SREENIVASA INSTITUTE OF TECHNOLOGY AND MANAGEMENT STUDIES
(Autonomous)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING-DATA SCIENCE
II B.Tech. - IV Sem

23CSD242T	**STATISTICAL METHODS FOR DATASCIENCE (CSD)**	**L T P C** **2 1 0 3**

PRE-REQUISITES:NIL

COURSE EDUCATIONAL OBJECTIVES:

1. Understand and distinguish between discrete and continuous random variables, and their probability density functions.
2. Explore properties and applications of Binomial, Poisson, and Normal distributions.
3. Define population, sample, parameter, and statistic, and understand characteristics of a good estimator.
4. Grasp the concepts of consistency, unbiasedness, and sufficiency in estimators.
5. Explain the importance of these concepts in statistical inference.

UNIT I: BASIC CONCEPTS (9)

Random variables (discrete and continuous), probability density functions, properties, mathematical expectation. Probability distributions: Binomial, Poisson and Normal-their properties. Population, sample, parameter and statistic; characteristics of a good estimator; Consistency – Invariance property of Consistent estimator, Sufficient condition for consistency; Unbiasedness; Sufficiency.

UNIT 2: POINT ESTIMATION (9)

Point Estimation- Estimator, Estimate, Methods of point estimation – Maximum likelihood method (the asymptotic properties of ML estimators are not included), Large sample properties of ML estimator (without proof)- applications, Method of moments, method of least squares, method of minimum chi-square and modified minimum chi-square- Asymptotic Maximum Likelihood Estimation and applications

UNIT 3: INTERVAL ESTIMATION (9)

Confidence limits and confidence coefficient; Duality between acceptance region of a test and a confidence interval; Construction of confidence intervals for population proportion (small and large samples) and between two population proportions(large samples); Confidence intervals for mean and variance of a normal population; Difference between the mean and ratio of two normal populations.

UNIT 4: TESTING OF HYPOTHESES (9)

Types of errors, power of a test, most powerful tests; Neyman-Pearson Fundamental Lemma and its applications; Notion of Uniformly most powerful tests; Likelihood Ratio tests: Description and property of LR tests - Application to standard distributions.

UNIT 5: SMALL SAMPLE TESTS (9)

Student's t-test, test for a population mean, equality of two population means, paired t-test, F- test for equality of two population variances, Chi-square test for goodness of fit and test for independence of attributes, $\chi 2$ test for testing variance of a normal distribution.

.

TOTAL HOURS: 45

COURSE OUTCOMES:

On successful completion of the course, students will be able to		**Pos**
CO1	Understand the basic concepts of Statistics.	**PO1, PO2, P03,**
CO2	Analyze the data and draw conclusion about collection of data under study using Point estimation	**PO1, PO2, P03,**
CO3	Analyze data and draw conclusion about collection of data under study using Interval estimation.	**PO1, PO2, P03, PO4**
CO4	Analyze to test various hypotheses included in theory and types of errors for large samples.	**PO1, PO2, P03, PO4**
CO5	Apply the different testing tools like t-test, F-test, chi-square test to analyze the relevant real life problems.	**PO1, PO2, P03, PO4**

TEXTBOOKS:

1. Miller and Freunds, Probability and Statistics for Engineers,7/e, Pearson, 2008.
2. Manoj Kumar Srivastava and Namita Srivastava, Statistical Inference – Testing of Hypotheses, Prentice Hall of India, 2014

REFERENCE BOOKS:

1. S.C. Gupta and V.K. Kapoor, Fundamentals of Mathematical Statistics, 11/e, Sultan Chand & Sons Publications, 2012.
2. S. Ross, a First Course in Probability, Pearson Education India, 2002.
3. W. Feller, an Introduction to Probability Theory and its Applications, 1/e, Wiley, 1968.
4. Robert V Hogg, Elliot A Tannis and Dale L.Zimmerman, Probability and Statistical Inference, 9th edition, Pearson publishers,2013.

RESOURCE WEBSITE

1. https://onlinecourses.nptel.ac.in/noc21_ma74/preview
2. https://onlinecourses.nptel.ac.in/noc22_mg31/preview

CO-PO MAPPING:

CO\PO	**PO1**	**PO2**	**PO3**	**PO4**	**PO 5**	**PO6**	**PO7**	**PO8**	**PO9**	**PO1 0**	**PO11**	**PO1 2**
CO.1	3	3	3		-	-	-	-	-	-	-	-
CO.2	3	3	3		-	-	-	-	-	-	-	-
CO.3	3	3	3	2	-	-	-	-	-	-	-	-
CO.4	3	3	3	2	-	-	-	-	-	-	-	-
CO.5	3	3	3	2	-	-	-	-	-	-	-	-
CO*	**3**	**3**	**3**	**2**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**

UNIT-1 : BASIC CONCEPTS

Introduction:

Random experiment: If an experiment is conducted any number of times under essential identical conditions, the result is not certain and is anyone of the several possible outcomes then that experiment is called Random experiment.

Eg: While throwing a dice the possible outcomes are $\{1, 2, 3, 4, 5, 6\}$ but we can get only either 1 (or) 2 (or) 3 (or) 4 (or) 5 (or) 6 but not certain outcome.

Event: The subset of a sample space is called event.

Eg: If $S = \{1, 2, 3, 4, 5, 6\}$ is a sample space, then

$E_1 = \{1, 3, 5\}$, $E_2 = \{2, 4, 6\}$, $E_3 = \{1, 5\}$ ..... are called events.

Sample space: The set of all possible outcomes is called a sample space.

Eg: While throwing a dice the sample space is $S = \{1, 2, 3, 4, 5, 6\}$.

Probability: In a random experiment, let E be an event.

Probability of E is defined by,

$$P(E) = \frac{\text{No. of favourable outcomes}}{\text{Total no. of possible outcomes}}.$$

Eg: While tossing a coin, if E is the event of getting a head, then $P(E) = \frac{1}{2}$.

→ Let No. of favourable outcomes denoted by m.
and total no. of outcomes possible be denoted by n.
then $P(E) = \frac{m}{n}$.

→ Let $\bar{E}$ denotes the event of non-happening of E, then,
$\bar{E} = n - m$, then, $P(\bar{E}) = \frac{n-m}{n} = \frac{n}{n} - \frac{m}{n} = 1 - \frac{m}{n} = 1 - P(E)$

$P(\bar{E}) = 1 - P(E)$

$P(E) + P(\bar{E}) = 1$

→ $0 \le P(E) \le 1 \Rightarrow 0 \le P(\bar{E}) \le 1$

→ If $P(E) = 1$, then the event is called sure event & certain event.

→ If $P(E) = 0$, then that event is called impossible event.

Q: Five digit numbers are formed with 0, 1, 2, 3, 4 (repetition of digits is not allowed). Find the probability of getting 2 in ten's place and 0 in the units place always.

Sol: Total no. of 5 digit nos using 0, 1, 2, 3, 4 is

$$n = 4 \times 4 \times 3 \times 2 \times 1 = 96 \quad (\text{or}) \quad 5! - 4! = 96.$$

Let E be the event of getting 2 in 10's place and 0 in units place

Then, No. of favourable outcomes of E =

$$= m = \underset{3}{\square}\ \underset{2}{\square}\ \underset{1}{\boxed{2}}\ \underset{1}{\square}\ \underset{1}{\boxed{0}} = 6.$$

$$\therefore P(E) = \frac{m}{n} = \frac{6}{96} = \frac{1}{16}.$$

Random variables (Discrete & Continuous):

Introduction:

In a random experiment, the sample space may be numerical or non-numerical (descriptive).

For example, while throwing a dice the outcomes are numerical i.e., $\{1, 2, 3, 4, 5, 6\}$. But the outcomes of tossing a coin is $\{H, T\}$ i.e., non-numerical.

It is inconvinient to deal with these numerical and non-numerical outcomes mathematically.

For ex: Head → we can assign to '0'

Tail → we can assign to '1'.

Thus we need a variable for the assigned values which we call as random variable.

Random variable: A real variable X whose value is determined by the outcome of a random experiment is called random variable.

Random variables are usually denoted by capital letters and particular values which the random variable takes are denoted by small letters.

Eg: Consider a random experiment consisting of tossing a coin twice, then the sample space,

$$S = \{HH, HT, TH, TT\} = \{s_1, s_2, s_3, s_4\}$$

Define a function with random variable, $X: S \to R$ by

$$X(S) = \text{"no. of heads"}$$

Then, $X(S_1) = X(HH) = 2$

$X(S_2) = X(HT) = 1$

$X(S_3) = X(TH) = 1$

$X(S_4) = X(TT) = 0$

$\therefore$ Range of $X = \{0, 1, 2\}$.

Types of Random variables:

Random variables are of 2 types.

1. Discrete Random variables
2. Continuous Random variable.

Discrete Random variable: A random variable which takes the values only on set $\{0, 1, 2, \dots, n\}$ is called a discrete random variable.

Ex: ① The no. of students in a class

② The no. of defectives in a sample of electric bulbs.

③ The no. of printing mistakes in each page of a book.

Continuous Random Variable: A random variable X takes all possible values in a given interval is called continuous random variable.

Ex: ① Height of the students in class

② Age of the students in class

③ Weight of the students in class

④ Temperature today at your location.

Discrete Probability Distribution Function:

Let X be random variable. The probability distribution function is defined as the probability that the outcomes of an experiment will be one of the outcomes $X(S) \leq x_i$, $x \in \mathbb{R}$.

It is denoted by $F(x)$ or $F_X(x)$, is defined by

$F(x) = F_X(x) = P(X \leq x)$.

Eg: while tossing a coin twice, $S = \{HH, HT, TH, TT\}$.

$X : S \to \mathbb{R}$, defined as $X(S) =$ "no. of heads"

$X = x_i$	0	1	2
$P(X = x_i)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

$P(X \leq 0) = \frac{1}{4}$; $P(X \leq 1) = \frac{1}{4} + \frac{2}{4} = \frac{3}{4}$;

$P(X \leq 2) = \frac{1}{4} + \frac{2}{4} + \frac{1}{4} = 1$.

Probability density function: The derivative of the probability distribution function is called probability density function and is denoted by $f_x(x)$.

i.e., $f_x(x) = \frac{d}{dx} F_x(x)$.

Expectation: The expectation or the expected value of x is denoted by $E(x)$ and is defined by

$$E(x) = \sum_{i=1}^{n} p_i x_i.$$

Mean: It is denoted by μ and is defined as

$$\mu = \frac{\sum p_i x_i}{\sum p_i} = \sum p_i x_i = E(x).$$

Variance: The variance is denoted by σ^2 and is defined by $\sigma^2 = E(x_i - \bar{x})^2$, where $\bar{x}$ is the mean.

$$= \sum p_i (x_i - \bar{x})^2$$

$$= \sum p_i (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \sum p_i (x_i^2 - 2x_i\mu + \mu^2)$$

$$= \sum x_i^2 p_i - 2\mu \sum p_i x_i + \mu^2 \sum p_i$$

$$= \sum p_i x_i^2 - 2\mu(\mu) + \mu^2(1)$$

$$= \sum p_i x_i^2 - 2\mu^2 + \mu^2$$

$$\therefore \sigma^2 = \sum p_i x_i^2 - \mu^2$$

Standard deviation: $\sigma = \sqrt{\sigma^2} = \sqrt{\sum p_i x_i^2 - \mu^2}$

Note: $E(x^2) = \sum p_i x_i^2$.

Problems:

① Let x denote the number of heads in a single toss of 4 fair coins. Determine (i) $P(x < 2)$ (ii) $P(1 < x \leq 3)$ for the following data.

x	0	1	2	3	4
$P(x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

Sol: $P(x < 2) = P(x = 0) + P(x = 1) = \frac{1}{16} + \frac{4}{16} = \frac{5}{16}$

$P(1 < x \leq 3) = P(x = 2) + P(x = 3) = \frac{6}{16} + \frac{4}{16} = \frac{10}{16} = \frac{5}{8}$.

(2) The random variable X has the probability distribution as follows

x	0	1	2	3	4	5	6	7
P(x)	0	k	2k	2k	3k	k^2	$2k^2$	$7k^2+k$

Determine (i) k (ii) $P(X<6)$, $P(X\geq 6)$, $P(0<X<5)$
(iii) mean (iv) variance.

Sol: (i) We have $\sum P(x) = 1$

$\Rightarrow 0+k+2k+2k+3k+k^2+2k^2+7k^2+k = 1$

$10k^2+9k-1=0$

$k = -1, \frac{1}{10}$

$\therefore \boxed{k = \frac{1}{10} = 0.1}$ $\because 0 \leq P(x) \leq 1$.

(ii) $P(X<6) = P(x=0)+P(x=1)+\cdots+P(x=5)$

$= 0+k+2k+2k+3k+k^2$

$= 8k+k^2$

$= \frac{8}{10} + \frac{1}{100} = 0.8+0.01$

$\boxed{P(X<6) = 0.81}$

(or) $P(X<6) = 1-P(X\geq 6)$

$= 1-P(x=6)-P(x=7$

$= 1-2k^2-7k^2-k$

$= 1-9k^2-k = 1-0.09-0$

$= 1-0.19 = 0.81$

$P(X\geq 6) = 1-P(X<6) = 1-0.81 = \boxed{0.19}$

$P(0<X<5) = P(x=1)+P(x=2)+P(x=3)+P(x=4)$

$= k+2k+2k+3k = 8k = \boxed{0.8}$

(iii) Mean, $\mu = \sum_{i=0}^{7} P_i x_i = 0+k(1)+2k(2)+2k(3)+3k(4)+k^2(5)$
$+2k^2(6)+(7k^2+k)(7)$

$= k+4k+6k+12k+5k^2+12k^2+49k^2+7k$

$\mu = 66k^2+30k = 0.66+3 = \boxed{3.66}$

(iv) Variance, $\sigma^2 = \sum P_i x_i^2 - \mu^2$

$\sigma^2 = 0+k(1)^2+2k(2)^2+2k(3)^2+3k(4)^2+k^2(5)^2+(2k^2)(6)^2+(7k^2+k)(7)^2 - \mu^2$

$= k+8k+18k+48k+25k^2+72k^2+343k^2+49k-(3.66)^2$

$= 440k^2+124k-(3.66)^2 = 4.40+12.4-(3.66)^2$

$\boxed{\sigma^2 = 3.4044}$.

(3) For the discrete probability distribution

x	0	1	2	3	4	5	6
P(x)	0	2k	2k	3k	k^2	$2k^2$	$7k^2+k$

Find (i) k (ii) mean (iii) variance.

Sol: (i) $\sum P(x) = 1 \Rightarrow 10k^2+8k-1=0$, $k = 0.1099$

(ii) Mean $\mu = \sum P_i x_i = 56k^2+21k = 2.9842$.

(iii) variance, $\sigma^2 = \sum P_i x_i^2 - \mu^2 = 318k^2+73k-8.9056$

$= 2.9581$.

Continuous Probability Distribution Function:

It is denoted by $F(x)$ and is defined as

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)\,dx.$$

Properties of continuous probability density function:

(i) $f(x) \ge 0$

(ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

(iii) Expectation or mean:

$$\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)\,dx, \quad -\infty < x < \infty$$

(or)

$$= \int_{a}^{b} x\, f(x)\,dx, \quad a < x < b.$$

(iv) Median (M):

$$\int_{a}^{M} f(x)\,dx = \int_{M}^{b} f(x)\,dx = \frac{1}{2}.$$

(v) Mode is the value of x for which $f(x)$ is maximum.

i.e., $f'(x) = 0$ and $f''(x) < 0$.

(vi) Variance, $\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$.

(vii) Mean deviation about the mean μ is $= \int_{-\infty}^{\infty} |x-\mu|\, f(x)\,dx$.

Problems:

① If a random variable has the probability density function $f(x)$ as $f(x) = \begin{cases} 2e^{-2x}, & \text{for } x > 0 \\ 0, & \text{for } x < 0 \end{cases}$,

find the probability that it will take value

(i) between 1 and 3 (ii) greater than 0.5.

Sol: (i) $P(1 \le X \le 3) = \int_{1}^{3} f(x)\,dx = \int_{1}^{3} 2e^{-2x}\,dx = 2\left[\frac{e^{-2x}}{-2}\right]_{1}^{3}$

$$= -\left[e^{-6} - e^{-2}\right] = e^{-2} - e^{-6}.$$

(ii) $P(X \ge 0.5) = \int_{0.5}^{\infty} f(x)\,dx = \int_{1/2}^{\infty} 2e^{-2x}\,dx = -2\left[\frac{e^{-2x}}{2}\right]_{1/2}^{\infty}$

$$= \frac{-2}{2}\left[e^{-\infty} - e^{-1}\right] = -\left[\frac{1}{e^{\infty}} - e^{-1}\right] = e^{-1}.$$

(2) Probability density function of a random variable x is

$$f(x) = \begin{cases} \frac{1}{2}\sin x, & \text{for } 0 \le x \le \pi \\ 0, & \text{elsewhere} \end{cases}$$

. Find the mean, mode and median of the distribution and also find the probability between 0 and $\pi/2$.

Sol: (i) Mean $= \int_{-\infty}^{\infty} x f(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{\pi} \frac{1}{2}\sin x \cdot x\,dx + \int_{\pi}^{\infty} 0\,dx.$

$$= \frac{1}{2}\int_0^{\pi} x \sin x\,dx.$$

$$= \frac{1}{2}\left[-x\cos x - \int -\cos x\,dx\right]_0^{\pi}$$

$$= \frac{1}{2}\left[-x\cos x + \sin x\right]_0^{\pi} = \frac{1}{2}(-\pi\cos\pi + \sin\pi + 0 - \sin 0)$$

$$= \frac{\pi}{2}.$$

$\int f(x)g(x)\,dx = f(x)\int g(x)\,dx - \int f'(x)\int g(x)\,dx\,dx.$

(ii) Mode: $f(x) = \frac{1}{2}\sin x$

$$f'(x) = \frac{\cos x}{2}$$

$f'(x) = 0 \Rightarrow \frac{\cos x}{2} = 0 \Rightarrow x = \pi/2$

For $x = \pi/2$, $f''(x) = -\frac{\sin x}{2}$ at $x = \pi/2$

$\Rightarrow f''(x) = -\frac{1}{2} < 0.$

$\therefore f(x)$ is maximum at $x = \pi/2$.

i.e., Mode $= \pi/2$.

(iii) Median: $\int_a^M f(x)\,dx = \int_M^b f(x)\,dx = \frac{1}{2}.$

$$\int_0^M \frac{1}{2}\sin x\,dx = \frac{1}{2} \Rightarrow \frac{1}{2}\left[-\cos x\right]_0^M = \frac{1}{2}$$

$$\Rightarrow \frac{1}{2}\left[-\cos M + \cos 0\right] = \frac{1}{2} \Rightarrow \frac{1}{2}\left[-\cos M + 1\right] = \frac{1}{2}$$

$$\Rightarrow -\cos M + 1 = 1$$

$$\Rightarrow \cos M = 0$$

$$M = \pi/2.$$

$\therefore$ Meadian $= \pi/2$.

Here we observe that Mean = Mode = Median $= \pi/2$.

(iv) $P(0 \le x \le \pi/2) = \int_0^{\pi/2} \frac{1}{2}\sin x\,dx = \frac{1}{2}\left[-\cos x\right]_0^{\pi/2} = \frac{1}{2}\left[-\cos\pi/2 + \cos 0\right]$

$$= \frac{1}{2}.$$

③ For the continuous probability distribution function

$f(x) = kx^2e^{-x}$, when $x \geq 0$, find.

(i) x (ii) Mean (iii) Variance.

Sol: (i) $\int_{-\infty}^{\infty} f(x)\,dx = 1 \Rightarrow \int_{0}^{\infty} kx^2e^{-x} = 1 \Rightarrow k\left[-x^2e^{-x} + \int 2xe^{-x}dx\right]_0^{\infty} =$

$\Rightarrow k\left[-x^2e^{-x} + 2x(-e^{-x}) + 2\int e^{-x}dx\right]_0^{\infty} = 1$

$\Rightarrow k\left[-x^2e^{-x} - 2xe^{-x} - 2e^{-x}\right]_0^{\infty} = 1$

$\Rightarrow k[-\infty - \infty - 0 + \infty + \infty + 0 + 0 + 2] = 1$

$2k = 1$

$\Rightarrow \boxed{k = \frac{1}{2}}$.

(ii) Mean $= \int_{-\infty}^{\infty} x\,f(x)\,dx = \int_{0}^{\infty} x\left(\frac{1}{2}x^2e^{-x}\right)dx$.

$= \int_{0}^{\infty} \frac{1}{2}x^3e^{-x}\,dx$

$= \frac{1}{2}\int_{0}^{\infty} x^3e^{-x}\,dx$

$= \frac{1}{2}\left[x^3\left(\frac{e^{-x}}{-1}\right) - (3x^2)\left(\frac{e^{-x}}{1}\right) + (6x)\left(\frac{e^{-x}}{-1}\right) - 6\left(\frac{e^{-x}}{1}\right)\right]_0^{\infty}$

$\int uv = [uv_1 - u'v_2 + u''v_3 - u'''v_4 + \cdots]$

when the derivative tends to 0. i.e., $x^3, 3x^2, 6x, 6, 0$

Bernoulli's rule

$= \frac{1}{2}[-6](e^{-\infty} - e^{-0})$

$= -3\left(\frac{1}{\infty} - \frac{1}{1}\right) = 3$.

(iii) Variance $= \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$

$= \int_{0}^{\infty} x^2 \cdot \frac{1}{2}x^2e^{-x} - \mu^2$

$= \frac{1}{2}\left[x^4\frac{e^{-x}}{-1} - (4x^3)\frac{e^{-x}}{1} + 12x^2\frac{e^{-x}}{-1} - 24x\frac{e^{-x}}{1} + 24\frac{e^{-x}}{-1}\right]_0^{\infty} - \mu^2$

$= \frac{1}{2}[-24] - 9 = 12 - 9 = 3$.

Binomial distribution: For a discrete random variable x, Binomial distribution is a discrete probability distribution defined as

$$P(X=r) = p(r) = \begin{cases} {}^nC_r\, p^r q^{n-r}, & r = 0, 1, 2, \ldots, n \\ 0, & \text{otherwise.} \end{cases}$$

where n = no. of trials.
p = probability of success of the event
q = probability of failure $= 1-p$.
r = no. of times the event occured.

Ex: ① The number of defective bolts in a box containing 'n' bolts.
② The number of post graduates in a group of 'n' men.

Conditions of BD:
1. Trials are repeated 'n' times.
2. There are only 2 possible outcomes
   i.e., success or failure for each trial.

Mean of BD: $\mu = E(X) = np$.
Variance of BD: $\sigma^2 = npq$
Standard deviation of BD: $\sigma = \sqrt{npq}$.

