events occur with frequencies \( O_1, O_2, \ldots, O_n \) called observed frequencies. Out of \( N \) observations the expected (theoretical) frequency of each possible event can be evaluated from the knowledge of the probability distribution suggested for the population. Let they are denoted by \( E_1, E_2, \ldots, E_n \). Our problem is to verify whether the suggestion regarding the probability distribution of the population is acceptable. That is to test the compatibility (or discrepancy) of observed and theoretical frequencies or to determine whether the deviations, if any, of the observed frequencies from the theoretical frequencies, are small enough to be regarded as due to fluctuation of random sampling or whether they indicated that the data could not have possibly come from a population giving rise to theoretical frequencies.

A measure of discrepancy existing between the observed and expected frequencies can be found by using the test statistic \( \chi^2 \) and is given by

\[
\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},
\]

which follows chi-square distribution with \((n-1)\) degrees of freedom.

Consider significance level \( \alpha \), then find \( \chi^2_\alpha \) from the table of chi-square distribution for \((n-1)\) d.f. such that, \( P(\chi^2_{(n-1)} > \chi^2_\alpha) = \alpha \).

If the calculated value of \( \chi^2 \) greater than \( \chi^2_\alpha \), it is to conclude that the data could not have possibly come from a population giving rise to theoretical frequencies.

That is our hypothesis regarding the probability distribution of the population is rejected. In other words the assumed probability distribution is not fitting in good to the population considered or the fit is not good.

**Conditions for validity of \( \chi^2 \) – test:**

1. The sample size \( n \) should be large (say > 50)
2. The theoretical frequencies of each class should be at least 5, if that of any class is less than 5, that class should be combined with the adjacent class and the corresponding frequency is added together. This process (pooling) should be repeated till the frequency in all classes are greater than 5.

3. The degree of freedom of $\chi^2$ is one less than the total number of classes. If $r$ parameters are estimated using the observations for the calculation of the theoretical frequencies, then the degree of freedom of $\chi^2$ is $n-r-1$, ($n$ is the total number of classes after pooling).

**Problem 1:** When the first proof of 392 pages of a book of 1200 pages were read, the distribution of printing mistakes were found to be as follows:

<table>
<thead>
<tr>
<th>No. of mistakes per page ($x$)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of pages ($f$)</td>
<td>275</td>
<td>72</td>
<td>30</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

**Solution:**

Let $X$ denote the number of printing mistakes per page. To fit a Poisson distribution for the given data, first to identify the parameter $\lambda$. For a Poisson random variable $X$, $E(X) = \lambda$. Hence, equate the sample mean to the population mean $E(X)$, to get an estimate of the parameter $\lambda$.

For the given data, Mean, $\bar{x} = \frac{1}{N} \sum_i f_i x_i$

$$\sum_i f_i x_i = 0 \times 275 + 1 \times 72 + \ldots + 6 \times 1 = 189, \quad N = 392$$

$$\Rightarrow \bar{x} = \frac{189}{392} = 0.482.$$

Equating sample mean and population mean, we get $\lambda = 0.482$.

Hence, $P(X = x) = \frac{e^{-0.482}0.482^x}{x!} ; x = 0, 1, 2, \ldots.$
<table>
<thead>
<tr>
<th>x</th>
<th>Observed frequency</th>
<th>$P(X = x)$</th>
<th>Expected frequency $N \times P(X=x)$ (rounded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>275</td>
<td>$e^{-0.482} \times 0.482^0/0! = 0.6175$</td>
<td>242</td>
</tr>
<tr>
<td>1</td>
<td>72</td>
<td>$e^{-0.482} \times 0.482^1/1! = 0.2976$</td>
<td>117</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>$e^{-0.482} \times 0.482^2/2! = 0.0717$</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>$e^{-0.482} \times 0.482^3/3! = 0.0115$</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>$e^{-0.482} \times 0.482^4/4! = 0.00138$</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>$e^{-0.482} \times 0.482^5/5! = 0.00013$</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>$e^{-0.482} \times 0.482^6/6! = 0.00001$</td>
<td>0</td>
</tr>
<tr>
<td>N = 500</td>
<td></td>
<td>500</td>
<td></td>
</tr>
</tbody>
</table>

Now,
From the table, $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 50.753$. We know, here $\chi^2$ follow chi-square distribution with (4-1-1) = 2 degrees of freedom. (4 classes are considered after pooling and 1 d.f. is lost due to the estimation of the parameter for calculating theoretical frequencies)

From chi-square table for 2 d.f., and for significance level $\alpha = 0.05$, $\chi^2_\alpha = 5.99$.

Here, $\chi^2 > \chi^2_\alpha$, that is the data is not matching with the hypothesis considered. Hence we reject the hypothesis that $X$ follows Poisson distribution.

**Problem 2:** A survey of 800 families with four children each revealed the following distribution:

<table>
<thead>
<tr>
<th>Number of Boys:</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Girls:</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Number of families:</td>
<td>32</td>
<td>178</td>
<td>290</td>
<td>236</td>
<td>64</td>
</tr>
</tbody>
</table>
Is this result consistent with the hypothesis that male and female births are equally probable?

Solution:

Here we have to check whether the given data support that male and female births are equally probable with probability of male \( p = 0.5 \).

Let \( X \) denote the number of boys in a family.

Expected number of families with \( x \) male children out of 800 families with 4 Childs, and with probability of male \( p = 0.5 \), can be found by \( 800 \times P(X = x) \), where \( P(X = x) \) can be get using binomial probability law as,

\[
P(X = x) = ^{4}C_{x} (0.5)^{x}(0.5)^{4-x} , x = 0,1,2,3,4.
\]

Expected number of families with no male child = \( 800 \times P(X = 0) \)

\[
= 800 \times ^{4}C_{0} (0.5)^{0}(0.5)^{4-0} = 800 \times (0.5)^{4} = 50.
\]

Expected number of families with one male child = \( 800 \times P(X = 1) \)

\[
= 800 \times ^{4}C_{1} (0.5)^{1}(0.5)^{4-1} = 800 \times 4 \times (0.5)^{4} = 200.
\]

Expected number of families with two male child = \( 800 \times P(X = 2) \)

\[
= 800 \times ^{4}C_{2} (0.5)^{2}(0.5)^{4-2} = 800 \times 6 \times (0.5)^{4} = 300.
\]

Expected number of families with two male child = \( 800 \times P(X = 3) \)

\[
= 800 \times ^{4}C_{3} (0.5)^{3}(0.5)^{4-3} = 800 \times 4 \times (0.5)^{4} = 200.
\]

Expected number of families with two male child = \( 800 \times P(X = 4) \)

\[
= 800 \times ^{4}C_{4} (0.5)^{4}(0.5)^{4-4} = 800 \times (0.5)^{4} = 50.
\]