Derivation of mean of Binomial Distribution:

$$\text{Mean} = \sum p_i x_i = \sum_{r=0}^{n} x_r\, {}^nC_r\, p^r q^{n-r}.$$

$$= 0\times({}^nC_0\, p^0 q^n) + 1\times({}^nC_1\, p^1 q^{n-1}) + 2\times({}^nC_2\, p^2 q^{n-2}) + 3\times({}^nC_3 p^3 q^{n-3}) + \cdots + n\times({}^nC_n\, p^n q^0).$$

$$= 0 + 1\times npq^{n-1} + 2\times\frac{n(n-1)}{2\times1} p^2 q^{n-2} + 3\times\frac{n(n-1)(n-2)}{3\times2\times1} p^3 q^{n-3} + \cdots + n\times p^n$$

$$= npq^{n-1} + n(n-1)p^2q^{n-2} + \frac{n(n-1)(n-2)}{2\times1} p^3 q^{n-3} + \cdots + np^n.$$

$$= np\left[q^{n-1} + (n-1)pq^{n-2} + \frac{(n-1)(n-2)}{2\times1} p^2 q^{n-3} + \cdots + p^{n-1}\right]$$

$$= np\left[q^{n-1} + {}^{n-1}C_1\, pq^{n-2} + {}^{n-1}C_2\, p^2 q^{n-3} + \cdots + p^{n-1}\right]$$

$$= np\,(q+p)^{n-1}$$

$$= np\,(1)^{n-1}$$

$\therefore$ Mean, $\mu = np$.

Problems:

① A fair coin is tossed 6 times. Find the probability of getting four heads using Binomial distribution.

Sol: While tossing a coin, if head is the outcome.

p = propability of success = $1/2$.

q = probability of failure = $1/2$.

Here $n = 6$, $r = 4$.

$P(r) = {}^nC_r\, p^r q^{n-r}$

$P(4) = {}^6C_4\, p^4 q^{6-4} = {}^6C_2\, (1/2)^4 (1/2)^2$

$= \frac{6 \times 5}{2 \times 1} \times \left(\frac{1}{2}\right)^6 = 3 \times 5 \times \frac{1}{64}$

$= 0.2344.$

② Determine the probability of getting the sum 6 exactly 3 times in 7 throws in with a pair of fair Dice.

Sol: While throwing a pair of fair dice,

the outcomes that have sum 6 exactly are

$(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)$.

No. of possible outcomes = 5.

p = probability of success in one throw = $\frac{5}{36}$.

$q = 1 - p = 1 - \frac{5}{36} = \frac{31}{36}$.

Here $n = 7$, $r = 3$.

$P(r) = {}^nC_r\, p^r q^{n-r}$.

$P(3) = {}^7C_3 \left(\frac{5}{36}\right)^3 \left(\frac{31}{36}\right)^4 = 0.0516.$

③ In eight throws of a die 5 & 6 is considered a success. Find the mean number of successes, variance and standard deviation.

Sol: $n = 8$, $p = \frac{2}{6} = \frac{1}{3}$, $q = \frac{2}{3}$.

Mean, $\mu = np = 8\left(\frac{1}{3}\right) = \frac{8}{3} = 2.6667$.

Variance, $\sigma^2 = npq = 8\left(\frac{1}{3}\right)\left(\frac{2}{3}\right) = \frac{16}{9} = 1.77778$.

S.D $= \sqrt{\frac{16}{9}} = \frac{4}{3} = 1.3333 \ldots$

(4) A dice is thrown 6 times. If getting an even number is a success, find the probabilities of (i) atleast one success (ii) ≤3 success (iii) 4 successes.

Sol: $p = \frac{1}{2}, q = \frac{1}{2}, n = 6$

$P(r) = {}^nC_r p^r q^{n-r}$.

(i) $P(r \geq 1) = 1 - P(r=0) = 1 - P(0) = 1 - {}^6C_0 (\frac{1}{2})^0 (\frac{1}{2})^{6-0} = 1 - \frac{1}{64} = \frac{63}{64}$

$\therefore P(r \geq 1) = 0.9844$

(ii) $P(r \leq 3) = P(r=0) + P(r=1) + P(r=2) + P(r=3)$

$= {}^6C_0 (\frac{1}{2})^0 (\frac{1}{2})^{6-0} + {}^6C_1 (\frac{1}{2})^1 (\frac{1}{2})^{6-1} + {}^6C_2 (\frac{1}{2})^2 (\frac{1}{2})^{6-2}$

$+ {}^6C_3 (\frac{1}{2})^3 (\frac{1}{2})^{6-3}$

$= (\frac{1}{2})^6 [{}^6C_0 + {}^6C_1 + {}^6C_2 + {}^6C_3] = \frac{1}{64}[1 + 6 + 15 + 20]$

$= \frac{42}{64} = 0.6562.$

(iii) $P(r=4) = {}^6C_4 (\frac{1}{2})^4 (\frac{1}{2})^{6-4} = \frac{15}{64} = 0.2344$

Recurrence relation for the B.D:

$P(r) = {}^nC_r p^r q^{n-r}$

$P(r+1) = {}^nC_{r+1} p^{r+1} q^{n-(r+1)}$

$\frac{P(r+1)}{P(r)} = \frac{{}^nC_{r+1} p^{r+1} q^{n-r-1}}{{}^nC_r p^r q^{n-r}}$

$\Rightarrow \frac{P(r+1)}{P(r)} = \frac{\frac{n!}{(n-(r+1))!(r+1)!} p^r \cdot p \cdot q^{n-r} \cdot q^{-1}}{\frac{n!}{(n-r)! r!} p^r \cdot q^{n-r}}$

$= \frac{(n-r)(n-r-1)! r!}{(n-r+1)! (r+1) r!} \cdot \frac{p}{q}$

$\frac{P(r+1)}{P(r)} = \frac{n-r}{r+1} \cdot \frac{p}{q}$

$\therefore \boxed{P(r+1) = \left(\frac{n-r}{r+1}\right) \frac{p}{q} \cdot P(r)}$

Binomial frequency distribution:

Expected (or) theoritical frequency,

$f(x) = N P(x)$

where N = sum of frequency of each success = Σf

$P(r) = {}^nC_r p^r q^{n-r}$.

(or)

$N(q+p)^n$.

The possible number of successes and their frequencies is called a Binomial Frequency distribution

Q: Fit a BD to the following data

x	0	1	2	3	4	5
f	2	14	20	24	22	8

Sol: Here $n = 5$,

$$N = \Sigma f = 2+14+20+24+22+8 = 100$$

$$\text{Mean} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{2(0)+14(1)+20(2)+24(3)+22(4)+8(5)}{100}$$

$$= \frac{284}{100} = 2.84.$$

Mean of BD $= np$

$\therefore np = 2.84$

$5p = 2.84$

$p = 0.568.$

Hence, the BD is fitted as,

$$N(q+p)^n = 100\,(0.432+0.568)^5$$

$$= 100\left[{}^5C_0\,(0.568)^0\,(0.432)^{5-0} + {}^5C_1\,(0.568)^1\,(0.432)^{5-1} + {}^5C_2\,(0.568)^2\,(0.432)^{5-2} + {}^5C_3\,(0.568)^3\,(0.432)^{5-3} + {}^5C_4\,(0.568)^4\,(0.432)^{5-4} + {}^5C_5\,(0.568)^5\,(0.432)^0\right]$$

$$= 100\,[0.015 + 0.0989 + 0.260 + 0.341 + 0.224 + 0.059]$$

$$= 1.5 + 9.89 + 26 + 34.1 + 22.4 + 5.9$$

$$= 2 + 10 + 26 + 34 + 22 + 6.$$

The respective terms of the binomial give the theoritical frequencies.

Since frequencies are always integers.

$\therefore$ By converting them to the nearest integers, we get 2, 10, 26, 34, 22, 6.

x	0	1	2	3	4	5
f	2	14	20	24	22	8
Expected (or) Theoritical frequency	2	10	26	34	22	6.

Q: Four coins are tossed 160 times. The number of times x heads occcur is given below.

x	0	1	2	3	4
No. of times	8	34	69	43	6

Fit a BD to the data on the hypothesis that coins are unbiased.

Sol: $p = \frac{1}{2}, q = \frac{1}{2}, n = 4$

$N = \sum f = 8+34+69+43+6 = 160$

By the BD $p(x) = {}^nC_x p^x q^{n-x}$

$\Rightarrow P(0) = {}^4C_0 (\frac{1}{2})^0 (\frac{1}{2})^{4-0} = (\frac{1}{2})^4 = \frac{1}{16}$.

we have the recurrence relation,

$P(x+1) = \frac{(n-x)}{x+1} \frac{p}{q} \cdot P(x) = \frac{4-x}{x+1} \cdot P(x)$.

No. of heads	Observed frequency	Probability, $P(x)$	Expected or theoritical frequency $f(x) = N p(x)$
0	8	$P(0) = \frac{1}{16}$	$f(0) = NP(0) = 160 \times \frac{1}{16} = 10$
1	34	$P(1) = P(0+1) = \frac{4-0}{0+1} P(0) = \frac{1}{4}$	$f(1) = NP(1) = 160 \times \frac{1}{4} = 40$
2	69	$P(2) = P(1+1) = \frac{4-1}{1+1} P(1) = \frac{3}{8}$	$f(2) = NP(2) = 160 \times \frac{3}{8} = 60$
3	43	$P(3) = P(2+1) = \frac{4-2}{2+1} P(2) = \frac{1}{4}$	$f(3) = NP(3) = 160 \times \frac{1}{4} = 40$
4	6	$P(4) = P(3+1) = \frac{4-3}{3+1} P(3) = \frac{1}{16}$	$f(4) = NP(4) = 160 \times \frac{1}{16} = 10$

x	0	1	2	3	4
Observed Frequency	8	34	69	43	6
Expected or theoretical frequency.	10	40	60	40	10

Poisson Distribution: For a discrete random variable, the Poisson Distribution is defined as

$$P(X=r) = P(r) = \begin{cases} \frac{e^{-\eta}\eta^r}{r!}, & r = 0, 1, 2, 3, \ldots \\ 0, & \text{otherwise} \end{cases}$$

where $\eta > 0$.

Note: For rare events we use PD.

Eg: ① No. of persons born blind per year in a large city.
② No. of defective bolts manufactured by reputed company.

Conditions:
1. No. of trials (n) is large
2. Probability of success (p) is very small (close to zero)
3. $np = \eta$ is finite.

Properties:

① Sum of the probabilities = 1

i.e., $\sum_{r=0}^{\infty} P(r) = 1$

$$LHS = \sum_{r=0}^{\infty} P(r) = \sum_{r=0}^{\infty} e^{-\eta}\frac{\eta^r}{r!} = e^{-\eta}\left[1 + \frac{\eta}{1!} + \frac{\eta^2}{2!} + \frac{\eta^3}{3!} + \cdots\right]$$

$$= e^{-\eta} \cdot e^{\eta} = 1 = RHS.$$

② Mean of PD $= \sum_{r=0}^{\infty} rP(r)$

$$= \sum_{r=0}^{\infty} r \cdot \frac{e^{-\eta}\eta^r}{r!}$$

$$= \sum_{r=0}^{\infty} \cancel{r} \cdot \frac{e^{-\eta}\eta^r}{\cancel{r}(r-1)!}$$

$$= e^{-\eta} \sum_{r=0}^{\infty} \frac{\eta^r}{(r-1)!} \qquad \text{Put } r-1 = t, \; r = t+1$$

$$= e^{-\eta} \sum_{t=0}^{\infty} \frac{\eta^{t+1}}{t!}$$

$$= e^{-\eta} \cdot \eta \sum_{t=0}^{\infty} \frac{\eta^t}{t!}$$

$$= e^{-\eta} \cdot \eta \cdot e^{\eta}$$

$\therefore$ Mean of PD $= \eta$

③ Variance of PD $= \eta$ $\left[\sum_{r=0}^{\infty} r^2 P(r) - \mu^2 = \eta\right]$

④ Mode of PD is $\eta - 1$ and η.

Recurrence relation for the PD:

$$P(r) = \frac{e^{-\lambda}\lambda^r}{r!}, \quad P(r+1) = \frac{e^{-\lambda}\lambda^{r+1}}{(r+1)!}$$

$$\frac{P(r+1)}{P(r)} = \frac{\frac{e^{-\lambda}\lambda^{r+1}}{(r+1)!}}{\frac{e^{-\lambda}\lambda^r}{r!}} = \frac{e^{-\lambda}\lambda^r \cdot \lambda}{(r+1)\,r!} \times \frac{r!}{e^{-\lambda}\lambda^r} = \frac{\lambda}{r+1}$$

$$P(r+1) = \frac{\lambda}{r+1} P(r)$$

Poisson frequency distributions:

To find expected or theoretical frequency we use,

$$f(r) = NP(r).$$

Problems ① If the probability that an individual suffers a bad reaction from a certain injection is 0.001. Determine the probability that out of 2000 individuals (i) exactly 3 (ii) more than 2 individuals (iii) none (iv) more than one individual suffer a bad reaction.

Sol: $p = 0.001$ (very small), $n = 2000$ (very large)

$\lambda = np = (0.001)(2000) = 2$ (finite)

∴ Poisson distribution is applicable.

By PD, $P(r) = \frac{e^{-\lambda}\lambda^r}{r!} = \frac{e^{-2}2^r}{r!}$

(i) $P(3) = \frac{e^{-2}2^3}{3!} = 0.1804$

(ii) $P(r>2) = P(3) + P(4) + \cdots + P(2000)$

$= 1 - [P(0) + P(1) + P(2)]$

$= 1 - \left[\frac{e^{-2}2^0}{0!} + \frac{e^{-2}2^1}{1!} + \frac{e^{-2}2^2}{2!}\right] = 0.3233$

(iii) $P(0) = \frac{e^{-2}2^0}{0!} = 0.1353$

(iv) $P(r>1) = 1 - [P(0) + P(1)] = 0.594$

② If a bank received on the average of 6 bad cheques per day, find the probability that it will receive 4 bad cheques on any given day.

Sol: $\lambda = 6$, $P(\lambda) = \frac{e^{-\lambda}\lambda^r}{r!}$

$$P(4) = \frac{e^{-6}6^4}{4!} = 0.1389$$

③ Using recurrence relation formula find the probabilities when $x = 0, 1, 2, 3, 4$ and 5, if the mean of the PD is 3.

Sol: Given, $\lambda = 3$

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-3}3^x}{x!}$$

$$P(0) = \frac{e^{-3}3^0}{0!} = 0.0498.$$

By recurrence relation.

$$P(x+1) = \frac{\lambda}{x+1}P(x) \Rightarrow P(x+1) = \frac{3}{x+1}P(x)$$

For $x=0$, $P(1) = P(0+1) = \frac{3}{0+1}P(0) = 3(0.0498) = 0.1494$.

$$P(2) = P(1+1) = \frac{3}{1+1}P(1) = \frac{3}{2}(0.1494) = 0.2241$$

$$P(3) = P(2+1) = \frac{3}{2+1}P(2) = \frac{3}{3}(0.2241) = 0.2241$$

$$P(4) = P(3+1) = \frac{3}{3+1}P(3) = \frac{3}{4}(0.2241) = 0.1681$$

$$P(5) = P(4+1) = \frac{3}{4+1}P(4) = \frac{3}{5}(0.1681) = 0.1008.$$

④ Fit a Poisson Distribution for the following data and calculate the expected frequencies.

x	0	1	2	3	4
f	109	65	22	3	1

Sol: Here $N = \Sigma f = 109+65+22+3+1 = 200$

$$\text{Mean} = \lambda = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{0(109)+1(65)+2(22)+3(3)+4(1)}{200}$$

$$= \frac{122}{200} = 0.61$$

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0.61}(0.61)^x}{x!}.$$

[2 Tables from next page].

Since frequencies are always integers.

∴ By converting into nearest integers we get

109, 66, 20, 4, 1.

x	f	$P(x)$	Expected frequencies $f(x) = NP(x)$
0	109	$P(0) = \frac{e^{-0.61}(0.61)^0}{0!} = 0.5433$	$f(0) = NP(0)$ $= 200(0.5433)$ $= 108.67$
1	65	$P(1) = \frac{e^{-0.61}(0.61)^1}{1!} = 0.3314$	$f(1) = NP(1) = 200(0.3314)$ $= 66.288$
2.	22	$P(2) = \frac{e^{-0.61}(0.61)^2}{2!} = 0.1010$	$f(2) = NP(2) = 200(0.01010)$ $= 20.218$
3.	3	$P(3) = \frac{e^{-0.61}(0.61)^3}{3!} = 0.0205$	$f(3) = NP(3) = 200(0.0205)$ $= 4.111$
4.	1	$P(4) = \frac{e^{-0.61}(0.61)^4}{4!} = 0.0031$	$f(4) = NP(4) = 200(0.0031)$ $= 0.6269$

Since frequencies are always integers,

∴ By converting them to nearest integers we get 109, 66, 20, 4, 1.

x	0	1	2	3	4
f	109	65	22	3	1
Expected Frequency	109	66	20	4	1

(5) The distribution of typing mistakes committed by typist is given below. Assuming the distribution be poisson, find the expected frequencies.

x	0	1	2	3	4	5
f	42	33	14	6	4	1

Sol: $N = \Sigma f = 42 + 33 + 14 + 6 + 4 + 1 = 100$

$$\text{Mean} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{0(42) + 1(33) + 2(14) + 3(6) + 4(4) + 5(1)}{100} = \frac{100}{100} = 1.$$

$$\Rightarrow \lambda$$

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-1}(1)^x}{x!} = \frac{e^{-1}}{x!}.$$

By Recurrence relation,

$$P(x+1) = \frac{\lambda}{x+1} P(x) = \frac{1}{x+1} P(x)$$

x	f	$P(x)$	Expected Frequency $f(x) = N \cdot P(x)$.
0	42	$P(0) = \frac{e^{-1}}{0!} = \frac{1}{e} = 0.3678$	$f(0) = NP(0) = 100(0.3678) = 36.78$
1	33	$P(1) = P(0+1) = \frac{1}{0+1}P(0) = 0.3678$	$f(1) = NP(1) = 100(0.3678) = 36.78$.
2	14	$P(2) = P(1+1) = \frac{1}{1+1}P(1) = 0.1839$	$f(2) = NP(2) = 100(0.1839) = 18.39$
3	6	$P(3) = P(2+1) = \frac{1}{2+1}P(2) = 0.0613$	$f(3) = NP(3) = 100(0.0613) = 6.13$
4	4	$P(4) = P(3+1) = \frac{1}{3+1}P(3) = 0.0153$	$f(4) = NP(4) = 100(0.0153) = 1.53$
5	1	$P(5) = P(4+1) = \frac{1}{4+1}P(4) = 0.0030$	$f(5) = NP(5) = 100(0.0030) = 0.30$

Since frequencies are always integers
∴ By converting them to nearest integers we get 37, 37, 18, 6, 2, 0.

x	0	1	2	3	4	5
f	42	33	14	6	4	1
Expected Frequency	37	37	18	6	2	0

Normal Distribution:

Binomial and Poisson distributions are discrete distributions but normal distribution is continuous distribution. ND was first discovered by English Mathematician De-Moivre (1733) and further it is developed by French Mathematician Laplace (1774) and independently by Karl Friedrich Gauss (1777-1855). ND is also known as Gaussian Distribution.

Defn: A random variable X is said to have a ND and its density function is

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty, \; -\infty < \mu < \infty, \; \sigma > 0 \quad \text{---(1)}$$

(or)

$$f = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad \left[\because \text{Let } \frac{x-\mu}{\sigma} = z\right]$$

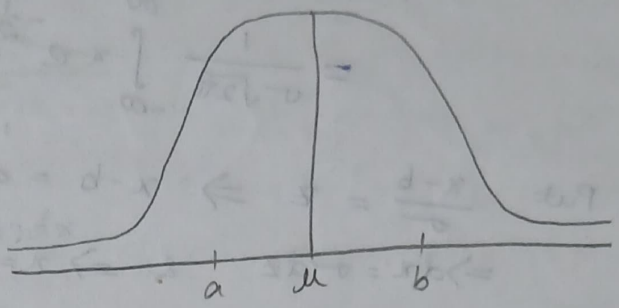
where μ = mean, σ = Standard Deviation.

Properties:

① The random variable x which follows ND is said to be normal random variable (or) normal variate. The eqn ①, represents normal curve and the total area bounded by the curve and the x-axis is 1.

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$P(a<x<b) = \int_{a}^{b} f(x)\,dx.$$



② ND is limiting case of the BD under the following conditions

(i) the no. of trials (n) is independently large i.e., $n \to \infty$.

(ii) Neither p nor q is very small

By B.D, $P(x) = {}^{n}C_x\, p^x q^{n-x}$, $x = 0, 1, 2, \ldots, n$.

Let us consider the variate

$$z = \frac{x-\mu}{\sigma}$$

$$= \frac{x-np}{\sqrt{npq}} \qquad x = 0, 1, 2, \ldots, n.$$

$$\underline{x=0} \Rightarrow z = \frac{-np}{\sqrt{npq}} = \frac{-\sqrt{np}\,\sqrt{np}}{\sqrt{np}\,\sqrt{q}} = -\sqrt{\frac{np}{q}}.$$

$$n \to \infty \Rightarrow z \to -\infty.$$

$$\underline{x=n} \Rightarrow z = \frac{n-np}{\sqrt{npq}} = \frac{n(1-p)}{\sqrt{npq}} = \frac{nq}{\sqrt{nq}\,p} = \sqrt{\frac{nq}{p}}$$

$$n \to \infty \Rightarrow z \to \infty.$$

$\therefore$ For $n \to \infty$, z takes the values $-\infty$ to ∞.

Hence the distribution of x will be continuous distribution over the range $-\infty$ to ∞.

③ Mean of ND:

Consider the normal distribution, with b and σ are parameters

$$f(x, b, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-b}{\sigma}\right)^2}.$$

$$\text{Mean} = \mu = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$$

$$= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-b}{\sigma}\right)^2} dx.$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x\, e^{-\frac{1}{2}\left(\frac{x-b}{\sigma}\right)^2} dx$$

Put $\frac{x-b}{\sigma} = z \Rightarrow x - b = \sigma z$

$\Rightarrow dx = \sigma\, dz$ & $\Rightarrow x = b + \sigma z$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (b + \sigma z)\, e^{-\frac{z^2}{2}} (\sigma\, dz)$$

$$= \frac{b}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\, e^{-\frac{z^2}{2}} dz$$

(Even function) (Odd function)

$$= \frac{2b}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{z^2}{2}} dz + 0.$$

$$= \frac{2b}{\sqrt{2}\sqrt{\pi}} \frac{\sqrt{\pi}}{\sqrt{2}} = b.$$

$\therefore$ Mean $= \mu = b$.

(4) Variance of ND $= \sigma^2$ and SD $= \sigma$

(5) Mean = Median = Mode $= \mu$

(6) Mean deviation $= \frac{4\sigma}{5}$.

Standard ND: The Normal distribution for $\mu = 0$ and $\sigma = 1$ is called Standard Normal distribution.