Now, the expected and observed frequencies of number of boys is,

<table>
<thead>
<tr>
<th>Number of Boys:</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Number of families (( O_i )):</td>
<td>32</td>
<td>178</td>
<td>290</td>
<td>236</td>
<td>64</td>
</tr>
<tr>
<td>Expected Number of families (( E_i )):</td>
<td>50</td>
<td>200</td>
<td>300</td>
<td>200</td>
<td>50</td>
</tr>
<tr>
<td>No. of male births</td>
<td>Observed frequency ($O_i$)</td>
<td>Expected frequency ($E_i$)</td>
<td>$(O_i - E_i)^2$</td>
<td>$(O_i - E_i)^2 / E_i$</td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>----------------</td>
<td>---------------------</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>32</td>
<td>50</td>
<td>324</td>
<td>6.48</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>178</td>
<td>200</td>
<td>484</td>
<td>2.42</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>290</td>
<td>300</td>
<td>100</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>236</td>
<td>200</td>
<td>1296</td>
<td>6.48</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>64</td>
<td>50</td>
<td>196</td>
<td>3.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>800</td>
<td></td>
<td>19.63</td>
<td></td>
</tr>
</tbody>
</table>

From the table, $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 19.63$. We know, here $\chi^2$ follow chi-square distribution with (5-1) = 4 degrees of freedom.

From chi-square table for 4 d.f., and for significance level $\alpha = 0.05$, $\chi^2_{\alpha} = 9.49$.

Here, $\chi^2 > \chi^2_{\alpha}$, that is the data is not matching with the hypothesis considered. Hence we reject the hypothesis that male and female births are equally probable.

**Problem 3:** A sample analysis of examination results of 200 MBA’s was made. It was found that 46 students had failed, 68 secured a third division, 62 secured a second division and the rest were placed in first division. Are these figures commensurate with the general examination result which is in the ratio 4 : 3 : 2 : 1 for various categories respectively?

**Solution:**

Let us consider the null hypothesis as the data commensurate with the general examination result. Under this hypothesis our expected number of students in each category can be calculated as follows:
Expected number of failed students out of 200 = $200 \times \frac{4}{4+3+2+1} = 80$

Expected number of students with third division out of 200 = $200 \times \frac{3}{4+3+2+1} = 60$

Expected number of students with second division out of 200 = $200 \times \frac{2}{4+3+2+1} = 40$

Expected number of students with first division out of 200 = $200 \times \frac{1}{4+3+2+1} = 20$

Calculation for chi-square value follows:

<table>
<thead>
<tr>
<th>category</th>
<th>Observed frequency $(O_i)$</th>
<th>Expected frequency $(E_i)$</th>
<th>$(O_i - E_i)^2 / E_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed</td>
<td>46</td>
<td>80</td>
<td>14.45</td>
</tr>
<tr>
<td>Third divn.</td>
<td>68</td>
<td>60</td>
<td>1.067</td>
</tr>
<tr>
<td>Second divn.</td>
<td>62</td>
<td>40</td>
<td>12.1</td>
</tr>
<tr>
<td>First divn.</td>
<td>24</td>
<td>20</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td>28.417</td>
</tr>
</tbody>
</table>

From the table, $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 28.417$.

At 5\% significance level, for (4-1) = 3 d.f., from chi-square table, $\chi^2_{a} = 7.815$.

Here $\chi^2 > \chi^2_{a}$. Hence we reject the hypothesis that the data commensurate with the general examination result.

5.7 $\chi^2$ – test of independence:

Consider two characteristics A and B. To test whether A and B are independent in nature. Assume the following table giving frequency corresponds to A and B, which is divided into different classes,
Under the hypothesis, that A and B are independent,

\[ P(\text{An observation to come in } A_i\text{th class}) = \frac{f_{i_1}}{f_-} \quad \text{(where } f_- = \sum_{i,j} f_{ij}) \]

\[ P(\text{An observation to come in } B_j\text{th class}) = \frac{f_{j_1}}{f_-} \]

then, the probability of an observation to come in \( A_i^{th} \) and \( B_j^{th} \) class is \( \frac{f_{i_1} \times f_{j_1}}{f_- \times f_-} \).

Hence the expected number of observations in \( A_i^{th} \) and \( B_j^{th} \) class

\[ = \frac{f_{i_1} \times f_{j_1}}{f_- \times f_-} = \frac{f_{i_1} \times f_{j_1}}{f_-} \]

Under the hypothesis,

\[ \chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \]
\[ \chi^2 = \sum_{i,j} \left[ \frac{f_{ij} - f_{i\cdot} \times f_{\cdot j}}{f_{i\cdot} \times f_{\cdot j}} \right]^2, \] which follows chi-square distribution with \((l-1)(m-1)\) degrees of freedom. If the hypothesis that the characteristics A and B are independent is acceptable, then, calculated value of \(\chi^2\), will be close to zero. If \(\chi^2 > \chi^2_\alpha\), we reject the hypothesis that the characteristics are independent. (\(\chi^2_\alpha\) is the table value of chi-square distribution for \((l-1)(m-1)\) d.f. such that, \(P(\chi^2_{(l-1)(m-1)} > \chi^2_\alpha) = \alpha\)).

Problem 1: For a \(2 \times 2\) contingency table, where the frequencies are \(a, b, c\) and \(d\), as given by,

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>A₂</td>
<td>(c)</td>
<td>(d)</td>
</tr>
</tbody>
</table>
|     | \(a+c\) | \(b+d\) | \(N\); \(N = a + b + c + d\). Show that, \(\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)}\).

Solution:

\(\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}\). The expected frequencies of each cell is calculated as,

- Expected frequency of (1,1)th cell = \(\frac{(a+b)(a+c)}{N}\)
- Expected frequency of (1,2)th cell = \(\frac{(a+b)(b+d)}{N}\)
- Expected frequency of (2,1)th cell = \(\frac{(c+d)(a+c)}{N}\)
- Expected frequency of (1,1)th cell = \(\frac{(c+d)(b+d)}{N}\)

\(\chi^2 = \frac{(a - \frac{(a+b)(a+c)}{N})^2}{\frac{(a+b)(a+c)}{N}} + \frac{(b - \frac{(a+b)(b+d)}{N})^2}{\frac{(a+b)(b+d)}{N}} + \frac{(c - \frac{(c+d)(a+c)}{N})^2}{\frac{(c+d)(a+c)}{N}} + \frac{(d - \frac{(c+d)(b+d)}{N})^2}{\frac{(c+d)(b+d)}{N}}\)
but, \[
\frac{(a - \frac{(a+b)(a+c)}{N})^2}{\frac{(a+b)(a+c)}{N}} = \frac{1}{N} \times \frac{(a(a+b+c+d)-(a+b)(a+c))^2}{(a+b)(a+c)}
\]

\[
= \frac{1}{N} \times \frac{(ad-bc)^2}{(a+b)(a+c)}
\]

Similarly, the other terms are becomes, \[
\frac{1}{N} \times \frac{(ad-bc)^2}{(a+b)(b+d)}
\]

and \[
\frac{1}{N} \times \frac{(ad-bc)^2}{(c+d)(a+c)}
\]