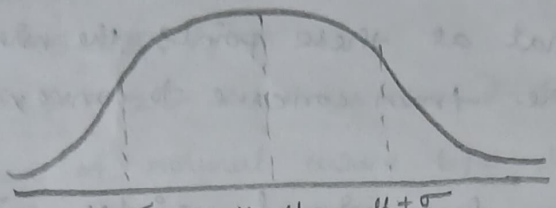
i.e., Z is Standard Normal variate

where $Z = \frac{X - \mu}{\sigma}$, μ = mean, σ = S.D.

[i.e., $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ i.e., $f(z) \simeq N(0,1)$

$$P(a < x < b) = \int_{a}^{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

is not possible to integrate. But we can solve by Numerical integration, then we get only approximate solution not exact sol. For different values of 'x' along x-axis we can make approximate values and they are represented in the table.



$$z = \frac{x-\mu}{\sigma}.$$

$$P(\mu-\sigma < X < \mu+\sigma) = \int_{\mu-\sigma}^{\mu+\sigma} f(x)\,dx$$

$$P(-1 < z < 1) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-z^2/2}\,dz \quad \text{(Even func}^{c}\text{)}$$

$$= \frac{2}{\sqrt{2\pi}} \int_{0}^{1} e^{-\frac{z^2}{2}}\,dz$$

$$= 2 \int_{0}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\,dz$$

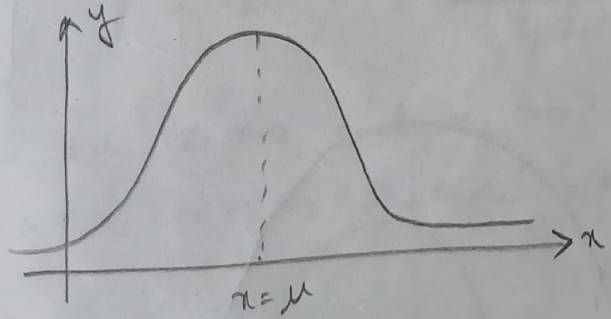
$$= 2 \int_{0}^{1} \phi(z)\,dz.$$

$$= 2(0.3413) \text{ from table}$$

$$= 0.6826.$$

Characteristics of Normal Distributions:

① The graph of ND, $y = f(x)$ in the xy-plane is called as normal curve.



② The curve is bell shaped and symmetrical about the line $x=\mu$ and the 2 tails on the right and the left sides of the mean(μ) extends to infinity.

③ Area under the curve normal curve represents the total Population.

④ At $x=\mu$, mean = median = mode (coincide). So normal curve has only one maximum point.

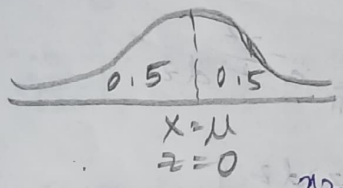
(5) x-axis is asymptode to the curve i.e, the curve and x-axis do not touch as it tends to ∞ but it comes closer to x-axis.

(6) The points of inflection to the normal curve are at $x = \mu \pm \sigma$. This means that at these points the normal curve changes its curvature from concave to convex and vice versa.

(7) If x and y are independent normal variates with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2 respectively, then $(x+y)$ is also a normal variate with mean $(\mu_1 + \mu_2)$ and variance $(\sigma_1^2 + \sigma_2^2)$.

(8) Area of the normal curve where $X \geq \mu$ is 0.5.
Area of the normal curve where $X \leq \mu$ is 0.5.

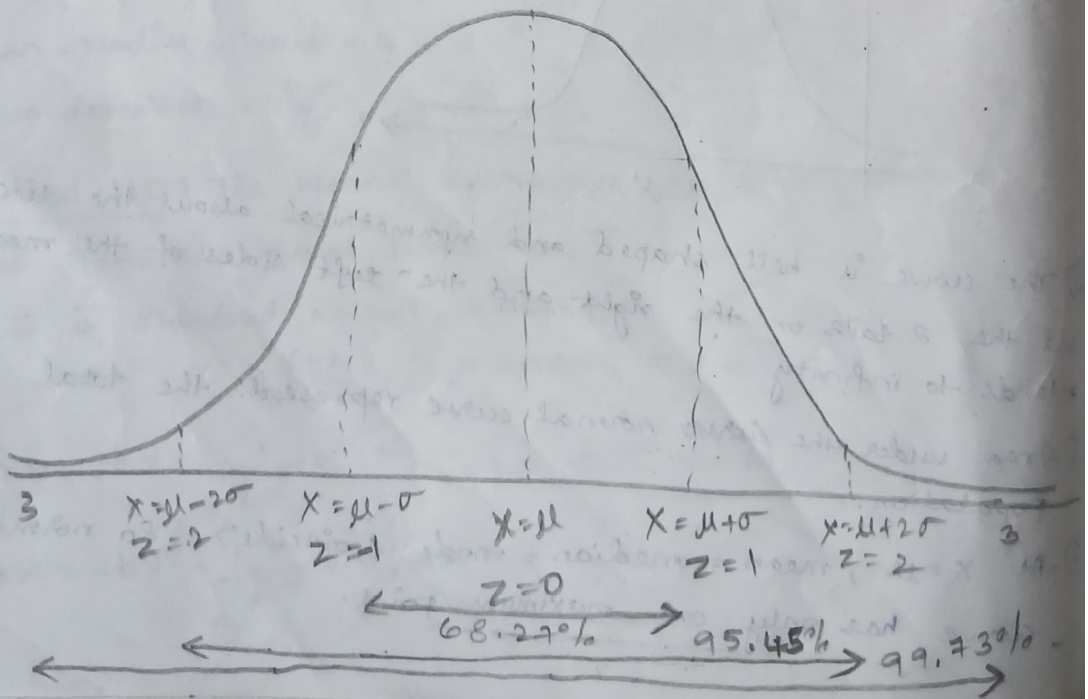
(9)



(9) $$P(x_1 \leq X \leq x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx. \quad \text{---(1)}$$

Eqn (1) depends on two parameters μ and σ, we get different values of μ and σ and it is an impracticable task to plot all such normal curves, so we put $Z = \frac{x-\mu}{\sigma}$, Here Z is called as the standard variate.

(10)



(i) Area of normal curve b/w $\mu-\sigma$ and $\mu+\sigma$ is 68.27%.

$P(\mu-\sigma < X < \mu+\sigma) = P(-1 < Z < 1) = 0.6827$

(ii) Area of normal curve b/w $\mu-2\sigma$ and $\mu+2\sigma$ is 95.45%.

$P(\mu-2\sigma < X < \mu+2\sigma) = P(-2 < Z < 2) = 0.9543$

(iii) Area of normal curve b/w $\mu-3\sigma$ and $\mu+3\sigma$ is 99.73%

$P(\mu-3\sigma < X < \mu+3\sigma) = P(-3 < Z < 3) = 0.9973.$

How to find Probability density of normal curve:

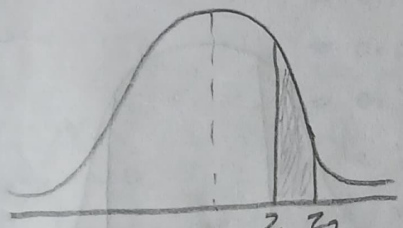
Step 1: Find z_1 and z_2 corresponding to the values of x_1 & x_2 by using $Z = \frac{x-\mu}{\sigma}$.

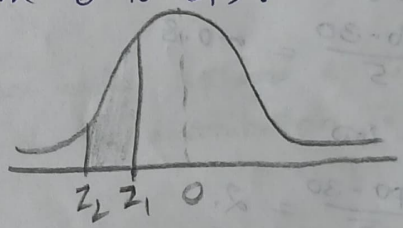
Step 2: To find $P(x_1 \le X \le x_2) = P(z_1 \le Z \le z_2)$.

Case 1: If both z_1 and z_2 are +ve (or -ve) then

$P(x_1 \le X \le x_2) = |A(z_2) - A(z_1)|$

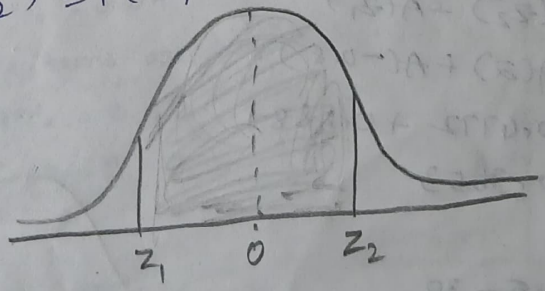
= (Area under the normal curve from 0 to z_2) − (Area under the normal curve from 0 to z_1).





Case 2: If $z_1 < 0$ and $z_2 > 0$

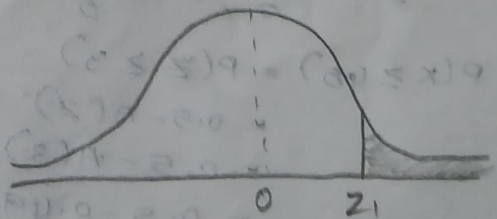
$P(x_1 \le X \le x_2) = P(z_1 \le Z \le z_2) = A(z_2) + A(z_1)$



To find $P(Z > z_1)$.

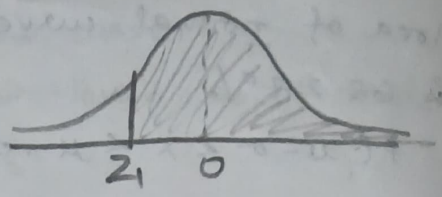
Case 3: If $z_1 > 0$ then

$P(Z > z_1) = 0.5 - A(z_1)$



Case 4: If $z_1 < 0$ then

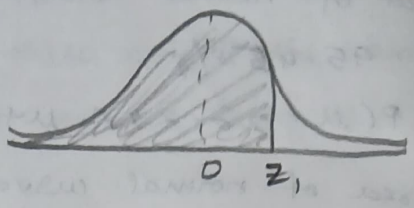
$P(z > z_1) = 0.5 + A(z_1)$



To find $P(z < z_1)$

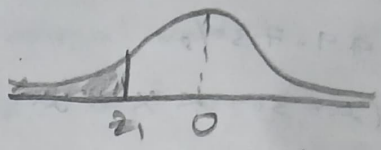
Case 5: If $z_1 > 0$ then

$P(z < z_1) = 0.5 + A(z_1)$



Case 6: If $z_1 \leq 0$ then

$P(z < z_1) = 0.5 - A(-z_1)$.



Problems:

① If X is a normal variate with mean 30 and S.D 5.

Find the probabilities that

(i) $26 \leq X \leq 40$ (ii) $X \geq 45$.

Sol: Mean $\mu = 30$, $\sigma = 5$.

(i) $26 \leq X \leq 40$.

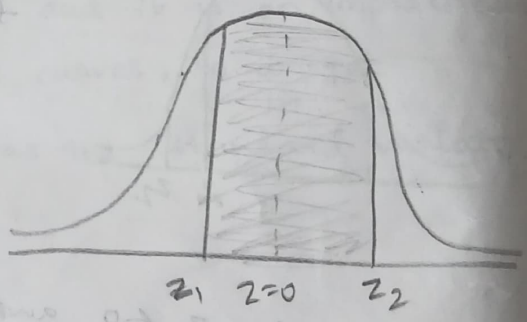
$$z = \frac{x - \mu}{\sigma} = \frac{x - 30}{5}.$$

when $x_1 = 26$

$$z_1 = \frac{26 - 30}{5} = -0.8$$

when $x_2 = 40$

$$z_2 = \frac{40 - 30}{5} = 2.$$



$P(x_1 \leq x \leq x_2) = P(z_1 \leq z \leq z_2)$

$P(26 \leq X \leq 40) = P(-0.8 \leq z \leq 2)$

$= A(z_2) + A(z_1)$

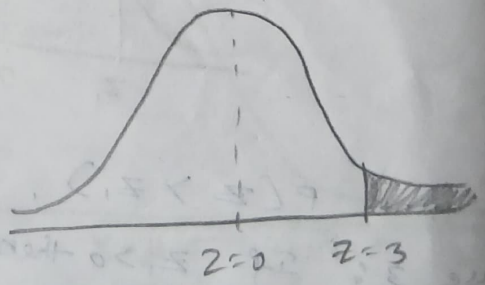
$= A(2) + A(-0.8)$

$= 0.4772 + 0.2881$

$= 0.7653$

(ii) $X \geq 45$

$$z = \frac{x - \mu}{\sigma} = \frac{45 - 30}{5} = 3.$$



$P(x \geq 45) = P(z \geq 3)$

$= 0.5 - A(z)$

$= 0.5 - A(3)$

$= 0.5 - 0.4987 = 0.0013$

(2) For a normally distributed variate with mean 1 and SD 3, find the probabilities that

(i) $3.43 \le x \le 6.19$ (ii) $-1.43 \le x \le 6.19$

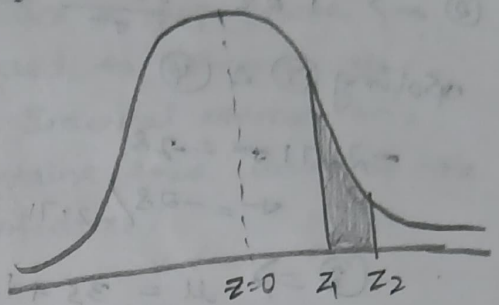
Sol: Given, $\mu = 1$, $\sigma = 3$.

(i) $3.43 \le x \le 6.19$

$$z = \frac{x-\mu}{\sigma} = \frac{x-1}{3}.$$

$$z_1 = \frac{3.43-1}{3} = +0.81$$

$$z_2 = \frac{6.19-1}{3} = 1.73.$$



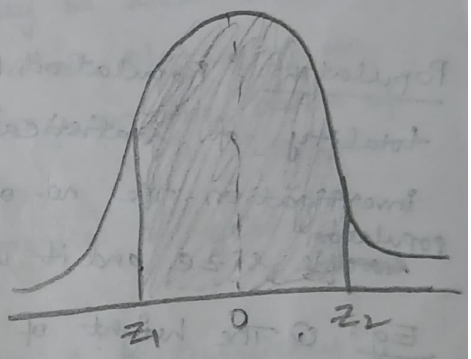
$$P(3.43 \le x \le 6.19) = P(0.81 \le z \le 1.73)$$
$$= A(1.73) - A(0.81)$$
$$= 0.4582 - 0.2910$$
$$= 0.1672$$

(ii) $-1.43 \le x \le 6.19$

$$z = \frac{x-\mu}{\sigma} = \frac{x-1}{3}.$$

$$z_1 = \frac{-1.43-1}{3} = -0.81$$

$$z_2 = \frac{6.19-1}{3} = 1.73.$$



$$P(-1.43 \le x \le 6.19) = P(-0.81 \le z \le 1.73)$$
$$= A(z_1) + A(z_2)$$
$$= A(-0.81) + A(1.73)$$
$$= 0.2910 + 0.4582$$
$$= 0.7492.$$

(3) For a ND, 7% of the items are under 35 and 89% are under 63. Determine the mean and variance of the distribution.

Sol: Let μ = mean
σ = S.D.

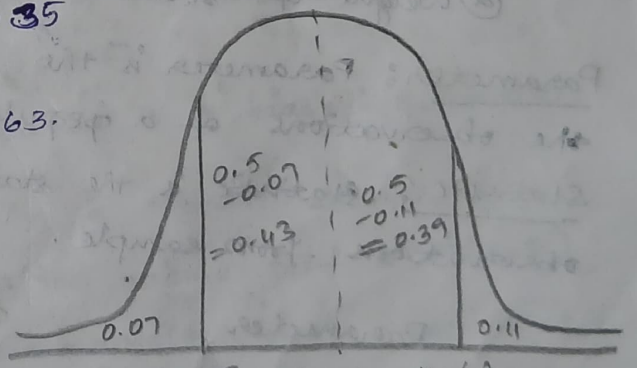
Given 7% of the items are under 35

$P(x < 35) = 7\% = 0.07$

Given, 89% of the items are under 63.

$P(x < 63) = 89\% = 0.89$.

$P(x > 63) = 1 - 0.89 = 0.11$



When $x = 35$

$$z_1 = \frac{x-\mu}{\sigma} = \frac{35-\mu}{\sigma} \quad \text{---(1)}$$

When $x = 63$

$$z_2 = \frac{63-\mu}{\sigma} \quad \text{---(2)}.$$

From figure

$P(0<z<z_2) = 0.39 \Rightarrow z_2 = 1.23$ (From table)

$P(0<z<z_1) = 0.43 \Rightarrow z_1 = -1.48$ (From table)

~~From~~

① $\Rightarrow -1.48 = \frac{35-\mu}{\sigma} \Rightarrow \mu - 1.48\sigma = 35$ — ③

② $\Rightarrow 1.23 = \frac{63-\mu}{\sigma} \Rightarrow \mu + 1.23\sigma = 63$ — ④

Solving ③ & ④

$-2.71\sigma = -28$

$\sigma = -28/-2.71 = 10.332.$

③ $\Rightarrow \mu = 35 + 1.48(10.332)$

$\therefore$ Mean, $\boxed{\mu = 50.29}$

$\therefore \boxed{\text{Variance}, \sigma^2 = (10.332)^2 = 106.75}$

Population: Population (or universe) is the aggregate (or) totality of statistical data forming a subject of investigation. The no. of observations in population is called population ~~sample~~ size and it is denoted by N.

Eg: ① The height of Indians

② The population of Nationalized banks in India.

Sample: A sample is a subset of population and no. of observations in the sample is called the sample size and it is denoted by n.

Eg: ① Height of ~~students~~ students in class of a college.

② Weight of students in class of college.

Parameter: Parameter is the statistical measure of all the observations of a population.

Statistic: Statistic is the statistical measure based on observations from sample.

Parameter		Statistic
μ	Mean	$\bar{x}$
σ	S.D	s
σ^2	Variance	s^2
N	Sample size	n
ρ	rank correlation	r

Estimator: The technique of finding an estimator to produce an estimate of the unknown parameter on the basis of a sample is called estimator.

There are 2 types of estimations.

(1) Point estimation

(2) Interval estimation.

In Point estimation; here we determine a appropriate single statistic whose value is used to estimate the unknown parameter, where as In Interval estimation; we determine an interval that contains true value of the unknown parameter with certain confidence.

Characteristic of good estimator:

Many functions of sample observations may be proposed as estimators of the same parameter. For ex, either the mean or median & mode of the sample values may be used to estimate the parameter. If of the Normal Distribution. Naturally we have to choose one among these estimators on the basis of certain criteria.

Prof. Ronald A. Fisher gave the 4 properties of good estimator.

(1) Unbiasedness (2) Consistency (3) Efficiency (4) Sufficiency.

1) Unbiasedness: If $x_1, x_2, \dots, x_n$ is a random sample of size n take from population whose probability density function is $f(x, \theta)$, where θ is the population parameter, then, an estimator $T = t(x_1, x_2, \dots, x_n)$ is said to be unbiased estimator of the parameter θ if and only if $E(T) = \theta$ for all $\theta \in \Theta$.

Thus, if $E(T) \neq \theta$ then we called it a biased estimator.

2) Consistency: If $x_1, x_2, \dots, x_n$ is random sample of size 'n' taken from a population whose pdf is $f(x, \theta)$, where θ is the population parameter then, consider a sequence of estimators say

$T_1 = t_1(x_1)$, $T_2 = t_2(x_1, x_2)$, $T_3 = t_3(x_1, x_2, x_3)$, ...

....., $T_n = t_n(x_1, x_2, \dots, x_n)$.

The value of estimator tends to get closer to parameter θ as sample size increases.

i.e., $T_n \to \theta$ as $n \to \infty$ for every $\theta \in \Theta$.

Sufficient conditions for consistency:

If $\{T_n\}$ is a sequence of estimators for all $\theta \in \Theta$, and

(i) $E(T_n) \to \theta$ as $n \to \infty$

i.e., estimator T_n is unbiased.

(ii) $\text{Var}(T_n) \to 0$ as $n \to \infty$

i.e., variance of estimator T_n converges to zero as $n \to \infty$.

The condition (ii) is called Invariance property of consistent estimator.

Then estimator T_n is consistent estimator of θ.

③ Efficiency: If T_1 and T_2 are two estimators of an parameter θ. Then T_1 is said to be more efficient than T_2 if

$\text{Var}(T_1) < \text{Var}(T_2)$ for all 'n'.

Eg: T_1: sample mean

T_2: sample median.

To estimate population parameter mean (μ), sample mean is more efficient than sample median if variance (mean) < variance (median).

(or)

sample median is more efficient than sample mean if variance (median) < variance (mean).

Most efficient estimator: In class of estimators of parameter, if there exists one estimator whose variance is minimum among the class then it is called most efficient estimator of that parameter.

If T^* is the most efficient estimator having variance $\text{Var}(T^*)$ and T is any other estimator having variance $\text{Var}(T)$. Then Absolute efficiency e defined as

$$e = \frac{\text{Var}(T^*)}{\text{Var}(T)} < 1$$

(4) **Sufficiency:** An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter.

More precisely, if $T = t(x_1, x_2, \ldots, x_n)$ is an estimator of a parameter θ such that the conditional distribution of $x_1, x_2, \ldots, x_n$ given T, is independent of θ, then T is sufficient estimator of θ.

Mathematically,

$$f(x_1, x_2, \ldots, x_n / T = t) = g(x_1, x_2, \ldots, x_n)$$

where g does not depend on parameter θ.

Sufficient condition is used if it is difficult to estimate statistic T.

UNIT-2 : POINT ESTIMATION

Point estimation: A statistical method where a single value from sample data is used to estimate a single value of population parameter is called Point Estimation.

Estimator (Statistic) → Estimation Process → Estimate (numerical value / parameter)

Estimator: A function of sample (statistic) which is used to calculate the estimate is known as estimator.

Estimate: The numerical value produced by the estimator is known as estimate.

Methods of point estimation:

Some of the important & frequently used methods are:

1. Method of maximum likelihood
2. Method of moments
3. Method of least squares
4. Method of Minimum chi-square
5. Method of Minimum Variance.

(1) Method of Maximum Likelihood:

Concept of ML

(1) One of the most important methods of point estimation is the method of maximum likelihood.

(2) It was initially introduced by Prof. C. F. Gauss but later on it was used as general method of estimation by Prof. Ronald A. Fisher in 1912.

(3) The principle of maximum likelyhood estimation is to find or estimate or choose the value of unknown parameter which would most likely generate the observed data.

(4) This method is explained in the following example:

Eg: Coin is tossed 5 times and observed that the 3 heads and 2 tails. Instead of assuming probability of getting head is $p = 0.5$, we find estimate the value of p that makes the observed data most likely.

Since the no. of heads follows the BD.

$r=3,\ n=5$

$P(X=r) = P(3) = {}^{n}C_r\, p^r q^{n-r}$

$P(X=3) = P(3) = {}^{5}C_3\, p^3 (1-p)^{5-3} = {}^{5}C_3\, p^3 (1-p)^2$.

For $p=0.1$

$P(3) = {}^{5}C_3 (0.1)^3 (0.9)^2 = 0.0081$

Similarly we find different values of p in the following table.

S.No.	p	Probability / likelihood func^c
1	0.1	0.0081
2	0.2	0.0512
3	0.3	0.1323
4	0.4	0.2304
5	0.5	0.3125
6	0.6	0.3456 → maximum.
7	0.7	0.3087
8	0.8	0.2048
9	0.9	0.0729

From the above table, we conclude that p is more likely to be 0.6, bcoz at $p=0.6$ the probability is maximum (or) likelihood func^c is maximum.

Likelihood function:

If $x_1, x_2, \ldots, x_n$ is a random sample taken from a population with joint probability density function $f(x_1, x_2, \ldots, x_n, \theta)$ of sample values. then the likelihood function is denoted $L(\theta)$ and is defined as

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

For discrete case $L(\theta) = P[x=x_1]\, P[x=x_2] \ldots P[x=x_n]$.

For continuous case $L(\theta) = f(x_1, \theta) \cdot f(x_2, \theta) \ldots f(x_n, \theta)$.

Procedure of Maximum likelihood method:

By the definition of likelihood function,

$L(\theta) = f(x_1, x_2, \ldots, x_n, \theta)$, let $\hat{\theta}$ be the parameter

$L(\hat{\theta}) = \sup L(\theta)$.

For maximum, we take

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \quad \text{i.e.,} \quad \frac{\partial L}{\partial \theta} = 0$$

and $\frac{\partial^2 L}{\partial \theta^2} < 0$ at stationary points $\theta_1, \theta_2, \ldots, \theta_n$.

But likelihood L is the product of different functions, so differentiating is very difficult. Hence we apply log for likelyhood fun^n. Since L is always positive, log L remains finite. When L is maximum, log L also attains the maximum.

Now we find $\frac{\partial (\log L)}{\partial \theta}$

and consider $\frac{\partial}{\partial \theta}(\log L) = 0$.

Solving the above eqn we get the stationary points $\theta_1, \theta_2, \ldots$.

Then we find the second derivative $\frac{\partial^2}{\partial \theta^2}(\log L)$

At $\theta = \theta_1$,

if $\frac{\partial^2}{\partial \theta^2}(\log L) < 0$ then θ_1 is the maximum possible likelihood value. Hence $\hat{\theta} = \theta_1$.

Similarly we check for $\theta_2, \theta_3, \ldots$

Properties of MLE:

(1) A ML Estimator is not necessarily unique.

(2) A ML Estimator is not necessarily unbiased.

(3) ~~A ML Estimator is a function of~~ If a sufficient statistic exists, it is a function of ML estimator

(4) If ML estimator exist, then it is most sufficient.

(5) A ML Estimator is consistent most of the times (except rare case)

(6) If $T = t(x_1, x_2, \ldots, x_n)$ is a ML Estimator of θ and $g(\theta)$ is a one-to-one function of θ, then $g(T)$ is a MLE of $g(\theta)$. This is known as invariance property of ML Estimator.

Problems:

① The number of weekly accidents occurring on a mile stretch of a particular road follows Poisson Distribution with parameter λ then find the maximum likelihood estimate of parameter λ on the basis of the following data.

No. of accidents	0	1	2	3	4	5	6
Frequency	10	12	12	9	5	3	1

Sol: Given data follows PD with parameter λ

∴ By the defn of PD

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

By the defn of likelihood fun^n,

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

In discrete case.

$$L(\lambda) = L = P[X=x_1]\,P[X=x_2] \cdots P[X=x_n]$$

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

$$= \frac{e^{-n\lambda}\lambda^{x_1+x_2+x_3+\cdots+x_n}}{x_1!\,x_2!\cdots x_n!}$$

$$L = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n}(x_i!)}$$

Taking log on both sides,

$$\log L = \log\left[\frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n}(x_i!)}\right]$$

$$\log L = \log\left[e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}\right] - \log \prod_{i=1}^{n}(x_i!)$$

$$= \log e^{-n\lambda} + \log \lambda^{\sum_{i=1}^{n} x_i} - \log \prod_{i=1}^{n}(x_i!)$$

$$\therefore \log L = -n\lambda + \sum_{i=1}^{n} x_i \log\lambda - \log \prod_{i=1}^{n}(x_i!)$$

Partially differentiating w.r.to λ is

$$\frac{\partial}{\partial \lambda} \log L = -n(1) + \sum_{i=1}^{n} x_i \cdot \frac{1}{\lambda} = 0$$

$$\Rightarrow \frac{\partial}{\partial \lambda} \log L = -n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i \quad \text{---(1)}$$

Consider $\frac{\partial}{\partial \lambda} \log L = 0$

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

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

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

$$\bar{x} = \lambda \quad \left[\text{where } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ is the sample mean}\right]$$

$$\therefore \boxed{\lambda = \bar{x}}$$

Again partially differentiating (1) w.r.t λ, we get

$$\frac{\partial^2}{\partial \lambda^2} \log L = -0 + \left(-\frac{1}{\lambda^2}\right) \cdot \sum_{i=1}^{n} x_i$$

$$= -\frac{1}{\lambda^2} \sum_{i=1}^{n} x_i$$

At $\lambda = \bar{x}$

$$= -\frac{1}{\bar{x}^2} \sum_{i=1}^{n} x_i$$

$$< 0$$

$\therefore \bar{x}$ is the MLE of parameter λ.

To find sample mean,

$$\bar{x} = \frac{1}{\sum f_i} \sum f_i x_i = \frac{(0\times10)+(1\times12)+(2\times12)+(3\times9)+(4\times5)+(5\times3)+(6\times1)}{10+12+12+9+5+3+1}$$

$$\bar{x} = \frac{104}{52} = 2.$$

Hence MLE for parameter λ is $\bar{x} = 2$.

(2) For random sampling from normal population $N(\mu, \sigma^2)$, find the maximum likelihood estimators for μ and σ^2.

Sol: Let $x_1, x_2, \ldots, x_n$ be a random sample of size 'n' taken from the population $N(\mu, \sigma^2)$ whose pdf is

$$f(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty, \; -\infty < \mu < \infty, \; \sigma > 0.$$

By the defn of likelihood func for a continuous case.

$$L(\mu, \sigma^2) = L = f(x_1, \mu, \sigma^2) \cdot f(x_2, \mu, \sigma^2) \cdots f(x_n, \mu, \sigma^2)$$

$$\Rightarrow L = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2} \cdots \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2}$$

$$= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \cdot e^{-\frac{1}{2}\left[\frac{(x_1-\mu)^2}{\sigma^2} + \frac{(x_2-\mu)^2}{\sigma^2} + \cdots + \frac{(x_n-\mu)^2}{\sigma^2}\right]}$$

$$= \left[\frac{1}{2\pi\sigma^2}\right]^{n/2} \cdot e^{\frac{-1}{2\sigma^2}\left[(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2\right]}$$

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

Taking log on both sides

$$\log L = \log\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} + \log e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

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

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

$$\therefore \log L = \frac{n}{2}\left[-\log 2\pi - \log\sigma^2\right] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \quad \text{---} (1)$$

For parameter μ:

Partially differentiating ① w.r.t to μ,

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

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

$$\frac{\partial}{\partial \mu}\log L = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \quad \text{—— ④}$$

Consider $\frac{\partial}{\partial \mu}\log L = 0$

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

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

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

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

$$\Rightarrow \boxed{\mu = \bar{x}} \quad [\text{where } \bar{x} \text{ is the sample mean}].$$

Partially differentiating ④ w.r.t μ,

$$\frac{\partial^2}{\partial \mu^2}\log L = \frac{1}{\sigma^2}\sum_{i=1}^{n}(-1) = -\frac{n}{\sigma^2} < 0. \ (\text{At } \mu = \bar{x})$$

$\therefore \bar{x}$ is the MLE of μ.

For parameter σ^2:

Partially differentiating ① w.r.t to σ^2.

$$\frac{\partial}{\partial(\sigma^2)}\log L = \frac{n}{2}\left[-0 - \frac{1}{\sigma^2}\right] - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\left(-\frac{1}{\sigma^4}\right)$$

$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \quad \text{—— ③}$$

Consider $\frac{\partial}{\partial \sigma^2}(\log L) = 0$

$$\Rightarrow -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\Rightarrow \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{n}{2\sigma^2}$$

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

$\therefore \boxed{\sigma^2 = s^2}$ where s^2 is the sample variance

Partially differentiating (3) w.r.t σ^2,

$$\frac{\partial^2}{(\partial(\sigma^2)^2)}(\log L) = -\frac{n}{2}\left(-\frac{1}{\sigma^4}\right) + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\left(\frac{-2}{\sigma^6}\right)$$

$$= \frac{n}{2\sigma^4} - \sum_{i=1}^{n}(x_i-\mu)^2 \cdot \frac{1}{\sigma^6}.$$

At $\sigma^2 = s^2$

$$= \frac{n}{2s^4} - \frac{1}{s^6}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= \frac{n}{2s^4} - \frac{1}{s^6}\, n s^2$$

$$= \frac{n}{2s^4} - \frac{n}{s^4} = -\frac{n}{2s^4} < 0.$$

$\therefore$ s^2 is the MLE of σ^2.

(3) Prove that for a Binomial distribution with density function $P[X=x] = {}^nC_x\, p^x q^{n-x}$, $x = 1, 2, \ldots, n$, $q = 1-p$, the maximum likelihood estimate of p is $\frac{\bar{x}}{n}$.

Sol: Given, a binomial distribution with

$$P[X=x] = {}^nC_x\, p^x q^{n-x},\ x = 1, 2, \ldots, n,\ q = 1-p$$

$$\Rightarrow P[X=x] = {}^nC_x\, p^x (1-p)^{n-x}.$$

Let $x_1, x_2, \ldots, x_n$ be a random sample of above B.D.

By defn of likelihood func, for discrete case,

$$L(p) = L = P[x=x_1] \cdot P[x=x_2] \cdots P[x=x_n]$$

$$L = {}^nC_{x_1}\, p^{x_1} (1-p)^{n-x_1}\ {}^nC_{x_2}\, p^{x_2} (1-p)^{n-x_2} \cdots {}^nC_{x_n}\, p^{x_n} (1-p)^{n-x_n}.$$

$$L = \prod_{i=1}^{n} {}^nC_{x_i}\ p^{x_1+x_2+\cdots+x_n} (1-p)^{n-x_1+n-x_2+\cdots+n-x_n}$$

$$L = \left[\prod_{i=1}^{n} {}^nC_{x_i}\right] p^{\sum_{i=1}^{n} x_i} (1-p)^{n^2 - \sum_{i=1}^{n} x_i}.$$

Taking log on b.s.

$$\log L = \log \prod_{i=1}^{n} {}^{n}C_{x_i} + \sum_{i=1}^{n} x_i \log p + \left(n^2 - \sum_{i=1}^{n} x_i\right) \log(1-p)$$

Partially differentiating w.r.t p;

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

$$= \sum_{i=1}^{n} x_i \cdot \frac{1}{p} - \left(n^2 - \sum_{i=1}^{n} x_i\right)\left(\frac{1}{1-p}\right) \quad \text{---- (1)}$$

Consider $\frac{\partial \log L}{\partial p} = 0$

$$\Rightarrow \frac{1}{p}\sum_{i=1}^{n} x_i - \left(n^2 - \sum_{i=1}^{n} x_i\right)\left(\frac{1}{1-p}\right) = 0$$

$$\Rightarrow \frac{1}{p}\sum_{i=1}^{n} x_i = \frac{1}{1-p}\left(n^2 - \sum_{i=1}^{n} x_i\right)$$

$$\Rightarrow \frac{1}{p} n\bar{x} = \frac{1}{1-p}(n^2 - n\bar{x})$$ where $\bar{x}$ is the sample mean.

$$\Rightarrow n\bar{x} - pn\bar{x} = pn^2 - pn\bar{x}$$

$$\Rightarrow n\bar{x} = pn^2$$

$$\Rightarrow \boxed{p = \frac{\bar{x}}{n}}$$

Partially differentiating (1) w.r.t p;

$$\frac{\partial^2}{\partial p^2} \log L = \sum_{i=1}^{n} x_i \left(\frac{-1}{p^2}\right) - \left(n^2 - \sum_{i=1}^{n} x_i\right)\left(+\frac{1}{(1-p)^2}\right)$$

$$= -\frac{n\bar{x}}{p^2} - \frac{1}{(1-p)^2}(n^2 - n\bar{x})$$

At $p = \frac{\bar{x}}{n}$.

$$= -\frac{n\bar{x}\, n^2}{\bar{x}^2} - \frac{1}{\left(1-\frac{\bar{x}}{n}\right)^2}(n^2 - n\bar{x})$$

$$= -\frac{n^3}{\bar{x}} - \frac{n^2 \cdot n(n-\bar{x})}{(n-\bar{x})^2} = n^3\left[\frac{-1}{\bar{x}} - \frac{1}{n-\bar{x}}\right]$$

$$= -n^3\left[\frac{n}{\bar{x}(n-\bar{x})}\right] = -n^3\left[\frac{2\bar{x}-n}{\bar{x}(n-\bar{x})}\right]$$

$$= -n^3 \left[\frac{n - \bar{x} + \bar{x}}{\bar{x}(n - \bar{x})} \right] = -\frac{n^4}{\bar{x}(n-\bar{x})} < 0.$$

$\therefore \frac{\bar{x}}{n}$ is the MLE of p.

(4) Obtain the ML estimator for α and β for uniform or rectangular population whose pdf is given by

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha}, & \alpha \le x \le \beta \\ 0, & \text{elsewhere} \end{cases}$$

Sol: Given, uniform or rectangular population, whose pdf is given by $f(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \alpha \le x \le \beta \\ 0, & \text{elsewhere} \end{cases}$

Let $x_1, x_2, \ldots, x_n$ be a random sample from the uniform distribution.

Then likelihood fun^c,

$$L = f(x_1, \alpha, \beta) \cdot f(x_2, \alpha, \beta) \cdots, f(x_n, \alpha, \beta)$$

$$= \left(\frac{1}{\beta-\alpha}\right) \cdot \left(\frac{1}{\beta-\alpha}\right) \cdots \left(\frac{1}{\beta-\alpha}\right)$$

$$L = \left(\frac{1}{\beta-\alpha}\right)^n \quad \text{---} (1)$$

$$\log L = -n \log(\beta - \alpha)$$

Partially differentiating w.r.t α;

$$\frac{\partial}{\partial \alpha} \log L = \frac{+n}{\beta - \alpha}.$$

Partially differentiating w.r.t β,

$$\frac{\partial}{\partial \beta} \log L = \frac{-n}{\beta - \alpha}.$$

Equating to zero we get

$$\frac{n}{\beta - \alpha} = 0 \quad \& \quad \frac{-n}{\beta - \alpha} = 0.$$

From which we get $\alpha = \beta = \infty$. So the method of differentiating to find the maximum cannot be used.

So to maximise L, from eqn (1), we need maximum value of α and minimum value of β.

Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ be the ascending ordered arrangement, then ATP:

$$\alpha \leq x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \leq \beta.$$

So maximum value of α is $x_{(1)}$

minimum value of β is $x_{(n)}$

$\therefore$ $x_{(1)}$ is the MLE of α

$x_{(n)}$ is the MLE of β.

(5) Obtain the ML estimator of θ for the following distribution $f(x, \theta) = \frac{1}{\theta}$, $0 \leq x \leq \theta$, $\theta > 0$, if the sample values are 1.5, 1.0, 0.7, 2.2, 1.3 and 1.2.

Sol: Given, distribution is

$$f(x, \theta) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta, \quad \theta > 0.$$

Let $x_1, x_2, \cdots, x_n$ be a random sample from the above distribution. And the sample values are 1.5, 1.0, 0.7, 2.2, 1.3 and 1.2.

Then likelihood function is

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

$$L = \frac{1}{\theta} \cdot \frac{1}{\theta} \cdot \cdots \frac{1}{\theta}$$

$$L = \frac{1}{\theta^n}$$

$$\log L = -n \log \theta.$$

$$\frac{\partial}{\partial \theta} \log L = \frac{-n}{\theta}$$

$$\frac{\partial}{\partial \theta} \log L = 0$$

$$\frac{-n}{\theta} = 0$$

From which we get $\theta = \infty$. So the method of differentiation fails.

So to maximize L, from eqn (1), we need minimum value of θ.

Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ be the ascending ordered arrangement of the sample, then ATP.

$$0 \leq x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \leq \theta.$$

So ~~maximum~~ minimum value of θ is $x_{(n)}$.

$\therefore x_{(n)}$ is the MLE of θ.

Given sample is 1.5, 1.0, 0.7, 2.2, 1.3 and 1.2

Arranging in ascending order we get

$0.7 < 1.0 < 1.2 < 1.3 < 1.5 < 2.2$

$\therefore$ 2.2 is the MLE of θ.

Large sample properties of MLE:

1. Consistency
2. Asymptotic normality
3. Asymptotic efficiency
4. Invariance.

(2) Method of Moments:

1. The method of moments is the oldest but simple method for determining the point estimate of the unknown parameters.
2. It was discovered by Karl Pearson in 1894. This method was continued upto MLE was found.
3. The principle of this method consists of equating the sample moments to the corresponding moments of population parameters.
4. The process is explained below:

Let $x_1, x_2, \ldots, x_n$ be a random sample from a population whose pdf is $f(x, \theta)$ with k unknown parameters, say, $\theta_1, \theta_2, \ldots, \theta_k$.

Then,

r^{th} sample moment about origin is

$$M'_r = \frac{1}{n} \sum_{i=1}^{n} x_i^r.$$

and about mean is

$$M_r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^r.$$

r^{th} population moment about origin is

$$\mu_r' = E(X)^r$$

about mean is

$$\mu_r = E(X-\mu)^r.$$

Generally, the 1st moment about origin and rest central moments are equated to the corresponding sample moments.

i.e., For $r=1$, $\mu_1' = M_1'$.

For $r = 2, 3, \ldots, k$, $\mu_r = M_r$

By solving these k equations for unknown parameters, we get moment estimators.

Problems:

(1) Find the estimator for λ by the method of moments for the exponential distribution whose probability density function is given by $f(x,\lambda) = \frac{1}{\lambda} e^{-x/\lambda}$, $x>0, \lambda>0$.

Sol: Let $x_1, x_2, \ldots, x_n$ be a random sample taken from exponential distribution with pdf

$$f(x,\lambda) = \frac{1}{\lambda} e^{-x/\lambda}, \quad x>0, \lambda>0.$$

WKT, 1st population moment about origin is

$$\mu_1' = \lambda$$

1st sample moment about origin is

$$M_1' = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \quad [\text{where } \bar{x} \text{ is the sample mean}]$$

By the method of moments

$$\mu_1' = M_1'$$

$$\Rightarrow \boxed{\lambda = \bar{x}}$$

Hence moment estimator for λ is $\bar{x}$.

(2) If $x_1, x_2, \ldots, x_m$ is a random sample taken from binomial distribution (n, p) where n and p are unknown, obtain moment estimators for both n and p.

Sol: Given, $x_1, x_2, \ldots, x_m$ is a random sample taken from B.D. (n, p).

WKT, 1st Population moment about origin is

$$\mu_1' = E(X) = np$$

1st sample moment moment about origin is

$$M_1' = \frac{1}{m}\sum_{i=1}^{m} x_i = \bar{x} \quad \text{(where } \bar{x} \text{ is sample mean)}$$

2nd population moment about mean is

$$\mu_2 = E(X-\mu)^2 = npq$$

2nd sample moment about mean is

$$M_2 = \frac{1}{m}\sum_{i=1}^{m}(x_i-\bar{x})^2 = s^2 \quad \text{(where } s^2 \text{ is sample variance)}$$

By method of moments,

$$\mu_1' = M_1'$$

$$\Rightarrow np = \bar{x} \quad \text{---} (1)$$

$$\mu_2 = M_2$$

$$\Rightarrow npq = s^2 \quad \text{---} (2)$$

$$\frac{(1)}{(2)} \Rightarrow \frac{np}{npq} = \frac{\bar{x}}{s^2}$$

$$\Rightarrow q = \frac{s^2}{\bar{x}}.$$

we have $p = 1-q = 1-\frac{s^2}{\bar{x}} = \frac{\bar{x}-s^2}{\bar{x}}$

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

$$\Rightarrow n = \frac{\bar{x}^2}{\bar{x}-s^2}.$$

∴ The moment estimators for

p is $\frac{\bar{x}-s^2}{\bar{x}}$, n is $\frac{\bar{x}^2}{\bar{x}-s^2}$.

where $\bar{x}$ is sample mean and s^2 is sample variance.

(3) Show that moment estimator and maximum likelihood estimator are same of the parameter θ of the geometric distribution $G(\theta)$ whose pmf is $P[X=x] = \theta(1-\theta)^x$, $\theta>0, x=0,1,2,\ldots$

Sol: ~~Method of Moment Estimators~~

Let $x_1, x_2, \ldots, x_n$ be a sample from geometric distribution $G(\theta)$ with

$$P[X=x] = \theta(1-\theta)^x, \theta>0, x=0,1,2,\ldots$$

Moment estimator:

1st population moment about origin is

$$\mu_1' = \frac{1-\theta}{\theta}.$$

1st sample moment about origin is $M_1' = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$

where $\bar{x}$ is the sample mean.

By method of moments,

$$\mu_1' = M_1'$$

$$\Rightarrow \frac{1-\theta}{\theta} = \bar{x}$$

$$\Rightarrow 1-\theta = \bar{x}\theta$$

$$\Rightarrow 1 = (\bar{x}+1)\theta$$

$$\Rightarrow \theta = \frac{1}{\bar{x}+1}.$$

$\therefore$ Moment estimator of θ is $\frac{1}{\bar{x}+1}$.

Mean $= \sum p_i x_i$

$= \sum \theta(1-\theta)^{x_i} x_i$

$= \theta(1-\theta) + 2\theta(1-\theta)^2 + \dots + n\theta(1-\theta)^n$

$= \theta[1-\theta][1+2(1-\theta)+3(1-\theta)^2 + \dots]$

$= \frac{\theta[1-\theta]}{(1-(1-\theta))^2} = \frac{\theta(1-\theta)}{\theta^2}$

Maximum likelihood Estimator:

By def^n of likelihood fun^c,

$$L(\theta) = P[X=x_1] \cdot P[X=x_2] \cdots P[X=x_n]$$

$$= \theta(1-\theta)^{x_1} \cdot \theta(1-\theta)^{x_2} \cdots \theta(1-\theta)^{x_n}$$

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

$$\log L = n\log\theta + \left(\sum_{i=1}^{n} x_i\right)\log(1-\theta).$$

Partially differentiating w.r.t θ,

$$\frac{\partial}{\partial\theta}\log L = \frac{n}{\theta} - \frac{\sum_{i=1}^{n} x_i}{1-\theta}.$$

Considere $\frac{\partial}{\partial\theta}\log L = 0$

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

$$\Rightarrow \frac{n}{\theta} = \frac{\sum_{i=1}^{n} x_i}{1-\theta}.$$

$$\Rightarrow 1-\theta = \frac{\sum_{i=1}^{n} x_i}{n}\cdot\theta$$

$$\Rightarrow 1-\theta = \bar{x}\theta$$

$$\Rightarrow 1 = (\bar{x}+1)\theta$$

$$\Rightarrow \theta = \frac{1}{\bar{x}+1}$$ where $\bar{x}$ is the sample mean

Again partially differentiating w.r.t θ,

$$\frac{\partial^2}{\partial \theta^2}\log L = -\frac{n}{\theta^2} - \frac{\sum_{i=1}^{n} x_i}{(1-\theta)^2}.$$

At $\theta = \frac{1}{1+\bar{x}}$

$$= -n(\bar{x}+1)^2 - n\bar{x}\left(\frac{1}{1-\frac{1}{1+\bar{x}}}\right)^2$$

$$= -n(\bar{x}+1)^2 - \frac{n\bar{x}(\bar{x}+1)^2}{\bar{x}^2}$$

$$= -n(\bar{x}+1)^2 - \frac{n(\bar{x}+1)^2}{\bar{x}} < 0.$$

$\therefore \frac{1}{\bar{x}+1}$ is the MLE of θ.