Hence,

\[
\Rightarrow \chi^2 = \frac{1}{N} \left\{ \frac{(ad-bc)^2}{(a+b)(a+c)} + \frac{(ad-bc)^2}{(a+b)(b+d)} + \frac{(ad-bc)^2}{(c+d)(a+c)} + \frac{(ad-bc)^2}{(c+d)(b+d)} \right\}
\]

\[
= \frac{(ad-bc)^2}{N} \left[ \frac{1}{(a+b)(a+c)} + \frac{1}{(a+b)(b+d)} + \frac{1}{(c+d)(a+c)} + \frac{1}{(c+d)(b+d)} \right]
\]

\[
= \frac{(ad-bc)^2}{N} \left[ \frac{(b+d)+(a+c)}{(a+b)(a+c)(b+d)} + \frac{(a+c)+(b+d)}{(c+d)(a+c)(b+d)} \right]
\]

\[
= \frac{(ad-bc)^2}{N} \left[ \frac{N}{(a+b)(a+c)(b+d)} + \frac{N}{(c+d)(a+c)(b+d)} \right]
\]

\[
= (ad-bc)^2 \left[ \frac{(c+d)+(a+b)}{(a+b)(a+c)(b+d)(c+d)} \right]
\]

\[
\Rightarrow \chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(b+d)(c+d)(a+c)}
\]

**Problem 1:** Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas and the other from urban. The results are given in the adjoining table. Examine whether the nature of the area is related to voting preference in this election.
<table>
<thead>
<tr>
<th>Area</th>
<th>Votes for</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Rural</td>
<td>620</td>
<td>380</td>
</tr>
<tr>
<td>Urban</td>
<td>550</td>
<td>450</td>
</tr>
<tr>
<td>Total</td>
<td>1170</td>
<td>830</td>
</tr>
</tbody>
</table>

**Solution:**

Considering ‘the nature of the area and voting preference are independent’ as the null hypothesis. Under this hypothesis we get the expected frequency for each cell as follows.

Expected frequency for the (1,1) cell = \( \frac{1000 \times 1170}{2000} = 585 \)

Expected frequency for the (1,2) cell = \( \frac{1000 \times 830}{2000} = 415 \)

Expected frequency for the (2,1) cell = \( \frac{1170 \times 1000}{2000} = 585 \)

Expected frequency for the (2,2) cell = \( \frac{1000 \times 830}{2000} = 415 \)

The chi-square value = \( \sum \left( \frac{(O_i - E_i)^2}{E_i} \right) \)

\[ = \frac{(620 - 585)^2}{585} + \frac{(380 - 415)^2}{415} + \frac{(550 - 585)^2}{585} + \frac{(450 - 415)^2}{415} \]

\[ = 10.0891. \]

At 5% significance level, for (2-1)(2-1) = 1 d.f., from chi-square table, \( \chi^2_{0.05} = 3.841 \).
Here $\chi^2 > \chi^2_{\alpha}$. Hence we reject the hypothesis that the nature of the area and voting preference are independent. That is the nature of area is related with the voting preference.

**Problem 2:** The following contingency table gives the classification of 1000 workers in a factory according to the disciplinary action taken by the management and their promotional experience:

<table>
<thead>
<tr>
<th>Disciplinary action</th>
<th>Promotional experience</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Promoted</td>
<td>Not promoted</td>
</tr>
<tr>
<td>Offenders</td>
<td>30</td>
<td>670</td>
</tr>
<tr>
<td>Non-offenders</td>
<td>70</td>
<td>230</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>900</td>
</tr>
</tbody>
</table>

Test whether promotional experience and disciplinary action taken are associated or not.

**Solution:**

Considering ‘promotional experience and disciplinary action taken’ are independent as the null hypothesis, the expected frequency in each cell is calculated as follows.

Expected frequency for the (1,1) cell = $\frac{700 \times 100}{1000} = 70$

Expected frequency for the (1,2) cell = $\frac{900 \times 700}{1000} = 630$

Expected frequency for the (2,1) cell = $\frac{100 \times 300}{1000} = 30$

Expected frequency for the (2,2) cell = $\frac{900 \times 300}{1000} = 270$

The chi-square value = $\sum \frac{(O_i - E_i)^2}{E_i}$
\[
\chi^2 = \frac{(30-70)^2}{70} + \frac{(670-630)^2}{630} + \frac{(70-30)^2}{30} + \frac{(230-270)^2}{270} \\
= 84.66.
\]

At 5% significance level, for \((2-1)(2-1) = 1\) d.f., from chi-square table, \(\chi^2_a = 3.841\).

Here \(\chi^2 > \chi^2_a\). Hence we reject the hypothesis that promotional experience and disciplinary action taken' are independent. That is promotional experience and disciplinary actions are associated.

**Problem 3:** A group of 200 boys and 100 girls are selected for an IQ test and they are classified as given below. Examine whether there is any dependency between the intelligence levels and the gender

<table>
<thead>
<tr>
<th></th>
<th>Below average</th>
<th>Average</th>
<th>Above average</th>
<th>Genius</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>boys</td>
<td>86</td>
<td>60</td>
<td>44</td>
<td>10</td>
<td>200</td>
</tr>
<tr>
<td>girls</td>
<td>40</td>
<td>33</td>
<td>25</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>126</td>
<td>93</td>
<td>69</td>
<td>12</td>
<td>300</td>
</tr>
</tbody>
</table>

**Solution:**

Here to test whether the gender (boys and girls) and the IQ levels are independent.

Under the hypothesis the two characteristics given are independent, find the expected frequency in each cell. Then perform the chi-square test.

If gender and IQ are independent,

expected frequency in (1,1) cell = \(\frac{200\times126}{300} = 84\)

Expected frequency in (1,2)\(^a\) cell = \(\frac{200\times93}{300} = 62\)

Expected frequency in (1,3)\(^a\) cell = \(\frac{200\times69}{300} = 46\)
Expected frequency in \((1, 4)^{th}\) cell = \(\frac{200 \times 12}{300} = 8\)

Expected frequency in \((1, 5)^{th}\) cell = \(\frac{100 \times 126}{300} = 42\)

Expected frequency in \((1, 6)^{th}\) cell = \(\frac{100 \times 93}{300} = 31\)

Expected frequency in \((1, 7)^{th}\) cell = \(\frac{100 \times 69}{300} = 23\)

Expected frequency in \((1, 8)^{th}\) cell = \(\frac{100 \times 12}{300} = 4\), Now the expected frequencies

<table>
<thead>
<tr>
<th>Gender</th>
<th>Below average</th>
<th>Average</th>
<th>Above average</th>
<th>Genius</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>boys</td>
<td>84</td>
<td>62</td>
<td>46</td>
<td>8</td>
<td>200</td>
</tr>
<tr>
<td>girls</td>
<td>42</td>
<td>31</td>
<td>23</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>126</td>
<td>93</td>
<td>69</td>
<td>12</td>
<td>300</td>
</tr>
</tbody>
</table>