Hence $\frac{1}{\bar{x}+1}$ is both moment estimator and MLE of θ.

Properties of Moment Estimators:

1. The moment estimators can be obtained easily.
2. The moment estimators are not necessarily unbiased.
3. The moment estimators are consistent because by the law of large numbers a sample moment (raw & central) is a consistent estimator for the corresponding population moment.
4. The moment estimators are generally less efficient than maximum likelihood estimators.
5. The moment estimators are asymptotically normally distributed
6. The moment estimators may not be function of sufficient statistics.
7. The moment estimators are not unique.

Drawbacks of moment estimators:

1. This method is based on equating population moments with sample moments. But in some situations, like as cauchy distribution, the population moment does not exist therefore in such situations this method cannot be used.
2. This method does not, in general, give estimators with all the desirable properties of good estimator.
3. The property of efficiency is not possessed by these estimators.
4. The moment estimators are not unbiased in general.
5. Generally, the moment estimators and MLE are identical. But if they do differ, then MLEs are usually preferred.

④ Obtain the estimator of parameter λ when sample is taken from a Poisson distribution by the method of moments.

Sol: Let $x_1, x_2, \ldots, x_n$ be a random sample taken from PD with pdf, $P[X=x] = \frac{e^{-\lambda}\lambda^x}{x!}$.

1st population moment about origin is

$\mu_1' = \lambda$.

1st sample moment about origin is

$M_1' = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ [where $\bar{x}$ is the sample mean]

~~2nd population moment about mean is~~

~~$\mu_2 = E(X-\mu)^2 = \sigma^2$~~

~~2nd sample moment about mean is~~

~~$M_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2$ [where s^2 is the sample variance.~~

By method of moments

$\mu_1' = M_1'$

$\Rightarrow \lambda = \bar{x}$

$\therefore \bar{x}$ is the moment estimator of λ.

⑤ Obtain the moment estimators of the parameters μ and σ^2 when the sample is drawn from the normal population.

Sol: Let $x_1, x_2, \ldots, x_n$ be a random sample taken from ND.

1st population moment about origin is

$\mu_1' = \mu$.

1st sample moment about origin is

$M_1' = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ [where $\bar{x}$ is the sample mean].

2nd population moment about mean is

$\mu_2 = \sigma^2$ [$\because \mu_2 = E(X-\mu)^2$]

2nd sample moment about mean is

$M_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2$ [where s^2 is the sample variance]

By method of moments,

$\mu_1' = M_1' \Rightarrow \mu = \bar{x}$

$\mu_2 = M_2 \Rightarrow \sigma^2 = s^2$

$\therefore \bar{x}$ and s^2 are moment estimators of μ and σ^2.

(3) Method of Least Squares:

Method of least squares is obtained from the concept of ML.

Consider the ML of the parameter μ, when σ^2 is known on the basis of a random sample $y_1, y_2, \ldots, y_n$ of size n taken from a normal population $N(\mu, \sigma^2)$.

The pdf of Normal population is

$$f(y, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, \quad -\infty < y < \infty, \; -\infty < \mu < \infty, \; \sigma > 0$$

By the defn of likelihood func

$$L(\mu, \sigma^2) = L = f(y_1, \mu, \sigma^2) \cdot f_2(y_2, \mu, \sigma^2) \cdots f_n(y_n, \mu, \sigma^2)$$

$$L = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_1-\mu}{\sigma}\right)^2} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_2-\mu}{\sigma}\right)^2} \cdots \cdots \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_n-\mu}{\sigma}\right)^2}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\left[(y_1-\mu)^2 + (y_2-\mu)^2 + \cdots + (y_n-\mu)^2\right]}$$

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

$$L = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} (y_i-\mu)^2\right]}$$

$$\log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} (y_i-\mu)^2 \log e$$

$$\log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} (y_i-\mu)^2$$

By principle of MLE, we have to maximize $\log L$ w.r.t μ, and $\log L$ is maximum when $\sum_{i=1}^{n} (y_i-\mu)^2$ is minimum.

i.e., sum of squares $\sum_{i=1}^{n} (y_i-\mu)^2$ must be least.

The method of least squares is mostly used to estimate the parameters of linear function.

Suppose μ is linear func, of parameters $\theta_1, \theta_2, \ldots, \theta_k$.

$$\mu = x_1\theta_1 + x_2\theta_2 + \cdots + x_k\theta_k = \sum_{j=1}^{k} x_j\theta_j.$$

Then we have to minimize

$$\sum_{i=1}^{n} (y_i-\mu)^2 = \sum_{i=1}^{k}\left[y_i - \sum_{j=1}^{k} x_j\theta_j\right]^2$$

where $\theta_1, \theta_2, \ldots, \theta_k$ are parameters of observed sample observations $y_1, y_2, \ldots, y_n$.

Properties of least squares:

Least square estimators are not so popular. They possess some properties which are as follows:

1. Least square estimators are unbiased in case of linear models
2. Least squares estimators are minimum variance unbiased estimators (MVUE) in case of linear models.

(4) Method of Minimum Chi-Square:

The method of minimum chi-square makes the use of Pearson's chi-square statistic. This method can be used in case of discrete distributions or for grouped data from a continuous distribution.

WKT, Pearson's chi-square statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{[f_i - n p_i(\theta)]^2}{n p_i(\theta)} = \sum_{i=1}^{k} \frac{f_i^2}{n p_i(\theta)} - n.$$

where f_i are frequencies

p_i are probabilities which are function of unknown parameter θ_i.

So, to obtain the estimates of θ_i's, partially differentiate χ^2 statistic w.r.t θ_i $(i = 1, 2, \ldots, m)$ successively and equate to zero. Also check that the standard derivatives are non-negatives.

i.e., $$\frac{\partial \chi^2}{\partial \theta_i} = 0 \quad \text{for } i = 1, 2, \ldots, m$$

and $$\frac{\partial^2 \chi^2}{\partial \theta_i^2} \geq 0.$$

$\frac{\partial \chi^2}{\partial \theta_i} = 0$ provides m simultaneous eqn, solving them we get the value of $\theta_1, \theta_2, \ldots, \theta_m$ respectively.

gets the estimated values of $\theta_1, \theta_2, \ldots$ respectively.

Properties of Minimum Chi-square:

1. The minimum chi-square estimators are consistent.
2. The minimum χ^2-~~square~~ estimators are asymptotically normal.
3. Minimum χ^2 estimators are efficient
4. Minimum χ^2 estimators are not necessarily unbiased.

Uses:

Minimum χ^2 method of estimation is rarely used in practice. It is used only when it is difficult to solve the simultaneous equations obtained under maximum likelihood estimation method.

Method of Modified Minimum Chi-square :

The minimum chi-square method provides some computational difficulties for estimating the parameters since p_i's are occuring in the denominator. In such cases, one can use the method of modified minimum chi-square.

For a χ^2-statistic, the likelihood fun^n is

$$\log L = c - \frac{1}{2}(\chi')^2 + O(n^{-1/2}).$$

where, c is independent of p_i's.

$$(\chi')^2 = \sum_{i=1}^{k} \frac{(np_i - o_i)^2}{o_i}$$ is the modified chi-square statistic.

p_i are probabilities
o_i are observed frequencies
n is the sample size.

∴ If we neglect terms of order $O(n^{-1/2})$, then maximization of $\log L$ amounts to the minimization of $(\chi')^2$.

UNIT-III : INTERVAL ESTIMATION

Introduction: In point estimation, we learn how one can obtain point estimates of the unknown parameters of the population using sample observations. But, the major drawback of point estimation is that it does not specify how confident we can be that the estimated value is close to the true value of the parameter. This limitation is over come by the technique of Interval Estimation.

Ex: Suppose, the average annual income of 50 persons in a colony is Rs. 84240.

Statement 1: The average annual income of persons in the colony is between Rs. 80000 to Rs. 90000.

Statement 2: The annual income is 84240/-

We observe that statement ① is definitely more likely to be correct than statement ②.

Interval Estimation:

If we can find two values T_1, T_2 with the help of sample observations and have a interval $[T_1, T_2]$ such that it contains the true value of parameter θ, then $[T_1, T_2]$ is known as interval estimate of the parameter θ. This technique of estimation is known as Interval estimation.

Confidence Limits and confidence coefficient:

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a population with pdf $f(x, \theta)$.

Let $T_1 = t_1(x_1, x_2, \ldots, x_n)$ and $T_2 = t_2(x_1, x_2, \ldots, x_n)$ where $T_1 \leq T_2$.

Let $[T_1, T_2]$ is a random interval which includes the true value of parameter θ with probability $1-\alpha$.

i.e., $P[T_1 \leq \theta \leq T_2] = 1-\alpha$.

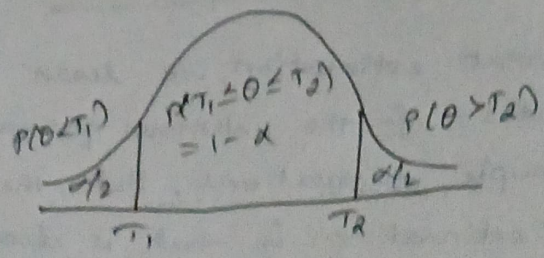
Here confidence coefficient $= 1-\alpha$.

T_1 = Lower Confidence Limit (LCL)

T_2 = Upper Confidence Limit (UCL)

Length of confidence interval $l = T_2 - T_1$.

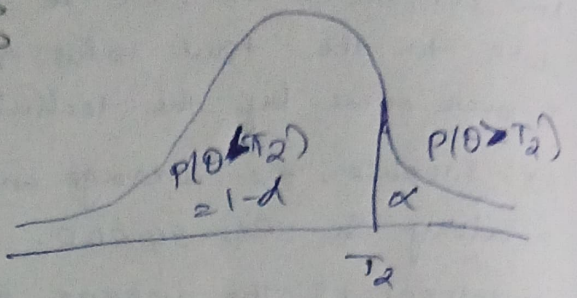
$[T_1, T_2]$ is known as $(1-\alpha)100\%$ confidence interval. Thus when $\alpha = 0.05$ we have a 95% confidence interval and when $\alpha = 0.01$ we have a 99% confidence interval



One sided confidence intervals

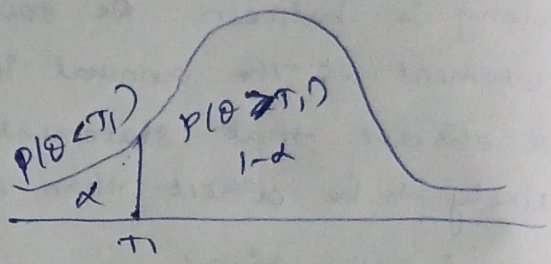
Left sided (or) left tailed confidence intervals.

$P(\theta < T_2) = 1-\alpha$.



Right sided (or) Right tailed confidence intervals

$P(\theta > T_1) = 1-\alpha$.



① Find the length of the following confidence intervals

(i) $P[-1.65 \le \mu \le 3] = 0.95$

(ii) $P[-1.68 \le \mu \le 2.7] = 0.95$

(iii) $P[-1.70 \le \mu \le 2.54] = 0.99$

(iv) $P[-1.96 \le \mu \le 1.96] = 0.95$.

Sol: length $L = T_2 - T_1$

(i) $L = 3 - (-1.65) = 4.65$

(ii) $L = 2.7 - (-1.68) = 4.38$

(iii) $L = 4.34$

(iv) $L = 3.92$

② Find the lower and upper confidence limits and also confidence coefficient of the following confidence intervals.

(i) $P(0 \le \theta \le 1.5) = 0.90$ (ii) $P(-1 \le \theta \le 2) = 0.95$

(iii) $P(-2 \le \theta \le 2) = 0.98$ (iv) $P(-2.5 \le \theta \le 2.5) = 0.99$.

Sol: we know that $P(T_1 \le \theta \le T_2) = 1-\alpha$.

let T_1 = lower confidence limit = LCL

T_2 = upper confidence limit = UCL

$1-\alpha$ = confidence coefficient = CC.

(i) LCL = 0, UCL = 1.5, CC = 0.90

(ii) LCL = -1, UCL = 2, CC = 0.95

(iii) LCL = -2, UCL = 2, CC = 0.98

(iv) LCL = -2.5, UCL = 2.5, CC = 0.99.

Confidence interval for population mean:

(1) **Large samples:** If the sample size $n \geq 30$, then the sample is said to be large sample.

(A) When population variance (σ^2) is known:

The confidence interval is

$$\bar{x} \mp z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = \left[\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

where $\bar{x}$ = sample mean
n = sample size
σ = standard Deviation of population
α = confidence level.

(B) When population variance (σ^2) is unknown.

The confidence interval is

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

where s = standard Deviation of sample.

(2) **Small samples:** If the sample size $n < 30$, then the sample is said to be small sample.

(A) When population variance is known:

The confidence interval is

$$\bar{x} \mp t_{(n-1, \alpha/2)} \cdot \frac{\sigma}{\sqrt{n}} = \left[\bar{x} - t_{(n-1, \alpha/2)} \cdot \frac{\sigma}{\sqrt{n}} , \bar{x} + t_{(n-1, \alpha/2)} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

(B) When population variance is unknown:

The confidence interval is

$$\bar{x} \mp t_{(n-1, \alpha/2)} \cdot \frac{s}{\sqrt{n}} = \left[\bar{x} - t_{(n-1, \alpha/2)} \cdot \frac{s}{\sqrt{n}} , \bar{x} + t_{(n-1, \alpha/2)} \cdot \frac{s}{\sqrt{n}} \right].$$

Table values to find $z_{\alpha/2}$.

	90%	95%	98%	99%
$1-\alpha$	0.90	0.95	0.98	0.99
α	0.10	0.05	0.02	0.01
$\alpha/2$	0.05	0.025	0.01	0.005
$z_{\alpha/2}$	1.645	1.96	2.33	2.58

① The mean life of the tyres manufactured by a company follows normal distribution with standard deviation of 3200 kms. A sample of 250 tyres is taken and it is found that the average life of the tyres is 50000 kms with a standard deviation of 3500 kms. Establish the 99% confidence interval within which the mean life of tyres of the company is expected to lie.

Sol: Given, $n = 250$, $\sigma = 3200$, $\bar{x} = 50000$, $s = 3500$.

$1 - \alpha = 99\% = 0.99$

$\alpha = 1 - 0.99 = 0.01$

$\alpha/2 = 0.005$

$z_{\alpha/2} = 2.58$

Here the population variance is known and it is a large sample.

∴ The confidence interval is

$$\bar{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

$$= \left[50000 - 2.58\left(\frac{3200}{\sqrt{250}}\right), 50000 + 2.58\left(\frac{3200}{\sqrt{250}}\right)\right]$$

$$= [49,477.84, 50,522.15]$$

∴ 99% confidence interval for mean life of tyres is [49,477.84, 50,522.15]

② Certain refined oil is packed in tins holding 15 kg each. The filling machine maintains this but have a S.D 0.30 kg. A sample of 200 tins is taken from the production line. If sample mean is 15.25 kg then find the 95% confidence interval for the average weight of oil tins.

Sol: Given,

$n = 200$, $\sigma = 0.30$, $\bar{x} = 15.25$

$1 - \alpha = 95\% = 0.95$

$\alpha = 1 - 0.95 = 0.05$

$\alpha/2 = 0.0025$

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

Here the population variance is known and it is a large sample.

∴ The confidence interval is

$$\bar{x} \mp z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} = \left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right]$$

$$= \left[15.25 - (1.96)\cdot\frac{0.30}{\sqrt{200}}, 15.25 + (1.96)\frac{0.30}{\sqrt{200}}\right]$$

$$= [15.20, 15.29]$$

∴ 95% confidence interval for average weight is [15.20, 15.29]

(3) Sample mean of weights (in kg) of 150 students of SITAMS is found to be 65 kg with S.D 12 kg. Find the 95% confidence limits in which the average weight of all students of SITAMS expected to lie.

Sol: Given, $n = 150$, $\bar{x} = 65$, $s = 12$,

$1-\alpha = 95\%$, $z_{\alpha/2} = 1.96$.

Confidence interval is

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

$$= \left[65 - (1.96)\frac{12}{\sqrt{150}}, 65 + (1.96)\frac{12}{\sqrt{150}}\right]$$

$$= [63.07, 66.92].$$

$\therefore$ 95% confidence interval for average weight is [63.07, 66.92]

(4) It is known that the average weight of students of a study centre of SITAMS follows normal distribution. To estimate the average weight, a sample of 10 students is taken from this study centre and measured their weights (in kg) which are given below:

48, 50, 62, 75, 80, 60, 70, 56, 52, 77.

Compute the 95% confidence interval for the average weight of students of study centre of SITAMS.

Sol: Given, $n = 10$

The weight of 10 students are 48, 50, 62, 75, 80, 60, 70, 56, 52, 77.

S.No.	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	48	$48-63 = -15$	$(-15)^2 = 225$
2	50	-13	169
3	62	-1	1
4	75	12	144
5	80	17	289
6	60	-3	9
7	70	7	49
8	56	-7	49
9	52	-11	121
10	77	14	196

$\Sigma x_i = 630$ $\quad$ $\Sigma (x_i - \bar{x})^2 = 1252$.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{630}{10} = 63$$

$$s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1} = \frac{1252}{10-1} = 139.11$$

$$s = \sqrt{139.11} = 11.79.$$

$1-\alpha = 95\%$

$\Rightarrow \alpha/2 = 0.025$.

Here, Population variance is unknown and it is a small sample. Therefore, the confidence interval is

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

$$= \left[63 - t_{(10-1, 0.025)}\cdot\frac{11.79}{\sqrt{10}}, 63 + t_{(9, 0.025)}\cdot\frac{11.79}{\sqrt{10}}\right]$$

$$= \left[63 - 2.26\left(\frac{11.79}{\sqrt{10}}\right), 63 + 2.26\left(\frac{11.79}{\sqrt{10}}\right)\right]$$

$$= [54.5740, 71.43]$$

$\therefore$ 95% confidence interval for the average weight is [54.5740, 7643]

(5) It is known that the average height of cadets of a centre follows normal distribution. A sample of 6 cadets of the centre was taken and measured their heights (in inch) which are given below:

70, 72, 80, 82, 78, 80.

From this data, estimate the 95% confidence limits for the average height of cadets of the particular centre.

Sol: Given, $n = 6$. Height of 6 students are 70, 72, 80, 82, 78, 80.

$1 - \alpha = 95\% = 0.95 \Rightarrow z_{\alpha/2} = 1.96$.

S.No	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	70	-7	49
2	72	-5	25
3	80	3	9
4	82	5	25
5	78	1	1
6	80	3	9

$\Sigma x_i = 462$ $\quad \Sigma (x_i - \bar{x})^2 = 118$

Sample mean, $\bar{x} = \frac{\Sigma x_i}{n} = \frac{462}{6} = 77$.

Sample variance, $s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1} = \frac{118}{6-1} = \frac{118}{5} = 23.6$

$s = \sqrt{23.6} = 4.85$.

Here population variance is unknown and it is a small sample.

∴ The confidence interval is

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

$$= \left[77 - t_{(6-1, 0.025)} \cdot \frac{4.85}{\sqrt{6}}, 77 + t_{(5, 0.025)} \frac{4.85}{\sqrt{6}}\right]$$

$$= \left[77 - 2.571 \left(\frac{4.85}{\sqrt{6}}\right), 77 + (2.571)\left(\frac{4.85}{\sqrt{6}}\right)\right]$$

$$= [71.890, 82.109]$$

∴ 95% confidence interval for average height is $[71.890, 82.109]$

Confidence interval for population proportion:

For many cases it is not possible to find variance or mean, in such cases we use population proportion.

Eg: 1) The proportion of female in population
2) The Diabetics patients in a hospital.
3) Number of defective bolts in a manufacturing company.
4) Number of science books in a library.

For this we use the confidence interval

$$p \mp z_{\alpha/2} \sqrt{\frac{pq}{n}} = \left[p - z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}, p + z_{\alpha/2} \sqrt{\frac{pq}{n}}\right]$$

where, $p = \frac{x}{n}$.

n = Total no. of sample observations.

x = Total no. of observations satisfying the particular character.

$q = 1-p$.

This method is applicable only if $np > 5$ and $nq > 5$

(1) A sample of 200 voters is chosen at random from all voters in a given city. 60% of them are in favour of a particular candidate. If large no. of voters caste their vote the find the 99% and 95% confidence intervals for the proportion of voters in favour of a particular candidate.

Sol: Given, $n = 200$, $p = 60\% = 0.6$

$q = 1-p = 1-0.6 = 0.4$.

For 99% confidence interval,

$1-\alpha = 99\% \Rightarrow z_{\alpha/2} = 2.58$.

Confidence interval is given by

$$p \mp z_{\alpha/2}\sqrt{\frac{pq}{n}} = \left[p - z_{\alpha/2}\sqrt{\frac{pq}{n}},\ p + z_{\alpha/2}\sqrt{\frac{pq}{n}}\right]$$

$$= \left[0.6 - (2.58)\sqrt{\frac{0.6\times0.4}{200}},\ 0.6 + (2.58)\sqrt{\frac{0.6\times0.4}{200}}\right]$$

$$= [0.5106,\ 0.6893]$$

For 95% confidence interval,

$1-\alpha = 95\% \Rightarrow z_{\alpha/2} = 1.96$.

Confidence interval is

$$p \mp z_{\alpha/2}\sqrt{\frac{pq}{n}} = \left[p - z_{\alpha/2}\sqrt{\frac{pq}{n}},\ p + z_{\alpha/2}\sqrt{\frac{pq}{n}}\right]$$

$$= \left[0.6 - (1.96)\sqrt{\frac{0.6\times0.4}{200}},\ 0.6 + (1.96)\sqrt{\frac{0.6\times0.4}{200}}\right]$$

$$= [0.5321,\ 0.6678]$$

(2) A random sample of 400 apples was taken from a large consignment and 80 were found to be bad. Obtain the 99% confidence limits for the proportion of bad apples in the consignment.

Sol: Given, $n = 400$, $x = 80$, $p = \frac{x}{n} = \frac{80}{400} = 0.2$.