The calculations for the value of \(\chi^2\) are as follows,

\[
\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}
\]

<table>
<thead>
<tr>
<th>Gender</th>
<th>IQcategory</th>
<th>(O_i)</th>
<th>(E_i)</th>
<th>(O_i - E_i)</th>
<th>((O_i - E_i)^2)</th>
<th>((O_i - E_i)^2) (E_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys</td>
<td>Below average</td>
<td>86</td>
<td>84</td>
<td>2</td>
<td>4</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>60</td>
<td>62</td>
<td>-2</td>
<td>4</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td>Above average</td>
<td>44</td>
<td>46</td>
<td>-2</td>
<td>4</td>
<td>0.087</td>
</tr>
<tr>
<td></td>
<td>Genius</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>0.500</td>
</tr>
<tr>
<td>Girls</td>
<td>Below average</td>
<td>40</td>
<td>42</td>
<td>-2</td>
<td>4</td>
<td>0.095</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>33</td>
<td>31</td>
<td>2</td>
<td>4</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td>Above average</td>
<td>(\frac{25}{2} = 27)</td>
<td>(\frac{25}{2} = 27)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Genius</td>
<td>(\frac{2}{2} = 2)</td>
<td>(\frac{2}{2} = 2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>300</td>
<td>300</td>
<td></td>
<td></td>
<td>0.923</td>
</tr>
</tbody>
</table>
The calculated value of $\chi^2 = 0.923$.

The table value of $\chi^2$ at 5% significance level for $(4-1)(2-1)-1= 2$ degrees of freedom (after the pooling of last two classes in one) is 5.991.

Here $\chi^2 < \chi^2_\alpha$. Hence the hypothesis that the two characteristics are independent is accepted. That is there is no dependency between gender and IQ level.

**Problem 4:** A marketing agency gives you the following information about the age groups of the sample informants and their liking for a particular model of scooter which a company plans to introduce:

<table>
<thead>
<tr>
<th>Age group of informants</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below 20</td>
<td></td>
</tr>
<tr>
<td>20 – 39</td>
<td></td>
</tr>
<tr>
<td>40 – 59</td>
<td></td>
</tr>
<tr>
<td>Liked</td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>605</td>
</tr>
<tr>
<td>420</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Disliked</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>395</td>
</tr>
<tr>
<td>220</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>1000</td>
</tr>
<tr>
<td>640</td>
<td></td>
</tr>
<tr>
<td>160</td>
<td></td>
</tr>
</tbody>
</table>

On the basis of the above data, can it be concluded that the model appeal is independent of the age group of the informants?

**Solution:**

Considering ‘the model appeal is independent of the age group’ as the null hypothesis, the expected frequency in each cell is calculated as follows.

Expected frequency for the (1,1) cell = $\frac{605 \times 200}{1000} = 121$

Expected frequency for the (1,2) cell = $\frac{605 \times 640}{1000} = 387.2$

Expected frequency for the (1,3) cell = $\frac{605 \times 160}{1000} = 96.8$

Expected frequency for the (2,1) cell = $\frac{395 \times 200}{1000} = 79$
Expected frequency for the (2,2) cell = \( \frac{395 \times 640}{1000} = 252.8 \)

Expected frequency for the (2,3) cell = \( \frac{395 \times 160}{1000} = 63.2 \)

<table>
<thead>
<tr>
<th>Observed frequency ((O_i))</th>
<th>Expected frequency ((E_i))</th>
<th>(\frac{(O_i - E_i)^2}{E_i})</th>
</tr>
</thead>
<tbody>
<tr>
<td>125</td>
<td>121</td>
<td>0.132</td>
</tr>
<tr>
<td>420</td>
<td>387.2</td>
<td>2.77</td>
</tr>
<tr>
<td>60</td>
<td>96.8</td>
<td>13.99</td>
</tr>
<tr>
<td>75</td>
<td>79</td>
<td>0.202</td>
</tr>
<tr>
<td>220</td>
<td>252.8</td>
<td>4.25</td>
</tr>
<tr>
<td>100</td>
<td>63.2</td>
<td>21.42</td>
</tr>
</tbody>
</table>

The calculated value of chi-square = \( \sum \frac{(O_i - E_i)^2}{E_i} = 42.764 \)

At 5% significance level, for \((3-1)(2-1) = 2\) d.f., from chi-square table, \(\chi^2_{0.05} = 5.991\).

Here \(\chi^2 > \chi^2_{0.05}\). Hence we reject the hypothesis the model appeal is independent of the age group.
EXERCISES

1. How do you determine the critical region for testing the mean of a population in large sample case?

2. What is large sample test? How can the equality of two population proportions be tested?

3. Explain the principle of a test of goodness of fit.

4. What are the conditions for using chi-square test for testing agreement between theoretical frequencies and observed frequencies?

5. Derive test statistic and test procedure to test the independence of two attributes in a $2 \times 2$ contingency table.

6. An examination was given to 50 students at college A and to 60 students at college B. At A, the mean grade was 75 with standard deviation of 9 and at B the mean grade was 79 with standard deviation 7. Is there significant difference between the performance of the students at A and those at B at 5% level of significance?

7. The manufacturer of television tubes knows from past experience that the average life of a tube is 2,000 hours with a standard deviation of 200 hours. A sample of 100 tubes has an average life of 1,950 hours. Test at the 0.05 level of significance, if this sample came from a normal population of mean 2,000 hours.

8. A sample of heights of 6,400 Englishmen has a mean of 67.85 inches and S.D. 2.56 inches, while a sample of heights of 1,600 Australians has a mean of 68.55 inches and S.D. of 2.52 inches. Do the data indicate that Australians are, on the average, taller than Englishmen?

9. The mean weekly sale of soap bars in departmental stores was 146.3 bars store. After an advertising campaign the mean sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation 17.2. Was the advertising campaign successful?

10. In a locality 110 persons were randomly selected and asked about their educational achievement. The results are given as follows:

<table>
<thead>
<tr>
<th>Sex</th>
<th>below SSLC</th>
<th>SSLC</th>
<th>Above SSLC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>15</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Female</td>
<td>25</td>
<td>15</td>
<td>10</td>
</tr>
</tbody>
</table>

Can you conclude in light of this sample, education depends on sex?
6.1. Small Sample Tests:

When the number of sample is large, by central limit theorem, almost all test statistics follows normal distribution. Then for testing the hypothesis, critical region can be obtained with the help of standard normal table. But when the sample is small, it is to use test statistics with known probability distributions to perform the testing of hypothesis.

6.2. Tests based on normally distributed test statistics:

(i) To test mean of a normal population:

Let \( H_0 : \mu = \mu_0 \) is to be tested, where \( \mu \) the mean of a normal population with the known standard deviation \( \sigma \). Let a sample of size \( n \) is taken from the normal population.