$q = 1-p = 1-0.2 = 0.8$,

$1-\alpha = 99\% \Rightarrow z_{\alpha/2} = 2.58$.

The confidence interval is given by

$$p \mp z_{\alpha/2}\sqrt{\frac{pq}{n}} = \left[p - z_{\alpha/2}\sqrt{\frac{pq}{n}},\ p + z_{\alpha/2}\sqrt{\frac{pq}{n}}\right]$$

$$= \left[0.2 - (2.58)\sqrt{\frac{0.2\times0.8}{400}},\ 0.2 + (2.58)\sqrt{\frac{0.2\times0.8}{400}}\right]$$

$$= [0.1484,\ 0.2516]$$

$\therefore$ 99% confidence limits for the proportion of bad apples are $[0.1484, 0.2516]$

Confidence interval for population variance

(A) When mean is known:

The confidence interval is

$$\left[\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{(n,\alpha/2)}} , \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{(n,1-\alpha/2)}} \right]$$

(B) When mean is unknown:

The confidence interval is

$$\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\chi^2_{(n-1,\alpha/2)}} , \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\chi^2_{(n-1,1-\alpha/2)}} \right]$$

(1) Diameter of steel ball bearing produced by a company is known to be normally distributed. To know the variation in the diameter of steel ball bearings, the product manger takes a random sample of 10 ball bearings from the lot having average diameter 5.0 cm and measures diameter (in cm) of each selected ball bearing. The results are given below:

S.No.	1	2	3	4	5	6	7	8	9	10
Diameter	5.2	5.0	5.1	5.0	5.2	4.9	5.0	5.0	5.1	5.1

Find the 95% confidence interval for variance in the diameter of steel ball bearings of the lot from which the sample is drawn.

Sol: Given, $n = 10$, $\mu = 5$, $1-\alpha = 95\%$ $\Rightarrow \alpha/2 = 0.025$.

S.No	x_i	$x_i-\mu$	$(x_i-\mu)^2$
1	5.2	0.2	0.04
2	5.0	0	0
3	5.1	0.1	0.01
4	5.0	0	0
5	5.2	0.2	0.04
6	4.9	-0.1	0.01
7	5.0	0	0
8	5.0	0	0
9	5.1	0.1	0.01
10	5.1	0.1	0.01

$\Sigma(x_i-\mu)^2 = 0.12$.

$$\chi^2_{(n,\alpha/2)} = \chi^2_{(10,0.025)} = 20.48$$

$$\chi^2_{(n,1-\alpha/2)} = \chi^2_{(10,0.975)} = 3.25.$$

The confidence interval is given by

$$\left[\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{(n,\alpha/2)}} , \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{(n,1-\alpha/2)}} \right] = \left[\frac{0.12}{20.48} , \frac{0.12}{3.25} \right]$$

$$= [0.00585, 0.0369]$$

$\therefore$ 95% confidence interval for variance in the diameter is $[0.00585, 0.0369]$

(2) A random sample of 10 members is taken from a factory in wages (in 100's) per month are given below:

48, 50, 62, 75, 80, 60, 70, 56, 52, 77.

Obtain the 95% confidence interval for the variance of wages of the workers of the factory.

Sol: Given, $n = 10$

$1 - \alpha = 95\% = 0.95$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$\chi^2_{(n-1, \alpha/2)} = \chi^2_{(9, 0.025)} = 19.02$

$\chi^2_{(n-1, 1-\alpha/2)} = \chi^2_{(9, 0.975)} = 2.70$.

S.No.	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	48	-15	225
2	50	-13	169
3	62	-1	1
4	75	12	144
5	80	17	289
6	60	-3	9
7	70	7	49
8	56	-7	49
9	52	-11	121
10	77	-14	196

$\Sigma x_i = 630$ $\quad \Sigma (x_i - \bar{x})^2 = 1252$

Sample mean,

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{630}{10} = 63$$

The confidence interval is given by

$$\left[\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\chi^2_{(n, \alpha/2)}}, \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\chi^2_{(n-1, 1-\alpha/2)}} \right] = \left[\frac{1252}{19.02}, \frac{1252}{2.70} \right]$$

$$= [63.85, 463.7].$$

(3) A study of variance in weights of soldiers was made and its known that the mean of weights of soildiers follow the normal distribution. A sample of 12 sotdiers is taken from the soldiers population and sample variance is found to be 60 pound. Estimate the 95% percent confidence interval for the variance of soldiers weight for population.

Sol: Given, $n = 12$, $s^2 = 60$

$$\Rightarrow \frac{\Sigma (x_i - \bar{x})^2}{n-1} = 60$$

$$\Sigma (x_i - \bar{x})^2 = 60(n-1) = 60(12-1) = 60(11) = 660.$$

$1 - \alpha = 95\% = 0.95$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$\chi^2_{(n-1,\alpha/2)} = \chi^2_{(11,0.025)} = 21.92$

$\chi^2_{(n-1,1-\alpha/2)} = \chi^2_{(11,0.975)} = 3.82$

The confidence interval is

$$\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\chi^2_{(n-1,\alpha/2)}} , \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\chi^2_{(n-1,1-\alpha/2)}} \right]$$

$$= \left[\frac{660}{21.92} , \frac{660}{3.82} \right]$$

$$= [30.10, 172.77]$$

Difference between the mean and ratio of two normal populations:

$$(\bar{x}_1 - \bar{x}_2) \mp z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$ [when population variance is known]

$$(\bar{x}_1 - \bar{x}_2) \mp t_{(n-1,\alpha/2)} \cdot s \cdot \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$ [when population variance is unknown]

where s = combined sample variance.

$\bar{x}_1$ and $\bar{x}_2$ are sample means of population 1 & 2

σ_1^2 and σ_2^2 are population variances of population 1 and population 2.

n_1 and n_2 are sample sizes of population 1 & population 2.

Ratio of two normal populations:

The confidence interval is given by $\left[\frac{1}{F_{(\alpha/2)}} \cdot \frac{s_1^2}{s_2^2} , \frac{1}{F_{(1-\alpha/2)}} \cdot \frac{s_1^2}{s_2^2} \right]$

where (n_1-1, n_2-1) are the degrees of freedom of F distribution

Q: If $x_1 = -5$, $x_2 = 4$, $x_3 = 2$, $x_4 = 6$, $x_5 = 1$, $x_6 = 4$, $x_7 = 0$, $x_8 = 10$ and $x_9 = 7$ are the sample observations taken from normal population $N(\mu, \sigma^2)$. Obtain confidence interval for σ^2.

Sol: Given, $x_1 = -5$, $x_2 = 4$, $x_3 = 2$, $x_4 = 6$, $x_5 = 1$, $x_6 = 4$, $x_7 = 0$, $x_8 = 10$ and $x_9 = 7$.

Sample mean, $\bar{x} = \frac{\sum x_i}{n} = \frac{27}{9} = 3$.

S.No	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	-5	-8	64
2	4	1	1
3	2	-1	1
4	6	3	9
5	-1	-4	16
6	4	1	1
7	0	-3	9
8	10	7	49
9	7	4	16
	$\Sigma x_i = 27$		$\Sigma (x_i - \bar{x})^2 = 166$

Sample variance, $s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1} = \frac{166}{9-1} = \frac{166}{8} = 20.75$.

The confidence interval is given by,

$$\left[\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\chi^2_{(n-1, \alpha/2)}} , \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\chi^2_{(n-1, 1-\alpha/2)}} \right] \quad \text{---} (1)$$

For 99%:

$\chi^2_{(n-1, \alpha/2)} = \chi^2_{(8, 0.005)} = 15.51$

$\chi^2_{(n-1, 1-\alpha/2)} = \chi^2_{(8, 0.995)} = 2.73$

(1) $\Rightarrow \left[\frac{166}{15.51} , \frac{166}{2.73} \right] = [10.702, 60.805]$

For 95%:

$\chi^2_{(n-1, \alpha/2)} = \chi^2_{(8, 0.025)} = 17.53$

$\chi^2_{(n-1, 1-\alpha/2)} = \chi^2_{(8, 0.975)} = 2.18$.

(1) $\Rightarrow \left[\frac{166}{17.53} , \frac{166}{2.18} \right] = [9.469, 76.14]$

UNIT- IV : TESTING OF HYPOTHESES

Hypotheses :-

In previous units, we find point estimate and interval estimate value. Rather than estimating the value of parameter, we need to decide that values are accepted & rejected from a statement about the parameter.

These statements is called Hypotheses and Decision making procedure about the hypothesis is called Testing of Hypotheses.

There are 2 types of Hypotheses.

1. Null Hypothesis
2. Alternate Hypothesis.

1. Null Hypothesis: It is denoted by H_0. H_0 represents the assumption of "There is no difference between estimated value and actual value of the parameter" i.e., H_0 & $\theta = \theta_0$

Here θ is the actual value, θ_0 is the estimated value.

2. Alternate Hypothesis: It is denoted by H_1. It is the statement that contradicts the null hypothesis, i.e.,

$H_1 : \theta \neq \theta_0$ (Two tailed)

$H_1 : \theta > \theta_0$ (Right - tailed)

$H_1 : \theta < \theta_0$ (Left - tailed)

Type - I error : Reject H_0 when it is true.

If the Null Hypothesis H_0 is true but it is rejected by test procedure, then the error made is called Type-I error or α error. It is denoted by α.

$$\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 \text{ when } H_0 \text{ is true}).$$

Type - II error : Accept H_0 when it is false.

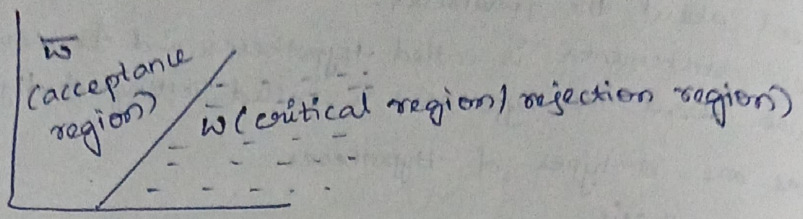
If the Null Hypothesis is false but it is accepted by the test, then the error made is called Type-II error or β-error. It is denoted by β.

$$\beta = P(\text{Type II error}) = P(\text{Accept } H_0 \text{ when } H_0 \text{ is false}).$$

	Accept (True)	Reject (False)
H_0 is true	Correct Decision	Type-I error (α)
H_0 is false	Type-II error (β)	Correct Decision.

Critical regions

If $x_1, x_2, \ldots, x_n$ be sample observations in sample space S. Let us divide the sample space S into two disjoint parts w and $\bar{w}$. The region w consists of the sample points for which the null hypothesis is rejected when it is true.



The region of the sample points for which the null hypothesis is rejected when it is true is called critical region.

Power of the test :

WKT, $\alpha = P$ (Type-I error)

$= P$ (Rejecting H_0 when H_0 is true)

$\alpha = P(x \in w / H_0)$

$\alpha = \int_w L_0 \, dx$

$\beta = P$(Type-II error)

$= P$(Accepting H_0 when H_0 is false)

$\beta = P(x \in \bar{w} / H_1)$

$\beta = \int_{\bar{w}} L_1 \, dx$

We have, $\int_w L_1 \, dx + \int_{\bar{w}} L_1 \, dx = 1$

$$\int_w L_1 \, dx = 1 - \int_{\bar{w}} L_1 \, dx = 1 - \beta.$$

$$\boxed{P(x \in w / H_1) = 1 - \beta.}$$

It is called power function of the test hypothesis H_0 against the alternate hypothesis H_1. The value of the power function at a parameter point is called the power of the test at that point.

Note: A hypothesis of the type $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$ are called simple hypothesis, the type $H_0 : \theta \neq \theta_0$, $H_1 : \theta > \theta_0$, $H_1 : \theta < \theta_0$, $H_0 : \theta \geq \theta_0$ etc.. are called composite hypothesis.

① Given the frequency function $f(x,\theta) = \frac{1}{\theta}, \ 0 \le x \le \theta$
$= 0$, otherwise

and that you are testing the null hypothesis $H_0: \theta = 1$ against $H_1: \theta = 2$, by means of a single observed value of x. What would be the sizes of the type-I and type-II errors, if you choose the interval.
(i) $0.5 \le x$ (ii) $1 \le x \le 1.5$
as the critical regions? Also obtain the power of the test.

Sol: Given, $f(x,\theta) = \frac{1}{\theta}, \ 0 \le x \le \theta$
$= 0$, otherwise

Null hypothesis $H_0: \theta = 1$
Alternate hypothesis $H_1: \theta = 2$.

(i) Given,
critical region $w = \{x : 0.5 \le x\}$
$\bar{w} = \{x : 0.5 \ge x\}$
$= \{x : x \le 0.5\}$

WKT, $\alpha = P(\text{Type-I error})$
$= P(x \in w / H_0)$
$= P(0.5 \le x / \theta = 1)$
$= \int_{0.5}^{1} f(x,\theta)$ At $\theta = 1 \ dx$.
$= \int_{0.5}^{1} f(x,1) \, dx$
$= \int_{0.5}^{1} 1 \cdot dx$.
$= [x]_{0.5}^{1}$
$= 1 - 0.5$
$\alpha = 0.5$.

Similarly $\beta = P(\text{Type-II error})$
$= P(x \in \bar{w} / H_1)$
$= P(x \le 0.5) / \theta = 2)$
$= \int_{0}^{0.5} f(x,\theta)$ At $\theta = 2 \ dx$.
$= \int_{0}^{0.5} \frac{1}{2} dx = \frac{1}{2}[x]_0^{0.5}$
$= \frac{1}{2} \times [0.5 - 0]$.
$\beta = 0.25$
Power of test $= 1 - \beta = 1 - 0.25 = 0.75$

(ii) Given, $w = \{x : 1 \le x \le 1.5\}$.

$$\alpha = P(x \in w / H_0)$$
$$= P(1 \le x \le 1.5 / \theta = 1)$$
$$= \int_1^{1.5} f(x,\theta)_{At\ \theta=1}\, dx$$
$$= \int_1^{1.5} 0\, dx$$

For $H_0 : \theta = 1$, $f(x,\theta) = 1$ for $x \ge 1$.

$$\alpha = 0.$$

$\overline{w} = \{x : x \in (0,1) \cup (1.5, \theta)\}$

$$\beta = P(x \in \overline{w} / H_1)$$
$$= \int_0^1 f(x,\theta)\, dx + \int_{1.5}^{\theta} f(x,\theta)\, dx \quad [At\ \theta = 2]$$
$$= \int_0^1 \frac{1}{\theta} dx + \int_{1.5}^{\theta} \frac{1}{\theta} dx \quad [At\ \theta = 2]$$
$$= \int_0^1 \frac{1}{2} dx + \int_{1.5}^{2} \frac{1}{2} dx$$
$$= \frac{1}{2}[x]_0^1 + \frac{1}{2}[x]_{1.5}^{2}$$
$$= \frac{1}{2}[1-0] + \frac{1}{2}[2-1.5]$$
$$= \frac{1}{2} + \frac{1}{2}[0.5]$$
$$= \frac{1}{2} + 0.25$$
$$= 0.5 + 0.25$$
$$\beta = 0.75$$

Power of a test $= 1 - \beta = 1 - 0.75 = 0.25$.

(2) If $x \ge 1$ is the critical region for testing $H_0 : \theta = 2$ against the alternative $\theta = 1$ on the basis of single observation from the population with $f(x,\theta) = \theta e^{-\theta x}$, $0 \le x \le \infty$. Obtain the type-I and type-II errors.

Sol: Given,

$w = \{x : x \ge 1\}$

$H_0 : \theta = 2$

$H_1 : \theta = 1$

$f(x,\theta) = \theta e^{-\theta x}$, $0 \le x \le \infty$.

$\overline{w} = \{x : x < 1\}$

$$\alpha = P(x \in w / H_0)$$

$$= P(x \geq 1 / \theta = 2)$$

$$= \int_1^\infty f(x,\theta)\,dx \quad [\text{At } \theta = 2]$$

$$= \int_1^\infty \theta e^{-\theta x}\,dx \quad [\text{At } \theta = 2]$$

$$= \int_1^\infty 2e^{-2x}\,dx$$

$$= 2\left[\frac{e^{-2x}}{-2}\right]_1^\infty$$

$$= \frac{2}{-2}\left[0 - e^{-2(1)}\right]$$

$$= -\left[\frac{1}{e^\infty} - e^{-2}\right] = -\left[\frac{1}{\infty} - e^{-2}\right]$$

$$= -[0 - e^{-2}]$$

$$\alpha = e^{-2}.$$

$$\beta = P(x \in \bar{w} / H_1)$$

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

$$= \int_0^1 f(x,\theta)\,dx \quad [\text{At } \theta = 1]$$

$$= \int_0^1 \theta e^{-\theta x}\,dx \quad [\text{At } \theta = 1]$$

$$= \int_0^1 1 e^{-1x}\,dx$$

$$= \left[\frac{e^{-x}}{-1}\right]_0^1$$

$$= -[e^{-1} - e^0]$$

$$= -\left[\frac{1}{e} - \frac{1}{e^0}\right]$$

$$= -\left[\frac{1}{e} - 1\right]$$

$$= 1 - \frac{1}{e}$$

$$\beta = \frac{e-1}{e}$$

(3) Let p be the probability that a coin will fall head in a single toss in order to test $H_0 : P = 1/2$ against $H_1 : P = 3/4$. The coin is tossed five times & H_0 is rejected if more than 3 heads are obtained. Find the probability of Type-I error and power of a test.

Sol: Given,

$H_0 : P = 1/2$.

$H_1 : P = 3/4$

$n = 5$

If the random variable X denotes the number of heads in small n tosses of a coin then,

$X \sim B(n, p)$ so that $P(X = x) = {}^nC_x p^x q^{n-x} = {}^5C_x p^x q^{5-x}$.

Given, H_0 is rejected if more than 3 heads are obtained.

Therefore, critical region is defined by

$w = \{x : x \geq 4\}$

$\bar{w} = \{x : x < 4\}$

$\alpha = P(x \in w / H_0)$

$= P[x \geq 4 / P = 1/2]$

$= P(X = 4) + P(X = 5)$ [At $P = 1/2$]

$= {}^5C_4 p^4 q^{5-4} + {}^5C_5 p^5 q^{5-5}$ [At $P = 1/2$, $q = 1 - P = 1 - 1/2 = 1/2$]

$= {}^5C_4 (\frac{1}{2})^4 (\frac{1}{2}) + {}^5C_5 (\frac{1}{2})^5 (\frac{1}{2})^0$

$\alpha = \frac{3}{16}$.

$\beta = P(x \in \bar{w} / H_1)$

$= P(x < 4 / P = 3/4)$

$= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$ [At $P = \frac{3}{4}$]

$= {}^5C_0 p^0 q^{5-0} + {}^5C_1 p^1 q^{5-1} + {}^5C_2 p^2 q^{5-2} + {}^5C_3 p^3 q^{5-3}$ [At $P = 3/4$, $q = 1 - P = 1 - 3/4$, $q = 1/4$]

$= {}^5C_0 (\frac{3}{4})^0 (\frac{1}{4})^5 + {}^5C_1 (\frac{3}{4})^1 (\frac{1}{4})^4 + {}^5C_2 (\frac{3}{4})^2 (\frac{1}{4})^3 + {}^5C_3 (\frac{3}{4})^3 (\frac{1}{4})^2$

$= \frac{47}{128}$

$\beta = 0.3671$

Power of a test $= 1 - \beta = 1 - 0.3671 = 0.6329$.

Most powerful test (MP Test): The critical region w is the most powerful critical region of size 2 for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ if $P(x \in w / H_0) = \int_w L_0 dx = \alpha$ — (1) and $P(x \in w / H_1) \geq P(x \in w_1 / H_1)$, for every other critical region w_1 satisfying eqn (1).

Note: neyman pearson lemma is only applicable for simple hypothesis $H_0 : \theta = \theta_0, H_1 : \theta = \theta_1$ not composite hypothesis $H_1 : \theta \neq \theta_1$

Uniformly Most Powerful Test (UMP Test):

The region w is called uniformly most powerful critical region of size α (and the corresponding test as uniformly most powerful test of level α) for testing $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_1$ that is $H_1: \theta = \theta_1 \neq \theta_0$ if

$P(x \in w / H_0) = \int_w L_0 dx = \alpha$ — (1) and

$P(x \in w / H_1) \geq P(x \in w_1 / H_1)$ for all $\theta \neq \theta_0$ and for every other region w_1 satisfying eqn (1).

Neyman J. and Pearson E.S. Lemma:

Let $k > 0$ be a constant and w be the critical region of size α such that

$$w = \left\{ x \in S : \frac{f(x, \theta_1)}{f(x, \theta_0)} > k \right\}$$

$$\overline{w} = \left\{ x \in S : \frac{f(x, \theta_1)}{f(x, \theta_0)} \leq k \right\} \quad (\theta)$$

$$\overline{w} = \left\{ x \in S : \frac{L_1}{L_0} > k \right\}, \quad \overline{w} = \left\{ x \in S : \frac{L_1}{L_0} \leq k \right\}.$$

where L_0 and L_1 are the likelihood functions of the sample observations $x = \{x_1, x_2, \ldots, x_n\}$ under H_0 and H_1 respectively. Then, w is the most powerful critical region of the test hypotheses $H_0: \theta = \theta_0$ against the alternative $H_1: \theta = \theta_1$.

(1) Suppose X is a single observation sample from a population with pdf $f(x) = \theta x^{\theta - 1}$ for $0 < x < 1$. Find a test with the best critical region, that is, find the most powerful test with significance level $\alpha = 0.05$ for testing the null hypothesis $H_0: \theta = 3$ against the alternative hypothesis $H_1: \theta = 2$.

Sol: Given, $f(\theta) = \theta x^{\theta - 1}$ for $0 < x < 1$.

Since both $H_0: \theta = 3$ and $H_1: \theta = 2$, are simple hypotheses. We can apply Neyman–Pearson Lemma to find the best critical region, which is

$$\frac{f(x, \theta_1)}{f(x, \theta_0)} > k$$

$$\frac{\theta_1 x^{\theta_1 - 1}}{\theta_0 x^{\theta_0 - 1}} > k$$

$$\frac{2x^{2-1}}{3x^{3-1}} > k$$

$$\frac{2x}{3x^2} > k$$

$$\frac{2}{3x} > k$$

$$\frac{2}{3k} > x$$

$\Rightarrow x < \frac{2}{3k} = k^*$ (say)

$x < k^*$

$\therefore w = \{x : x < k^*\}$

$\alpha = P(x \in w / H_0)$

$0.05 = P(x < k^*/\theta = 3)$

$$= \int_0^{k^*} f(x,\theta)\,dx \quad [\text{at } \theta = 3]$$

$$= \int_0^{k^*} \theta x^{\theta-1}\,dx \quad [\text{at } \theta = 3]$$

$$0.05 = \int_0^{k^*} 3x^{3-1}\,dx$$

$$= \int_0^{k^*} 3x^2\,dx$$

$$= 3\left[\frac{x^3}{3}\right]_0^{k^*}$$

$$= \frac{3}{3}\left[(k^*)^3 - 0\right]$$

$0.05 = (k^*)^3$

$k^* = (0.05)^{1/3} = 0.368$

$\therefore$ Best critical region is $w = \{x : x < 0.368\}$

(2) Examine whether a best critical region exists for a testing of null hypothesis $H_0: \theta = \theta_0$ against the alternate hypothesis $H_1: \theta = \theta_1 > \theta_0$ for the parameter θ of the distribution $f(x,\theta) = \frac{1+\theta}{(x+\theta)^2}$, $1 \le x < \infty$.