Then the sample mean \( \bar{x} \) follows \( N(\mu, \frac{\sigma}{\sqrt{n}}) \). That is \( t = \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma} \sim N(0,1) \)

If \( H_1 : \mu > \mu_0 \), as illustrated in the last chapter, for a given significance level \( \alpha \), we reject \( H_0 \), if \( t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} > t_\alpha \), where \( t_\alpha \) is from standard normal table such that \( P(t > t_\alpha) = \alpha \)

If \( H_1 : \mu < \mu_0 \), as illustrated in the last chapter, for a given significance level \( \alpha \), we reject \( H_0 \), if \( t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} < -t_\alpha \), where \( t_\alpha \) is from standard normal table such that \( P(t > t_\alpha) = \alpha \)
If \( H_1 : \mu \neq \mu_0 \), as illustrated in the last chapter, for a given significance level \( \alpha \), we reject \( H_0 \), if

\[ |t| = \left| \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \right| > t_{\alpha / 2}, \]

where \( t_{\alpha / 2} \) is from standard normal table such that \( P(|t| > t_{\alpha / 2}) = \alpha \).

(ii) To test the equality means of two normal population with known standard deviations

Let samples of sizes \( n_1 \) and \( n_2 \) are taken from two normal populations \( N(\mu_1, \sigma_1) \) and \( N(\mu_2, \sigma_2) \). To test \( H_0 : \mu_1 = \mu_2 \), where \( \sigma_1^2 \) and \( \sigma_2^2 \) are known.

The test statistic used is,

\[ t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \]

following \( N(0,1) \), under \( H_0 \).

Let \( \alpha \) is the significance level, then

If \( H_1 : \mu_1 > \mu_2 \), Reject \( H_0 \), if \( t > t_{\alpha} \), where \( t_{\alpha} \) is from standard normal such that, \( P(t > t_{\alpha}) = \alpha \).

If \( H_1 : \mu_1 < \mu_2 \), Reject \( H_0 \), if \( t < -t_{\alpha} \), where \( t_{\alpha} \) is from standard normal such that, \( P(t < -t_{\alpha}) = \alpha \).

If \( H_1 : \mu_1 \neq \mu_2 \), Reject \( H_0 \), if \( |t| > t_{\alpha / 2} \), where \( t_{\alpha / 2} \) is from standard normal such that, \( P(|t| > t_{\alpha / 2}) = \alpha \).

Problem 1: A sample of size 10 taken from a normal population. The sample mean is recorded as 47. Another sample of size 15 gives its mean as 41. Can the samples be regarded as drawn from the same population of standard deviation \( \sigma = 4 \). (\( \alpha = 0.05 \)).

Solution:

Two samples of sizes 10 and 15 were drawn from a normal population. The sample means are 47 and 41 respectively. Here to test whether the samples are from normal
population with standard deviation 4. That is to test whether the data compatible with the two population means are equal under common standard deviation $\sigma = 4$.

Hence $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

The test statistic used is 
\[
t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}} = \frac{47 - 41}{\sqrt{\frac{4^2}{10} + \frac{4^2}{15}}} = 3.6742.
\]

For, $\alpha = 0.05$, from std normal table $t_{\alpha/2} = 1.96$. Hence, here $|t| > t_{\alpha/2}$. Then reject $H_0$. That is, the two samples cannot be regarded as coming from same population.

**Problem 2:** A sample of size 10 of men and another sample of size 12 of women have mean IQ’s 101 and 98 respectively. Assuming that the IQ’s of men and women are independently and normally distributed with mean $\mu_1$ and $\mu_2$ and S.D’s 4 and 3. Examine whether men are on the average more intelligent than women at 5% level of significance?

**Solution:**

The test to perform is $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$

The test statistic used is 
\[
t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}} = \frac{101 - 98}{\sqrt{\frac{4^2}{10} + \frac{3^2}{12}}} = 1.96.
\]

For 5% significance level, from standard normal table we get $t_{\alpha} = 1.65$.

Here $t > t_{\alpha}$. Hence reject $H_0$. That is accepting that men are on the average more intelligent than women.
6.3. Tests based on statistics following t - distribution:

(i) To test mean of a normal population (When \( \sigma \) is unknown):

Consider a normal population \( N(\mu, \sigma) \). To test \( H_0: \mu = \mu_0 \). If the population standard deviation is unknown and since the sample size is small, the test based on normal distribution is not applicable.

Under the assumptions: (i) the population from the sample am taken is normal

(ii) Standard deviation of the population is unknown

(iii) The sample size is small (say<30)

the statistic , \( t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s} \) which follows t-distribution with \((n-1)\) degrees of freedom is considered as the test statistic.

Let \( \alpha \) is the significance level, then,

If \( H_1 : \mu_1 > \mu_2 \), Reject \( H_0 \), if \( t > t_\alpha \), where \( t_\alpha \) is from table of t - distribution for (n-1) d.f. such that, \( P( t_{(n-1)} > t_\alpha ) = \alpha \)

If \( H_1 : \mu_1 < \mu_2 \), Reject \( H_0 \), if \( t < -t_\alpha \), where \( t_\alpha \) is from table of t - distribution for (n-1) d.f. such that, \( P( t_{(n-1)} < -t_\alpha ) = \alpha \).

If \( H_1 : \mu_1 \neq \mu_2 \), Reject \( H_0 \), if \( |t| > \frac{t_\alpha}{2} \), where \( t_\alpha \) is from table of t - distribution for (n-1) d.f. such that, \( P( |t_{(n-1)}| > \frac{t_\alpha}{2} ) = \alpha \).

(ii) To test the equality means of two normal population with known standard deviations: (when population standard deviations \( \sigma_1 \) and \( \sigma_2 \) are unknown)

Let samples of sizes \( n_1 \) and \( n_2 \) are taken from two normal populations \( N(\mu_1, \sigma_1) \) and \( N(\mu_2, \sigma_2) \), where \( \sigma_1 \) and \( \sigma_2 \) are unknown. To test \( H_0 : \mu_1 = \mu_2 \).

Let the sample standard deviations are \( S_1 \) and \( S_2 \).
The test statistic used is, \( t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \), following \( t\)-distribution with \((n_1 + n_2 - 2)\) degrees of freedom.

Let \( \alpha \) is the significance level, then

If \( H_1: \mu_1 > \mu_2 \), Reject \( H_0 \), if \( t > t_\alpha \), where \( t_\alpha \) is from t- table for \( n_1 + n_2 - 2 \) d.f. such that, \( P( t_{(n_1+n_2-2)} > t_\alpha ) = \alpha \).

If \( H_1: \mu_1 < \mu_2 \), Reject \( H_0 \), if \( t < -t_\alpha \), where \( t_\alpha \) is from t- table for \( n_1 + n_2 - 2 \) d.f. such that, \( P( t_{(n_1+n_2-2)} < -t_\alpha ) = \alpha \).

If \( H_1: \mu_1 \neq \mu_2 \), Reject \( H_0 \), if \( |t| > t_{\alpha/2} \), where \( t_\alpha \) is from t- table for \( n_1 + n_2 - 2 \) d.f. such that, \( P( |t_{(n_1+n_2-2)}| > t_{\alpha/2} ) = \alpha \).

(iii) To test the equality means based on paired observations:

If to test the equality of mean effect of two different treatments on a population, let \( n \) different pairs of units of the population in such a way that each pair should contain units of the population which are homogenous in nature is selected. Apply treatments to each of the pair such that, the first treatment \( (X) \) is to the first unit of the pair and the second treatment \( (Y) \) to the second unit.

Collect the data \( u_1, u_2, \ldots, u_n \) for each pair such that, \( u_i = x_i - y_i \), where \( x_i \) is the value of the effect of treatment \( (X) \) on the first unit of the \( i \)th pair and \( y_i \) is the value of the effect of treatment \( (Y) \) on the second unit of the \( i \)th pair.

To test \( H_0 : \mu_1 = \mu_2 \), ie., \( H_0 : \mu_1 - \mu_2 = 0 \), where \( \mu_1 \) is the mean effect of first treatment and \( \mu_2 \) is the mean effect of second. If the mean effect of two treatments is same, we expect the value of \( \bar{u} \) as zero.

Assume \( u \) follow normal distribution with mean \( \mu_1 - \mu_2 \) and unknown standard deviation then the test statistic suggested to test \( H_0 : \mu_1 - \mu_2 = 0 \), is
Let $\alpha$ be the significance level, then,

If $H_1: \mu_1 > \mu_2$, Reject $H_0$, if $t > t_{\alpha}$, where $t_{\alpha}$ is from table of t-distribution for (n-1) d.f. such that, $P(t_{(n-1)}> t_{\alpha}) = \alpha$.

If $H_1: \mu_1 < \mu_2$, Reject $H_0$, if $t < -t_{\alpha}$, where $t_{\alpha}$ is from table of t-distribution for (n-1) d.f. such that, $P(t_{(n-1)}< -t_{\alpha}) = \alpha$.

If $H_1: \mu_1 \neq \mu_2$, Reject $H_0$, if $|t| > \frac{t_{\alpha}}{2}$, where $t_{\alpha}$ is from table of t-distribution for (n-1) d.f. such that, $P(|t_{(n-1)}| > \frac{t_{\alpha}}{2}) = \alpha$.

**Problem 1:** The heights of 10 males are found to be 70, 66, 59, 68, 62, 63, 61, 60, 59, 58 inches. Is it reasonable to think that the average height is more than 62 inches? Test at 5% level of significance.

**Solution:**

Let $X$ denotes the height of the males and assume it follow normal distribution. Now to test the hypothesis regarding the mean of the population. That is to test $H_0: \mu = 62$ against $H_1: \mu > 62$.

Sample mean $\bar{x}$ of the 10 samples given can be calculated. Population S.D. is unknown.

Hence the test statistic used is, $t = \frac{(\bar{x} - \mu_0)\sqrt{n-1}}{s}$ follows t-distribution with (n-1) d.f.

Here, $\sum_i x_i = 626 \Rightarrow \bar{x} = \frac{1}{n} \sum_i x_i = \frac{626}{10} = 62.6$

<table>
<thead>
<tr>
<th>$x$</th>
<th>70</th>
<th>66</th>
<th>59</th>
<th>68</th>
<th>62</th>
<th>63</th>
<th>61</th>
<th>60</th>
<th>59</th>
<th>58</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(x_i - \bar{x})$</td>
<td>7.4</td>
<td>3.4</td>
<td>-3.6</td>
<td>5.4</td>
<td>-0.6</td>
<td>0.4</td>
<td>-1.6</td>
<td>-2.6</td>
<td>-3.6</td>
<td>-4.6</td>
</tr>
<tr>
<td>$(x_i - \bar{x})^2$</td>
<td>54.76</td>
<td>11.56</td>
<td>12.96</td>
<td>29.16</td>
<td>0.36</td>
<td>0.16</td>
<td>2.56</td>
<td>6.76</td>
<td>12.96</td>
<td>21.16</td>
</tr>
</tbody>
</table>
\[ S^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{152.4}{10} = 15.24 \Rightarrow S = 3.9 \]

The test statistic is, 
\[ t = \frac{(\bar{x} - \mu_0)\sqrt{n-1}}{S} = \frac{(62.2 - 62)\sqrt{10-1}}{3.9} = 0.154. \]

Consider 5% level of significance. Then the value of \( t_\alpha \) from table of \( t \)-distribution for 9 degrees of freedom is 1.8331.

Here \( t = 0.154 < t_\alpha \). Then we accept \( H_0 \). Hence it is not reasonable to think that the average height is greater than 62 inches.

**Problem 2:** The heights of six randomly chosen sailors are in inches: 63, 65, 68, 69, 71, and 72. Those of 10 randomly chosen soldiers are 61, 62, 65, 66, 69, 69, 70, 71, 72, and 73. Test whether the data support the claim that the soldiers are on the average taller than sailors.

**Solution:**

First group contains \( n_1 = 12 \) soldiers and the second group contains \( n_2 = 15 \) sailors.

Let \( \mu_1 \) denote the average height of soldiers and \( \mu_2 \) be that of sailors. Now to test \( H_0 : \mu_1 = \mu_2 \) against \( H_1 : \mu_1 > \mu_2 \).

The test statistics is 
\[ t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1S_1^2 + n_2S_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \]

To calculate \( \bar{x}_1, \bar{x}_2 \) and \( S_1^2, S_2^2 \). The necessary steps are as follows,

For the group of soldiers \( \sum_i x_i = 408 \Rightarrow \bar{x}_i = \frac{408}{6} = 68 \)

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>63</td>
<td>65</td>
<td>68</td>
<td>69</td>
<td>71</td>
</tr>
<tr>
<td>( (x_i - \bar{x}) )</td>
<td>-5</td>
<td>-3</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>( (x_i - \bar{x})^2 )</td>
<td>25</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>9</td>
</tr>
</tbody>
</table>
\[ n_1S_1^2 = \sum (x_i - \bar{x}_1)^2 = 60 \]

Again, for the group of sailors, \[ \sum x_i = 678 \Rightarrow \bar{x}_2 = \frac{678}{10} = 67.8 \]

<table>
<thead>
<tr>
<th>( x )</th>
<th>61</th>
<th>62</th>
<th>65</th>
<th>66</th>
<th>69</th>
<th>69</th>
<th>70</th>
<th>71</th>
<th>72</th>
<th>73</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_i - \bar{x} )</td>
<td>-6.8</td>
<td>-5.8</td>
<td>-2.8</td>
<td>-1.8</td>
<td>1.2</td>
<td>1.2</td>
<td>2.2</td>
<td>3.2</td>
<td>4.2</td>
<td>5.2</td>
</tr>
<tr>
<td>( (x_i - \bar{x})^2 )</td>
<td>46.24</td>
<td>33.64</td>
<td>7.84</td>
<td>3.24</td>
<td>1.44</td>
<td>1.44</td>
<td>4.84</td>
<td>10.24</td>
<td>17.64</td>
<td>27.04</td>
</tr>
</tbody>
</table>

\[ n_2S_2^2 = \sum (x_i - \bar{x}_2)^2 = 153.6 \]

\[ t = \frac{68 - 67.8}{\sqrt{\frac{60 + 153.6}{6} + \frac{1}{10}}} = 0.0991 \]

At 5% significance level, from t-table, for \( (6+10-2)=14 \) degrees of freedom, \( t_\alpha = 1.76 \).