Sol: Given,

$$f(x,\theta) = \frac{1+\theta}{(x+\theta)^2}, \quad 1 \le x < \infty \quad \text{—①}$$

$H_0: \theta = \theta_0$

$H_1: \theta = \theta_1 > \theta_0$

By Neyman Pearson lemma.

$$\frac{L_1}{L_0} > k \Rightarrow L_1 > L_0 k \quad \text{—②}$$

By the definition of likelihood function,

$$L = f(x_1, \theta) \cdot f_2(x_2, \theta) \cdots f(x_n, \theta)$$

$$= \frac{1+\theta}{(x_1+\theta)^2} \cdot \frac{1+\theta}{(x_2+\theta)^2} \cdots \frac{1+\theta}{(x_n+\theta)^2}$$

$$L = \frac{(1+\theta)^n}{\prod_{i=1}^{n} (x_i+\theta)^2}$$

(2) $L_1 > L_0 \, k$

$$\frac{(1+\theta_1)^n}{\prod_{i=1}^{n} (x_i+\theta_1)^2} > \frac{(1+\theta_0)^n}{\prod_{i=1}^{n} (x_i+\theta_0)^2} \, k.$$

$$(1+\theta_1)^n \prod_{i=1}^{n} (x_i+\theta_0)^2 > (1+\theta_0)^n \prod_{i=1}^{n} (x_i+\theta_1)^2 \, k.$$

Taking log on Both sides

$$\log[1+\theta_1]^n + \log\left[\prod_{i=1}^{n} (x_i+\theta_0)^2\right] > \log(1+\theta_0)^n + \log\left[\prod_{i=1}^{n} (x_i+\theta_1)^2\right] + \log k$$

$$\Rightarrow \log(1+\theta_1)^n + \sum_{i=1}^{n} \log(x_i+\theta_0)^2 > \log(1+\theta_0)^n + \sum_{i=1}^{n} \log(x_i+\theta_1)^2 + \log k.$$

$$\Rightarrow \sum_{i=1}^{n} \log(x_i+\theta_0)^2 - \sum_{i=1}^{n} \log(x_i+\theta_1)^2 > \log(1+\theta_0)^n - \log(1+\theta_1)^n + \log k$$

$$\Rightarrow \sum_{i=1}^{n} \log\left[\frac{(x_i+\theta_0)^2}{(x_i+\theta_1)^2}\right] > \log\left(\frac{(1+\theta_0)^n}{(1+\theta_1)^n}\right) + \log k$$

$$\Rightarrow 2\sum_{i=1}^{n} \log\left[\frac{x_i+\theta_0}{x_i+\theta_1}\right] > \log\left[\frac{(1+\theta_0)^n}{(1+\theta_1)^n}\right] + \log k$$

The term $\sum_{i=1}^{n} \log\left[\frac{(x_i+\theta_0)^2}{(x_i+\theta_1)^2}\right]$ cannot be put in the form of a function of sample observations i.e, not depending on the hypothesis. Hence, there is no best critical region in this case.

(3) Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and variance 16. Find the test with best critical region that is find the most powerful test with sample size $n=16$ and significance level $\alpha = 0.05$ to test the simple null hypothesis $H_0: \mu = 10$ against the alternative hypotheses $H_1: \mu = 15$.

Sol: Given,

$H_0: \mu = 10$

$H_1: \mu = 15$

By the definition of normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Given $\sigma^2 = 16$, $\sigma = 4$

$$f(x) = \frac{1}{4\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{4}\right)^2}$$

$$= \frac{1}{\sqrt{4^2}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{4}\right)^2}$$

$$f(x) = \frac{1}{\sqrt{32\pi}} e^{-\frac{1}{32}(x-\mu)^2}$$

By the definition of likelihood function,

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

$$= \frac{1}{\sqrt{32\pi}} e^{-\frac{1}{32}(x_1-\mu)^2} \cdot \frac{1}{\sqrt{32\pi}} e^{-\frac{1}{32}(x_2-\mu)^2} \cdots \frac{1}{\sqrt{32\pi}} e^{-\frac{1}{32}(x_n-\mu)^2}$$

$$L = \left[\frac{1}{\sqrt{32\pi}}\right]^n e^{-\frac{1}{32}\sum(x_i-\mu)^2}$$

Given $n = 16$

$$L = \left[\frac{1}{\sqrt{32\pi}}\right]^{16} e^{-\frac{1}{32}\sum_{i=1}^{16}(x_i-\mu)^2}$$

$$L = \left[\frac{1}{32\pi}\right]^8 \cdot e^{-\frac{1}{32}\sum_{i=1}^{16}(x_i-\mu)^2}$$

By Neyman Pearson lemma,

$$\frac{L_1}{L_0} > k$$

$$\frac{\left[\frac{1}{32\pi}\right]^8 e^{-\frac{1}{32}\sum_{i=1}^{16}(x_i-\mu_1)^2}}{\left[\frac{1}{32\pi}\right]^8 e^{-\frac{1}{32}\sum_{i=1}^{16}(x_i-\mu_0)^2}} > k$$

$$\Rightarrow e^{\left[-\frac{1}{32}\sum_{i=1}^{16}(x_i-15)^2 + \frac{1}{32}\sum_{i=1}^{16}(x_i-10)^2\right]} > k$$

Taking log on both sides

$$\Rightarrow -\frac{1}{32}\sum_{i=1}^{16}(x_i-15)^2 + \frac{1}{32}\sum_{i=1}^{16}(x_i-10)^2 > \log k$$

$$\Rightarrow \frac{1}{32}\sum_{i=1}^{16}\left[-(x_i-15)^2 + (x_i-10)^2\right] > \log k$$

$$\Rightarrow \frac{1}{32}\sum_{i=1}^{16}\left[-(x_i^2+225-30x_i) + (x_i^2+100-20x_i)\right] > \log k$$

$$\Rightarrow \frac{1}{32}\sum_{i=1}^{16}\left[-x_i^2 - 225 + 30x_i + x_i^2 + 100 - 20x_i\right] > \log k$$

$$\Rightarrow \frac{1}{32}\sum_{i=1}^{16}\left[10x_i - 125\right] > \log k$$

$$\Rightarrow \frac{10}{32}\sum_{i=1}^{16} x_i - \frac{125}{32}\sum_{i=1}^{16}(1) > \log k$$

$$\Rightarrow \frac{5}{16}\sum_{i=1}^{16} x_i - \frac{125}{32}(16) > \log k$$

$$\Rightarrow 5\left[\frac{\sum_{i=1}^{16} x_i}{16}\right] - \frac{125}{2} > \log k$$

$$\Rightarrow 5\bar{x} - \frac{125}{2} > \log k$$

$$\Rightarrow 5\bar{x} > \log k + \frac{125}{2}$$

$$\Rightarrow \bar{x} > \frac{\log k}{5} + \frac{125}{10}$$

$$\Rightarrow \bar{x} > k^* \text{ (say)}.$$

$X \sim N(\mu, \sigma^2)$

then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$\bar{X} \sim N\left(\mu, \frac{16}{16}\right)$

$\bar{X} \sim N(\mu, 1)$

Standard normal variate

$$Z = \frac{\bar{X}-\mu}{\sigma} = \frac{\bar{X}-\mu}{1}$$

$$Z = \bar{X} - \mu$$

Given significance level $\alpha = 0.05$

$$\alpha = P[x \in w / H_0]$$

$$\alpha = P[\bar{x} > k^* / \mu = 10]$$

$$\alpha = P[Z > k^* - \mu / \mu = 10]$$

$$0.05 = P[Z > k^* - 10]$$

If $\alpha = 0.05$, the table value for critical region is 1.645

$$k^* - 10 = 1.645$$

$$k^* = 1.645 + 10$$

$$k^* = 11.645$$

Critical region is $w = \{x : \bar{x} > k^*\}$

$$\Rightarrow w = \{x : \bar{x} \geq 11.645\}.$$

Likelihood Ratio test: The Neyman-Pearson based on the magnitude of the ratio of two probability density functions provides the best test for testing simple hypothesis against simple alternative hypothesis. The best test in any given situation depends on the nature of population distribution and the form of the alternative hypothesis being considered.

General method of test construction is called the likelihood Ratio (LR) test and it is introduced by Neyman - Pearson for testing a hypothesis, simple or composite, against a simple (or) composite hypothesis.

(or)

A statistical method used to compare the goodness of fit of two hypothesis, a simple null hypothesis and a composite alternative hypothesis, to determine which better explains the data is called Likelihood Ratio test.

Properties of LR test:

In LR test, then P [type-I error] is controlled by suitable choosing the cut-off point λ_0. LR test is UMP if an UMP test at all exists.

The two asymptotic properties of LR tests are

(1) Under certain conditions, LR test is consistent.

(2) Under certain conditions, $-2\log\lambda$ has an asymptotic chi-square distribution.

The quantity λ is a function of the sample observations only and does not involve parameters. λ is a function of random variables and $\lambda > 0$ then $0 \le \lambda \le 1$.

The critical region for testing H_0 (against H_1) is an interval

$$0 < \lambda < \lambda_0 \quad \text{---(1)}$$

where $\lambda_0 < 1$, determined by the distribution λ and the desired P[type-I error]; i.e., λ_0 is given by the equation

$$P(\lambda < \lambda_0 / H_0) = \alpha \quad \text{---(2)}$$

Eqns (1) & (2) are called LR test for testing H_0.

Applications to distributions:

* Test for mean of a Normal distribution:

If a random sample of size n $(n \geq 30)$ has the sample mean $\bar{x}$, test the hypothesis that the population mean μ has the specified value μ_0.

Step 1: Null Hypothesis $H_0: \mu = \mu_0$

Step 2 : Alternate Hypothesis

$H_1: \mu \neq \mu_0$ (two-tailed test)

$H_1: \mu > \mu_0$ (right-tailed test)

$H_1: \mu < \mu_0$ (left-tailed test)

Step 3: Table value
Based on level of significance α, the critical values are given below.

Level of significance (α)	1%	5%	10%
Critical value for two-tailed test	2.58	1.96	1.645
Critical value for right-tailed test	2.33	1.645	1.28
Critical value for left-tailed test	-2.33	-1.645	-1.28

Step 4 : Test statistic

$$Z_{cal} = \frac{\bar{x} - \mu}{(\sigma/\sqrt{n})} \quad [\text{If } \sigma \text{ is known}]$$

$$Z_{cal} = \frac{\bar{x} - \mu}{(s/\sqrt{n})} \quad [\text{If } \sigma \text{ is unknown}]$$

Step 5 : Conclusion

If $Z_{cal} < Z_{tab}$ value then we accept H_0 otherwise we reject H_0.

① An oceanographer wants to check whether the depth of the ocean in a certain region is 57.4 fathoms, what can be conclude at the 0.05 level of significance if readings taken at 40 random locations in the given region yield at a mean of 59.1 fathoms with a SD of 5.2.

Sol: Given,

Population mean $\mu = 57.4$.

Sample size $n = 40$.

Sample mean $\bar{x} = 59.1$

Sample standard deviation $s = 5.2$

level of significance $\alpha = 0.05$.

Step 1 : Null Hypothesis $H_0: \mu = \mu_0$

$H_0: \mu = 57.4$

Step 2: Alternate hypothesis $H_1: \mu \neq 57.4$ (two tailed test)

Step 3: Critical value for $\alpha = 0.05$

$Z_{tab} = 1.96$.

Step 4: Test statistic

$$Z_{cal} = \frac{\bar{x} - \mu}{(S/\sqrt{n})}$$

$$= \frac{59.1 - 57.4}{5.2/\sqrt{40}}$$

$$Z_{cal} = 2.067$$

Step 5: Conclusion

$Z_{cal} = 2.067$

$Z_{tab} = 1.96$

$Z_{cal} > Z_{tab}$

$\therefore$ We reject H_0

Hence the average depth of ocean in a certain region is not equal to 57.4.

(2) According to norms establish mechanical aptitude test, persons who are 18 years old have an average height 73.2 with a S.D of 8.6. If 4 randomly selected persons of the age averaged by 76.7, test the hypothesis $\mu > 73.2$ at the 0.01 level of significance.

Sol: Given,

Population mean $\mu = 73.2$

Population S.D $\sigma = 8.6$

Sample size $n = 4$

Sample mean $\bar{x} = 76.7$

Level of significance $\alpha = 0.01$

Step 1: Null Hypothesis $H_0: \mu = 73.2$.

Step 2: Alternate Hypothesis $H_1: \mu > 73.2$. (right tailed test)

Step 3: For $\alpha = 0.01$, $Z_{tab} = 2.33$.

Step 4: Test statistic, $Z_{cal} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{76.7 - 73.2}{8.6/\sqrt{4}} = 0.8139$.

Step 5: $Z_{cal} = 0.8139$, $Z_{tab} = 2.33$.

$Z_{cal} < Z_{tab}$. $\therefore$ We accept H_0.

Hence the average height is equal to 73.2.

(3) A sample of 900 members has a mean of 3.4 cm and SD 2.61 cm, Is it is a sample from a large population of mean 3.25 cm and SD 2.61 cm. If the population is normal, test the mean with $\alpha = 0.05$.

Sol: Population mean, $\mu = 3.25$

Sample size $n = 900$

Sample mean $\bar{x} = 3.4$

Sample SD $s = 2.61$

Population SD $\sigma = 2.61$

level of significance $\alpha = 0.05$.

Step 1: Null Hypothesis $H_0: \mu = 3.25$

Step 2: Alternate Hypothesis $H_1: \mu \neq 3.25$ (two-tailed test)

Step 3: For $\alpha = 0.05$, $Z_{tab} = 1.96$.

Step 4: $Z_{cal} = \dfrac{\bar{x} - \mu}{(\sigma/\sqrt{n})} = \dfrac{3.4 - 3.25}{2.61/\sqrt{900}} = 1.724$

Step 5: $Z_{cal} = 1.724$

$Z_{tab} = 1.96$

$Z_{cal} < Z_{tab}$ ∴ we accept H_0.

Test for equality of means:

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means of 2 independent large random samples of sizes n_1 and n_2 drawn from 2 populations having means μ_1 and μ_2 and SD σ_1 and σ_2.

To test whether the two population means are equal, we follow the steps below.

Step 1: Null hypothesis $H_0: \mu_1 = \mu_2$

Step 2: Alternate hypothesis $H_1: \mu_1 \neq \mu_2$ (two-tailed test)

$H_1: \mu_1 > \mu_2$ (right-tailed test)

$H_1: \mu_1 < \mu_2$ (left-tailed test).

Step 3: Table value: Based on level of significance α, the critical values are given below.

level of significance (α)	1%	5%	10%
Critical value for two-tailed test	2.58	1.96	1.645.
Critical value for right-tailed test	2.33	1.645	1.28
Critical value for left-tailed test	-2.33	-1.645	-1.28

Step 4: Test statistic

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

→ If the samples have been drawn from same population then $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

$$Z_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

If σ is not known we can use $\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$

→ If σ_1^2 and σ_2^2 are not given but s_1^2 and s_2^2 are given then $Z_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Step 5: Conclusion

If $Z_{cal} < Z_{tab}$ then we accept H_0 otherwise we reject H_0.

① The mean life of sample of 10 electric bulbs was found to be 1456 hours with SD of 423 hours. A second sample of 17 bulbs chosen from different batch showed a mean life of 1280 hours with SD of 398 hours. Is there a significant difference between the means of 2 batches?

Sol: Given,

Sample size of 1st batch = $n_1 = 10$

Sample size of 2nd batch, $n_2 = 17$

Sample mean of 1st batch, $\bar{x}_1 = 1456$ hours

" " " 2nd " , $\bar{x}_2 = 1280$

Sample SD of 1st batch, $s_1 = 423$

" " " 2nd " , $s_2 = 398$

Step 1: Null hypothesis, $H_0 : \mu_1 = \mu_2$ where μ_1 & μ_2 are population means of two batches.

Step 2: Alternate hypothesis $H_1 : \mu_1 \neq \mu_2$ (two-tailed test)

Step 3: For $\alpha = 0.05$ (Assume), $Z_{tab} = 1.96$

Step 4: Test statistic $Z_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

$$= \frac{1456 - 1280}{\sqrt{\frac{(423)^2}{10} + \frac{(398)^2}{17}}}$$

$$Z_{cal} = 1.0669$$

Step 5: Conclusion,

$Z_{cal} = 1.0669$

$Z_{tab} = 1.96.$

$Z_{cal} < Z_{tab}$ ∴ we accept H_0.

∴ we conclude that there is no difference b/w the mean life of electric bulbs of 2 batches.

(2) Samples of students were drawn from two universities and from their weights in kg, mean and SD are calculated and shown below. Make a large sample test to test the significance of difference between the means.

	Mean	SD	Size of the sample
University - A	55	10	400
University - B	57	15	100

Sol: Given,

Size of first sample $n_1 = 400$

Mean " " " $\bar{x}_1 = 55$

SD " " " $S_1 = 10$

Size of second sample $n_2 = 100$

Mean " " " $\bar{x}_2 = 57$

SD " " " $S_2 = 15$

Step 1: Null hypothesis $H_0: \mu_1 = \mu_2$

Step 2: Alternate hypothesis $H_1: \mu_1 \neq \mu_2$ (two-tailed test)

Step 3: For $\alpha = 0.05$ (Assume), $Z_{tab} = 1.96$.

Step 4: Test statistic

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

$$= \frac{|55 - 57|}{\sqrt{\frac{100}{400} + \frac{225}{100}}}$$

$$= |-1.26|$$

$$Z_{cal} = 1.26$$

Step 5: Conclusion

$Z_{cal} = 1.26$, $Z_{tab} = 1.96$

$Z_{cal} < Z_{tab}$ ∴ we accept H_0

Hence there is no significant difference between the means.

(1) The means of 2 large samples of sizes 1000 and 2000 are 67.5 and 68 inches respectively. Can the samples be regarded as drawn from same population whose S.D is 2.5 inches?

Sol: Given,

Size of first sample $n_1 = 1000$

" " second " $n_2 = 2000$

Mean of first sample $\bar{x_1} = 67.5$

" " second " $\bar{x_2} = 68$.

SD of population, $\sigma = 2.5$

Step 1: Null hypothesis, $H_0: \mu_1 = \mu_2$

Step 2: Alternate hypothesis, $H_1: \mu_1 \neq \mu_2$ (two-tailed test)

Step 3: For $\alpha = 0.05$, $Z_{tab} = 1.96$.

Step 4: Test statistic

$$Z_{cal} = \frac{\bar{x_1} - \bar{x_2}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{|67.5 - 68|}{2.5\sqrt{\frac{1}{1000} + \frac{1}{2000}}}$$

$$= |-5.16|$$

$$Z_{cal} = 5.16$$

Step 5: Conclusion.

$Z_{cal} = 5.16$

$Z_{tab} = 1.96$

$Z_{cal} > Z_{tab}$ ∴ we reject H_0.

Hence the samples are not drawn from sample population.

Testing of hypotheses for population variance:

Here, we want to test the hypothesis for the specified value σ_0^2 of the population variance σ^2.

Step 1: Null hypothesis $H_0: \sigma^2 = \sigma_0^2$

Step 2: Alternate hypothesis $H_1: \sigma^2 \neq \sigma_0^2$ (two-tailed test)

$H_1: \sigma^2 > \sigma_0^2$ (right-tailed test)

$H_1: \sigma^2 < \sigma_0^2$ (left-tailed test)

Step 3: Table value Based on ~~critical value~~ α, critical values are given below.

Level of significance (α)	1%	5%	10%
Critical value for two-tailed test	2.58	1.96	1.645
Critical value for right-tailed test	2.33	1.645	1.28
Critical value for left-tailed test	-2.33	-1.645	-1.28

Note: With same Z_{cal} we can test SD

$H_0: \sigma = \sigma_0$

$H_1: \sigma \neq \sigma_0$

$\sigma > \sigma_0$

$\sigma < \sigma_0$

Step 4: Test Statistic

$$Z_{cal} = \frac{s^2 - \sigma^2}{\sigma^2\sqrt{\frac{2}{n}}}$$

Step 5: Conclusion: If $Z_{cal} < Z_{tab}$ we accept H_0 otherwise we reject H_0.

① A random sample of size 65 is taken from a population of big box. of screws and measured their lengths (in mm) which gives sample variance 9. Test the 2 years old population variance 10.5 is still maintained at present with 5% level of significance.

Sol: Given,

Sample size $n = 65$

Sample variance $s^2 = 9$.

Population variance $\sigma^2 = 10.5$

Step 1: Null hypothesis $H_0: \sigma^2 = 10.5$

Step 2: Alternate hypothesis $H_1: \sigma^2 \neq 10.5$ (Two-tailed test)

Step 3: Table value

For $\alpha = 0.05$, $Z_{tab} = 1.96$

Step 4: Test Statistic

$$Z_{cal} = \frac{s^2 - \sigma^2}{\sigma^2\sqrt{\frac{2}{n}}} = \frac{|9 - 10.5|}{10.5\sqrt{\frac{2}{65}}}$$

$$= |-0.81|$$

$$Z_{cal} = 0.81$$

Step 5: Conclusion

$Z_{cal} = 0.81$

$Z_{tab} = 1.96$

$Z_{cal} < Z_{tab}$ $\therefore$ we accept H_0

Hence the population variance 10.5 is still maintained at present.

(2) A random sample of size 120 bulbs is taken from a lot which gives the standard deviation of the life of electric bulbs as seven hours. Test the SD of life of bulbs of the lot is 6 hours at 5% level of significance.

Sol: Given,
Sample size $n = 120$
Sample SD, $s = 7 \Rightarrow s^2 = 49$.
Population SD, $\sigma = 6 \Rightarrow \sigma^2 = 36$.
level of significance, $\alpha = 0.05$.

Step 1: Null Hypothesis $H_0 : \sigma = 6$.
Step 2: Alternate hypothesis $H_1 : \sigma \neq 6$. (two-tailed test)
Step 3: Table value,
For $\alpha = 0.05$, $Z_{tab} = 1.96$.

Step 4: Test statistic,

$$Z_{cal} = \frac{s^2 - \sigma^2}{\sigma^2\sqrt{\frac{2}{n}}} = \frac{49 - 36}{36\sqrt{\frac{2}{120}}} = 2.79$$

Step 5: Conclusion
$Z_{cal} = 2.79$
$Z_{tab} = 1.96$
$Z_{cal} > Z_{tab}$ $\therefore$ we reject H_0.

$\therefore$ Hence SD of life of bulbs of the lot is not 6 hours.

Testing of Hypothesis for two population variances:

Here, we want to test two population variances σ_1^2 and σ_2^2, for this we follow the steps below.