Then it can be observed that, \( t < t_\alpha \). Hence we reject \( H_0 \). That is the soldiers are on average taller than sailors.

**Problem 3**: In a certain experiment to compare two types of animal foods A and B, the following results of increase in weights are observed in animals. The same sets of eight animals were used in both the foods.

<table>
<thead>
<tr>
<th>Animal number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase Food A ( (x_i) )</td>
<td>49</td>
<td>53</td>
<td>51</td>
<td>52</td>
<td>47</td>
<td>50</td>
<td>52</td>
<td>53</td>
<td>407</td>
</tr>
<tr>
<td>Increase Food B ( (y_i) )</td>
<td>52</td>
<td>55</td>
<td>52</td>
<td>53</td>
<td>50</td>
<td>54</td>
<td>54</td>
<td>53</td>
<td>423</td>
</tr>
</tbody>
</table>

Can we conclude that food B is better than food A?
Solution:

The above problem is to compare average gain in weight with two foods. Consider $\mu_1$ and $\mu_2$ are the average gain in weight by food A and food B. Let $u_i = x_i - y_i$. If food B is good, $\bar{u}$ should be negative. Hence the problem is to test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 < \mu_2$. That is to test $H_0 : \bar{u} = 0$ against $H_0 : \bar{u} < 0$. Test is performed using paired t-test.

Test statistic used is $t = \frac{(\bar{u} - 0)\sqrt{n-1}}{S_u}$.

<table>
<thead>
<tr>
<th>animals</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$</td>
<td>49</td>
<td>53</td>
<td>51</td>
<td>52</td>
<td>47</td>
<td>50</td>
<td>52</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>$y_i$</td>
<td>52</td>
<td>55</td>
<td>52</td>
<td>53</td>
<td>50</td>
<td>54</td>
<td>54</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>$u_i = x_i - y_i$</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
<td>-4</td>
<td>-2</td>
<td>0</td>
<td>-16</td>
</tr>
<tr>
<td>$u_i^2 = (x_i - y_i)^2$</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>16</td>
<td>4</td>
<td>0</td>
<td>44</td>
</tr>
</tbody>
</table>

$\bar{u} = \frac{1}{n} \sum u_i = \frac{-16}{8} = -2$

$S_u = \sqrt{\frac{1}{n} \sum (u_i)^2 - [\bar{u}]^2} = \sqrt{\frac{1}{8}(44) - (-2)^2} = 1.23$

$\Rightarrow t = \frac{(-2 - 0)\sqrt{8-1}}{1.23} = -4.30$

At 5% sig. level, from table of t-distribution for 7 degrees of freedom, $t_\alpha = 1.895$. Here calculated value of $t < -t_\alpha$. Hence reject $H_0$ that is we can conclude food B is better than food A.
6.4. Tests based on statistics following $\chi^2$ - distribution:

To test standard deviation of a normal population:

Let $H_0: \sigma = \sigma_0$, where $\sigma$ is the standard deviation of a normal population against $H_1: \sigma > \sigma_0$. Let $S$ be the sample standard deviation of a sample of size $n$ taken from the normal population. The test statistic suggested is,

$$\chi^2 = \frac{nS^2}{\sigma^2_0},$$

which follow chi-square distribution with $(n-1)$ degrees of freedom.

For a significance level $\alpha$, reject $H_0$, if,

$$\chi^2 > \chi^2_\alpha,$$

where $\chi^2_\alpha$ is from $\chi^2$ table for $(n-1) \text{ d.f.}$ such that,

$$P(\chi^2_{(n-1)} > \chi^2_\alpha) = \alpha.$$

**Problem:** It is believed that the weight of one of the product of a company is with variance greater than 0.16 gms. A sample of eleven items is taken. Their weights (in gms.) are measured as follows: 2.5, 2.3, 2.4, 2.3, 2.5, 2.7, 2.5, 2.6, 2.6, 2.7, and 2.5. Test the hypothesis at 1% level of significance.

**Solution:**

It is to test $H_0: \sigma = 0.16$ against $H_1: \sigma > 0.16$

The test statistic is, $\chi^2 = \frac{nS^2}{\sigma^2_0}$, following chi-square distribution with $(n-1)$ d.f.

Reject $H_0$, if $\chi^2 > \chi^2_\alpha$, with confidence coefficient $(1-\alpha)$.

Here, $\sum x_i = 27.6 \Rightarrow \bar{x} = \frac{1}{n} \sum x_i = \frac{27.6}{11} = 2.51$
\[ S^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{0.1891}{11} = 0.0172 \]

\[ \chi^2 = \frac{nS^2}{\sigma^2} = \frac{11 \times 0.0172}{0.16} = 1.1825 \]

At 1% significance level, from the table of chi-square distribution for 10 degrees of freedom, \( \chi^2_{0.01} = 23.2 \). Here \( \chi^2 < \chi^2_{0.01} \). Hence we accept \( H_0 \).

6.5. Tests based on statistics following F - distribution:

To test the equality of standard deviations of two normal populations:

To test \( H_0 : \sigma_1 = \sigma_2 \) against \( H_1 : \sigma_1 \neq \sigma_2 \), where \( \sigma_1 \) and \( \sigma_2 \) are the standard deviations of two normal populations. Let \( S_1 \) and \( S_2 \) be the standard deviations of the sample of sizes \( n_1 \) and \( n_2 \) is taken from the two populations.

The test statistics suggest is the ratio of the unbiased estimators of population variances, that is;

If \( \frac{n_1S_1^2}{(n_1-1)} > \frac{n_2S_2^2}{(n_2-1)} \),

\[ F = \frac{n_1S_1^2}{n_2S_2^2} = \frac{n_1(n_2-1)S_1^2}{n_2(n_1-1)S_2^2} \]
\[ F \text{ follows } F\text{-distribution with } (n_1-1, n_2-1) \text{ degrees of freedom under the null hypothesis.} \]

For a significance level \( \alpha \),

Reject \( H_0 \), if \( F > F_{\alpha} \), where \( F_{\alpha} \) is from the table of \( F\)-distribution for \( (n_1-1, n_2-1) \) degrees of freedom, such that 
\[ P(F_{(n_1-1, n_2-1)} > F_{\frac{\alpha}{2}}) = \frac{\alpha}{2}. \]

If \( \frac{n_2S_2^2}{(n_2-1)} < \frac{n_1S_1^2}{(n_1-1)} \),

Then, consider 
\[ F = \frac{n_2S_2^2}{n_1S_1^2} = \frac{n_2(n_1-1)S_2^2}{n_1(n_2-1)S_1^2}, \]

which follows \( F\)-distribution with \( (n_2-1, n_1-1) \) degrees of freedom.

Reject \( H_0 \), against \( H_1 : \sigma_1 \neq \sigma_2 \), if \( F > F_{\alpha} \), where \( F_{\alpha} \) is from the table of \( F\)-distribution for \( (n_2-1, n_1-1) \) degrees of freedom, such that 
\[ P(F_{(n_2-1, n_1-1)} > F_{\frac{\alpha}{2}}) = \frac{\alpha}{2}. \]

If \( H_1 : \sigma_1 > \sigma_2 \), reject \( H_0 \), if \( F > F_{\alpha} \), where \( P(F > F_{\alpha}) = \alpha \), and

If \( H_1 : \sigma_1 < \sigma_2 \), reject \( H_0 \), if \( F < F'_{\alpha} \), where \( P(F < F'_{\alpha}) = \alpha \).

**Problem:** Two random samples gave the following results:

<table>
<thead>
<tr>
<th>Sample</th>
<th>size</th>
<th>Sample means</th>
<th>Sample variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>14</td>
<td>9</td>
</tr>
</tbody>
</table>

Test whether the population variances are same at 10% level of significance.
Solution:

Let the two normal populations are $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$. If the two samples are from same normal population, $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$. Hence it is to test $H_0 : \mu_1 = \mu_2$ and $H_{10} : \sigma_1 = \sigma_2$.

To perform the test on $H_0 : \mu_1 = \mu_2$, we can use t-test with the small samples taken. And to perform the test on $H_0 : \sigma_1 = \sigma_2$, we can use F-test.

But to perform the t-test, the basic assumption is $\sigma_1 = \sigma_2$. Hence first we have to test $H_0 : \sigma_1 = \sigma_2$ against $H_1 : \sigma_1 \neq \sigma_2$.

$$\frac{n_1 S_1^2}{(n_1 - 1)} = \frac{10 \times 9}{(10-1)} = 10,$$ and $$\frac{n_2 S_2^2}{(n_2 - 1)} = \frac{12 \times 9}{(12-1)} = 9.82$$

Hence the test statistic is $$F = \frac{n_1 S_1^2}{n_2 S_2^2} = \frac{n_1 (n_2 - 1)}{n_2 (n_1 - 1)} S_1^2,$$ follow F-distribution with $(n_1 - 1, n_2 - 1)$ d.f.

Here calculated value of $F = \frac{10}{9.82} = 1.018$

For 10% of significance level, at $(10-1,12-1) = (9,11)$ d.f. From F-table, $F_{\alpha/2} = 2.90$.

Since calculated value is less than table value of F, we accept $H_0 : \sigma_1 = \sigma_2$.

Now to test, $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$, the test statistics is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}.$$
Here, \( t = \frac{15 - 14}{\sqrt{\frac{10 \times 9 + 12 \times 9}{10 + 12 - 2} \left( \frac{1}{10} + \frac{1}{12} \right)}} = 0.792. \)

For 10% significance level, at 20 d.f., \( t_{\alpha} = 2.086 \). It is observed that \(|t| < t_{\alpha}\).

Hence we accept \( H_0 : \mu_1 = \mu_2 \). Therefore it can be concluded that the two samples are coming from same normal population.

**EXERCISES**

1. Describe the method of testing a simple hypothesis \( H_0 : \mu = \mu_0 \) against the alternative \( H_1 : \mu \neq \mu_0 \) in a normal population \( N(\mu, 1) \) when sample size is small.

2. Stating the assumptions, explain t-test for testing mean of a normal population.

3. Stating your assumptions, explain t-test to test the equality of means of two independent normal populations.

4. Explain how t-test is used for paired comparison of difference of means.

5. Explain the uses of t-distribution and F-distribution in testing hypothesis.

6. Discuss the different applications of chi-square as a test statistic.

7. Describe the procedure for testing the equality of variances of two normal populations.

8. Two different diets ‘A’ and ‘B’ were administered to two different groups of pigs. The gains in weight by the diets are given below

   Diet ‘A’ : 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

   Diet ‘B’ : 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

Test whether the diets differ significantly as far as their effects on increasing weight is concerned.
9. Intelligence tests on groups of boys and girls gave the following results:

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean score</th>
<th>S.D.</th>
<th>Number tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys</td>
<td>70</td>
<td>20</td>
<td>250</td>
</tr>
<tr>
<td>Girls</td>
<td>75</td>
<td>15</td>
<td>150</td>
</tr>
</tbody>
</table>

Test whether there is significant difference between the average scores of boys and girls. Obtain 5% confidence interval for the difference in average scores.

10. Two independent groups of 10 children were tested to find how many digits they could repeat from memory after hearing them. The results are as follows:

- **Group A**: 8 6 5 7 6 8 7 4 5 6
- **Group B**: 10 6 7 8 6 9 7 6 7 7

Is the difference between the mean scores of the two groups significant?

11. Two random samples of size 8 and 11 drawn from two normal populations are characterized as follows:

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Sum of observations</th>
<th>Sum of squares of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>9.6</td>
<td>61.52</td>
</tr>
<tr>
<td>11</td>
<td>16.5</td>
<td>73.26</td>
</tr>
</tbody>
</table>

Examine whether the two samples came from populations having same variance at 5% level of significance.

12. The nicotine content (in mg.) of two samples of tobacco were found to be as follows:

- **Sample A**: 24 27 26 21 25
- **Sample B**: 27 30 28 31 22 36

Can it be said that the two samples come from the same normal population?
SYLLABUS

Course-III: Statistical Inference

Module 1: Sampling Distributions; Random sample from a population distribution, sampling distribution of a statistic, standard error, sampling from a normal population, sampling distributions of the sample mean and variance, Chi-square, Student’s t and F distributions- derivations, simple properties and inter relationships.


Module 3: Interval estimation: Interval estimates of mean, difference of means, variance, proportions and difference of proportions-Large and small sample cases.

Module 4: Testing of Hypothesis: Concept of testing hypothesis, simple and composite hypothesis, type I and type II errors, critical region, level of significance and power of a test. Neymann-Pearson approach-Large sample tests concerning mean, equality of means, properties, equality of proportions, Small sample tests based on t distribution for mean, equality of means and paired data. Tests based on F distribution for ratio of variances. Test based on chi square-distribution for variance, goodness of fit and for independence of attributes

Books for reference:

1. V.K.Rohatgi : An introduction to probability theory and Mathematical Statistics, Wiely Eastern

*********