Step 1: Null hypothesis, $H_0 : \sigma_1^2 = \sigma_2^2$
Step 2: Alternate hypothesis, $H_1 : \sigma_1^2 \neq \sigma_2^2$ (two-tailed test)
$H_1 : \sigma_1^2 > \sigma_2^2$ (right-tailed test)
$H_1 : \sigma_1^2 < \sigma_2^2$ (left-tailed test)

Step 3: Table value: Based on level of significance α, critical values are given below

level of significance (α)	1%	5%	10%
Critical value for two-tailed test	2.58	1.96	1.645
Critical value for right-tailed test	2.33	1.645	1.28
Critical value for left-tailed test	-2.33	-1.645	-1.28

Step 4: Test statistic

$$Z_{cal} = \frac{s_1^2 - s_2^2}{\sqrt{\frac{2s_1^4}{n_1} + \frac{2s_2^4}{n_2}}}$$

where s_1^2 = first sample variance

s_2^2 = second sample variance.

n_1 = size of 1st sample.

n_2 = size of 2nd sample.

Step 5: Conclusion

If $Z_{cal} < Z_{tab}$ we accept H_0 otherwise we reject H_0.

Q: A comparitive study of variation in weights (in pounds) of army soldiers and navy sailors was made. The sample variance of weight of 120 soldiers was 60 pound and sample variance of weights of 160 sailors is 70 pound. Test whether the soldiers and sailors have equal variation in their weights using 5% level of significance.

Sol: Given,

Size of 1st sample $n_1 = 120$

Size of 2nd sample $n_2 = 160$

1st sample variance $s_1^2 = 60$ pound

2nd sample variance $s_2^2 = 70$ pound.

Level of significance $\alpha = 0.05$.

Step 1: Null Hypothesis $H_0: \sigma_1^2 = \sigma_2^2$

Step 2: Alternate hypothesis $H_1: \sigma_1^2 \neq \sigma_2^2$ (two-tailed test)

Step 3: Table value for $\alpha = 0.05$ $Z_{tab} = 1.96$.

Step 4: Test statistic, $Z_{cal} = \frac{s_1^2 - s_2^2}{\sqrt{\frac{2s_1^4}{n_1} + \frac{2s_2^4}{n_2}}}$

$$= \frac{|60 - 70|}{\sqrt{\frac{2(60)^2}{120} + \frac{2(70)^2}{160}}}$$

$$= |-0.91|$$

$$Z_{cal} = 0.91$$

Step 5: Conclusion $Z_{cal} = 0.91$

$Z_{tab} = 1.96$

$Z_{cal} < Z_{tab}$ $\therefore$ we accept H_0

$\therefore$ The soldiers and sailors have equal variation in their weights.

Q: Two sources of raw material are under consideration by a bulb manufacturing company. Both sources seem to have similar characteristics but the company is not sure about their respective uniformity. A sample of 52 lots from source A yields variance 25 and a sample of 40 lots from source B yields variance of 12. Test whether the variance of source A significantly differs to variances of source B at $\alpha = 0.05$%.

Sol: Given,

Size of 1st sample $n_1 = 52$

Size of 2nd sample $n_2 = 40$

1st sample variance $s_1^2 = 25$

2nd sample variance $s_2^2 = 12$

level of significance $\alpha = 0.05$

Step 1: Null Hypothesis, $H_0 : \sigma_1^2 = \sigma_2^2$

Step 2: Alternate hypothesis, $H_1 : \sigma_1^2 \neq \sigma_2^2$ (two-tailed test)

Step 3: Table value, for $\alpha = 0.05$, $Z_{tab} = 1.96$.

Step 4: Test statistic, $Z_{cal} = \dfrac{s_1^2 - s_2^2}{\sqrt{\frac{2s_1^4}{n_1} + \frac{2s_2^4}{n_2}}}$

$$= \frac{25 - 12}{\sqrt{\frac{2(25)^2}{52} + \frac{2(12)^2}{40}}} = 2.36$$

Step 5: conclusion $Z_{cal} = 2.36$

$Z_{tab} = 1.96$

$Z_{cal} > Z_{tab}$ $\therefore$ we reject H_0.

Hence, variance of source A differs from variance of source B.

→ Test of significance for single proportion:

Suppose a large random sample size n has a sample proportion P of members possessing a certain attribute (i.e, proportion of successes). To test the hypotheses that the proportion P in the proportion has a specified value P_0.

Step 1: Null hypothesis $H_0 : P = P_0$.

Step 2: Alternate hypothesis $H_1 : P \neq P_0$ (two-tailed test)

$H_1 : P > P_0$ (right-tailed test)

$H_1 : P < P_0$ (left-tailed test)

Step 3: Table value: Based on level of significance α, critical values is given below.

Level of significance (α)	1%	5%	10%
Critical value for two-tailed test	2.58	1.96	1.645
Critical value for right-tailed test	2.33	1.645	1.28
Critical value for left-tailed test	-2.33	-1.645	-1.28

Step 4 : Test Statistic, $Z_{cal} = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$

where p = sample proportion

P = Population proportion

$Q = 1 - P$.

n = sample size.

$p = \frac{x}{n}$

x = No. of observations satisfying a particular character

Step 5 : Conclusion. If $Z_{cal} < Z_{tab}$ we accept H_0 otherwise we reject H_0.

Q : A manufacturer claims that only 4% of his products are defective. A random sample of 500 were taken among which 100 were defective. Test the hypothesis at 0.05 level.

Sol : Given, sample size $n = 500$

No. of defective $x = 100$

p = sample proportion $= \frac{x}{n} = \frac{100}{500} = 0.2$.

P = Population proportion $= 4\% = 0.04$

$Q = 1 - P = 1 - 0.04 = 0.96$.

Step 1 : Null hypothesis : $H_0 : P = 0.04$

Step 2 : Alternate hypothesis : $H_1 : P > 0.04$ (right-tailed test)

Step 3 : Table value : for $\alpha = 0.05$, $Z_{tab} = 1.645$.

Step 4 : Test statistic : $Z_{cal} = \frac{|p - P|}{\sqrt{\frac{PQ}{n}}} = \frac{|0.2 - 0.04|}{\sqrt{\frac{(0.04)(0.96)}{500}}} = |18.26|$

$Z_{cal} = 18.26$

Step 5 : Conclusion : $Z_{cal} = 18.26$

$Z_{tab} = 1.645$.

$Z_{cal} > Z_{tab}$ $\therefore$ we reject H_0.

Hence more than 4% of his products are defective.

5: In a sample of 1000 people in AP 540 are rice eaters and rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level of significance?

Sol: Given,

Sample size, $n = 1000$.

Sample proportion of rice eaters, $p = \frac{x}{n} = \frac{540}{1000} = 0.54$

Population proportion of rice eaters $P = \frac{1}{2} = 0.5$

$Q = 1 - P = 0.5$.

Step 1: Null hypothesis $H_0 : P = 0.5$

Step 2: Alternate hypothesis $H_1 : P \neq 0.5$

Step 3: Table value: For $\alpha = 0.01$, $Z_{tab} = 2.58$.

Step 4: Test statistic, $Z_{cal} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.54 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{1000}}} = 2.532$.

Step 5: Conclusion, $Z_{cal} = 2.532$

$Z_{tab} = 2.58$

$Z_{cal} < Z_{tab}$ ∴ We accept H_0.

Hence, both rice and wheat eaters are equally popular in the state at 1% level of significance.

→ Test for equality of two proportions (or) Test for significant difference between two proportions:

Let P_1 and P_2 be the proportions in two large random samples of sizes n_1 and n_2 drawn from two populations P_1 and P_2. To test whether the two population proportions P_1, P_2 are equal.

Step 1: Null hypothesis $H_0 : P_1 = P_2$

Step 2: Alternate hypothesis $H_1 : P_1 \neq P_2$ (two-tailed test)

$H_1 : P_1 > P_2$ (right-tailed test)

$H_1 : P_1 < P_2$ (left-tailed test)

Step 3: Table value. Based on level of significance α. Critical values are given below

level of significance (α)	1%	5%	10%
Critical value for two-tailed test	2.58	1.96	1.645
Critical value for right-tailed test	2.33	1.645	1.28
Critical value for left-tailed test	-2.33	-1.645	-1.28

Step 4 : Test statistic, $Z_{cal} = \frac{P_1 - P_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

where P_1 = first sample proportion
P_2 = second sample proportion
n_1 = first sample size
n_2 = second sample size

$$p = \frac{P_1 n_1 + P_2 n_2}{n_1 + n_2}$$

$$q = 1 - p.$$

Step 5 : Conclusion,
If $Z_{cal} < Z_{tab}$ then we accept H_0, otherwise we reject H_0.

Q : Random samples of 400 men and 600 women, were asked whether they would like have a flyover near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal are same, at 5% level.

Sol : Given,
Size of first sample $n_1 = 400$
Size of and sample $n_2 = 600$
Sample proportion of men $P_1 = \frac{200}{400} = \frac{x_1}{n_1}$

$= \frac{1}{2} = 0.5$

Sample proportion of women $P_2 = \frac{x_2}{n_2} = \frac{325}{600}$

$= 0.541.$

Step 1 : Null hypothesis $H_0 : P_1 = P_2$.
Step 2 : Alternate hypothesis $H_1 : P_1 \neq P_2$ (two-tailed test)
Step 3 : Table value for $\alpha = 0.05$, $Z_{tab} = 1.96$.
Step 4 : Calculated value $Z_{cal} = \frac{|P_1 - P_2|}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$$p = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2} = \frac{(400 \times 0.5) + (600 \times 0.541)}{400 + 600} = 0.525$$

$$q = 1 - p = 1 - 0.525 = 0.475.$$

$$Z_{cal} = \frac{|0.5 - 0.541|}{\sqrt{0.525 \times 0.475\left(\frac{1}{400} + \frac{1}{600}\right)}} = |-1.24| = 1.24$$

Step 5 : Conclusion $Z_{cal} = 1.24$, $Z_{tab} = 1.96$, $Z_{cal} < Z_{tab}$
∴ we accept H_0. Hence proportions of men and women in favour of proposal are same.

Q: In a city - A, 20% of a random sample of 900 school boys had a certain slight physical defect. In another city - B, 18.5% of a random sample of 1600 school boys have the same defect. Is there difference between the proportions significant at 0.05 level of significance?

Sol: Given,

First sample size $n_1 = 900$

2nd sample size $n_2 = 1600$.

$P_1 = 20\% = 0.2 \qquad P_2 = 18.5\% = 0.185.$

$$P = \frac{P_1 n_1 + P_2 n_2}{n_1 + n_2} = \frac{(0.2 \times 900) + (0.185 \times 1600)}{900 + 1600} = 0.19$$

$$q = 1 - P = 1 - 0.19 = 0.81$$

Step 1: Null hypothesis $H_0 : P_1 = P_2$

Step 2: Alternate hypothesis $H_1 : P_1 \neq P_2$.

Step 3: Table value, for $\alpha = 0.05$, $Z_{tab} = 1.96$.

Step 4: Calculated value,

$$Z_{cal} = \frac{|P_1 - P_2|}{\sqrt{Pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{|0.2 - 0.185|}{\sqrt{0.19 \times 0.81 \times \left(\frac{1}{900} + \frac{1}{1600}\right)}}$$

$$= |-0.918|$$

$$Z_{cal} = 0.918$$

Step 5: Conclusion $Z_{cal} = 0.918$

$Z_{tab} = 1.96$

$Z_{cal} < Z_{tab}$ $\therefore$ We accept H_0.

Hence there is no significant difference between the two proportions.

UNIT - V : [TESTING OF HYPOTHESES]

SMALL SAMPLE TESTS

Degrees of freedom:

The number of independent values to evaluate the test statistic is called degrees of freedom. It is denoted by "ν", which is read as "Nu".

Ex: (1) Choose any 4 numbers.

Then degrees of freedom = 4

i.e, we can take 4 values freely.

(2) Choose 4 numbers such that the sum should be 20.

Then degrees of freedom = 3

i.e., we can take 3 values freely but the 4th value we should take based on constraint.

$$2 + 3 + 5 + \square = 20$$

$$\square = 10,$$

$\therefore \nu = 3.$

Student's t-test for population mean:

Let sample mean = $\bar{x}$

sample size = n

Sample SD = s

Population mean = μ

Population SD = σ

To test the population mean we follow the steps below.

Step 1: Null Hypothesis. $H_0 : \mu = \mu_0$

Step 2: Alternate Hypothesis. $H_1 : \mu \neq \mu_0$ (two-tailed test)

$H_1 : \mu < \mu_0$ (left-tailed test)

$H_1 : \mu > \mu_0$ (right-tailed test)

Step 3: Based on level of significance α,

$t_{tab} = t_{\alpha/2}$ for two tailed test

$t_{tab} = t_{\alpha}$ for left & right tailed test

Degrees of freedom, $\nu = n - 1$.

Step 4: Test statistic:

$$t_{cal} = \frac{\bar{x} - \mu}{(\sigma/\sqrt{n})} \quad (\sigma \text{ is knowns})$$

$$t_{cal} = \frac{\bar{x} - \mu}{(s/\sqrt{n-1})} \quad (\text{if } \sigma \text{ is unknowns})$$

Step 5: Conclusion

If $t_{cal} < t_{tab}$ then we accept H_0, otherwise we reject H_0.

Student's t-test for difference of means:

Student's t-test for equality of two population means:

Let $\bar{x}_1$ and $\bar{x}_2$ be means of two independent small samples of sizes n_1 and n_2 drawn from two normal populations having means μ_1 and μ_2.

To test whether the population means are equal, we follow the steps below.

Step 1: Null Hypothesis $H_0: \mu_1 = \mu_2$

Step 2: Alternate Hypothesis $H_1: \mu_1 \neq \mu_2$ (Two-tailed test)

$H_1: \mu_1 > \mu_2$ (right-tailed test)

$H_1: \mu_1 < \mu_2$ (left-tailed test)

Step 3: Based on level of significance α,

$t_{tab} = t_{\alpha/2}$ for two-tailed test

$t_{tab} = t_\alpha$ for left & right tailed tests.

Degrees of freedom, $\nu = n_1 + n_2 - 2$

Step 4: Test Statistic: $t_{cal} = \dfrac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where $S^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$ [If s_1^2 & s_2^2 are given]

$S^2 = \dfrac{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2}{n_1 + n_2 - 2}$ [If s_1^2 & s_2^2 are not given]

Step 5: Conclusion.

If $t_{cal} < t_{tab}$ then we accept H_0, otherwise we reject H_0.

Q: A sample of 26 bulbs gives a mean life of 990 hours with a SD of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours. Is the sample upto the standard?

Sol: Given,

Sample size, $n = 26 (< 30$, small sample)

Sample mean, $\bar{x} = 990$

Population mean, $\mu = 1000$

Sample S.D., $s = 20$

Step 1: Null Hypothesis: $H_0: \mu = 1000$

Step 2: Alternate Hypothesis $H_1: \mu < 1000$ (left-tailed test)

Step 3: Here $\alpha = 0.05$ (Assume)

Degrees of freedom $= \nu = n - 1 = 26 - 1 = 25$

For left-tailed test.

$t_{tab} = t_\alpha = t_{0.05}$ with 25 degrees of freedom

$t_{tab} = 1.708$.

Step 4: Test statistic

$$t_{cal} = \frac{|\bar{x} - \mu|}{\left(\frac{s}{\sqrt{n-1}}\right)} = \frac{|990 - 1000|}{(20/\sqrt{25})} = |-2.5|$$

$t_{cal} = 2.5$.

Step 5: Conclusion

$t_{cal} = 2.5$

$t_{tab} = 1.708$

$t_{cal} > t_{tab}$ $\therefore$ We reject H_0

$\therefore$ The sample is not upto the standard.

Q: A random sample of 10 boys had the following IQ's: 70, 120, 110, 101, 88, 83, 95, 98, 107 and 100. Do these data support the assumption of population mean IQ of 100?

Sol: Given, $n = 100$.

Sample observations are 70, 120, 110, 101, 88, 83, 95, 98, 107 and 100.

S.No.	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	70	-27.2	739.84
2	120	22.8	519.84
3	110	12.8	163.84
4	101	3.8	14.44
5	88	-9.2	84.64
6	83	-14.2	201.64
7	95	-2.2	4.84
8	98	0.8	0.64
9	107	9.8	96.04
10	100	2.8	7.84
	$\Sigma x_i = 972$		$\Sigma(x_i - \bar{x})^2 = 1833.6$

Sample mean,

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{972}{10} = 97.2$$

Sample variance,

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

$$= \frac{1}{9}(1833.6)$$

$s^2 = 203.73$.

$s = \sqrt{203.73}$

$= 14.27$.

Step 1: Null Hypothesis, $H_0: \mu = 100$

Step 2: Alternate Hypothesis, $H_1: \mu \neq 100$. (two-tailed test)

Step 3: For $\alpha = 0.05$ (say).

Degrees of freedom $\nu = n-1 = 10-1 = 9$.

For two tailed test $t_{tab} = t_{\alpha/2} = t_{0.025}$

$$t_{tab} = 2.26.$$

Step 4: Test statistic,

$$t_{cal} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

$$= \frac{|97.2 - 100|}{(4.27/\sqrt{9})}$$

$$= |-0.5886|$$

$$t_{cal} = 0.5886.$$

Step 5: Conclusion

$t_{cal} = 0.5886$

$t_{tab} = 2.26$

$t_{cal} < t_{tab} \Rightarrow$ we accept H_0

∴ The data support the assumption of population mean of 100.

Q: The height of 10 members of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches? Test at 5% significance level assuming that for 9 degrees of freedom. ($t = 1.833$ at $\alpha = 0.05$).

Sol: Given, sample size, $n = 10$.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
70	4	16
67	1	1
62	-4	16
68	2	4
61	-5	25
68	2	4
70	4	16
64	-2	4
64	-2	4
66	0	0
$\Sigma x_i = 660$		$\Sigma (x_i - \bar{x})^2 = 90$

Sample mean, $\bar{x} = \frac{\Sigma x_i}{n}$

$$= \frac{660}{10} = 66.$$

Sample variance, s^2

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

$$= \frac{90}{9} = 10.$$

$$s = \sqrt{10} = 3.16.$$

Step 1: Null hypothesis: $H_0: \mu = 64$

Step 2: Alternate hypothesis: $H_1: \mu > 64$ (right-tailed test)

Step 3: $\alpha = 0.05$, d.f, $\nu = n - 1 = 9$.

For right-tailed test, $t_{tab} = t_\alpha = t_{0.05}$

$t_{tab} = 1.833$.

Step 4: Test statistic, $t_{cal} = \dfrac{\bar{x} - \mu}{(S/\sqrt{n-1})}$

$= \dfrac{66 - 64}{(3.16/\sqrt{9})} = 1.9$.

Step 5: Conclusion $t_{cal} = 1.9$

$t_{tab} = 1.833$.

$t_{cal} > t_{tab} \Rightarrow$ we reject H_0.

∴ The average height is greater than 64 inches.

Q: Samples of two types of electric bulbs were tested for length of life and following data were obtained.

	Type-I	Type-II
Sample size	$n_1 = 8$	$n_2 = 7$
Sample mean	$\bar{x}_1 = 1234$ hours	$\bar{x}_2 = 1036$ hours
Sample SD	$S_1 = 36$ hours	$S_2 = 40$ hours.

Is the difference in the means sufficient to warrant that type-I is superior to type-II regarding length of life.

Sol: Given,

Sample sizes, $n_1 = 8$, $n_2 = 7$.

Sample means, $\bar{x}_1 = 1234$, $\bar{x}_2 = 1036$.

Sample S.D.'s, $S_1 = 36$, $S_2 = 40$

Step 1: Null hypothesis, $H_0: \mu_1 = \mu_2$

Step 2: Alternate hypothesis, $H_1: \mu_1 > \mu_2$ (right tailed test)

Step 3: let $\alpha = 0.05$.

d.f, $\nu = n_1 + n_2 - 2 = 8 + 7 - 2 = 13$.

For right-tailed test, $t_{tab} = t_\alpha = t_{0.05} = 1.771$

Step 4: Test statistic, $t_{cal} = \dfrac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where $S^2 = \dfrac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} = \dfrac{8 \times 36^2 + 7 \times 40^2}{8 + 7 - 2} = 1659.08$

$$t_{cal} = \frac{12.34 - 10.36}{\sqrt{1659.8\left(\frac{1}{8}+\frac{1}{7}\right)}} \approx 9.39.$$

Step 5: Conclusion. $t_{cal} = 9.36$

$t_{tab} = 1.771$

$t_{cal} > t_{tab} \Rightarrow$ we reject H_0.

∴ The difference in the means is sufficient to warrant that type I is superior to type-II regarding length of life.

Q: A group of 5 patients treated with medicine A weighs 42, 39, 48, 60 and 41 kgs. Second group of 7 patients from the same hospital treated with medicine B weighs 38, 42, 56, 64, 68, 69 and 62 kgs. Do you agree with the claim that medicine B decreases the weight significantly?

Sol: Given, $n_1 = 5$, $n_2 = 7$.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
42	-4	16	38	-19	361
39	-7	49	42	-15	225
48	2	4	56	-1	1
60	14	196	64	7	49
41	-5	25	68	11	121
			69	12	144
			62	5	25
$\Sigma x_i = 230$		$\Sigma(x_i - \bar{x})^2 = 290$	$\Sigma y_i = 399$		$\Sigma(y_i - \bar{y})^2 = 926$

$$\bar{x} = \frac{\Sigma x_i}{n_1} = \frac{230}{5} = 46$$

$$\bar{y} = \frac{\Sigma y_i}{n_2} = \frac{399}{7} = 57$$

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

$$= \frac{290 + 926}{5 + 7 - 2} = \frac{1216}{10}$$

$$s^2 = 121.6$$

$$s = 11.03.$$

Step 1: Null Hypothesis: $H_0 : \mu_1 = \mu_2$

Step 2: Alternate Hypothesis: $H_1 : \mu_1 > \mu_2$ (right-tailed test)

Step 3: For $\alpha = 0.05$,

Degrees of freedom, $\nu = n_1 + n_2 - 2 = 5 + 7 - 2 = 10$

For right tailed test $t_{tab} = t_\alpha = t_{0.05}$

$t_{tab} = 1.812.$

Step 4: Test statistic, $t_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$$= \frac{|46 - 57|}{11.03\sqrt{\frac{1}{5} + \frac{1}{7}}} = |-1.71|$$

$$t_{cal} = 1.7.$$

Step 5: Conclusion, $t_{cal} < t_{tab} \Rightarrow$ we accept H_0.

∴ Medicine B does not decrease the weight significantly.

Q: The means of two random samples of sizes 9 and 7 are 196.42 and 198.82 respectively. The sum of squares of the deviations from the mean are 26.94 and 18.73 respectively. Can the sample be considered to have been drawn from the same normal population?.

Sol: Given,

$n_1 = 9, \quad n_2 = 7$

$\bar{x} = 196.42, \quad \bar{y} = 198.82$

$\sum (x_i - \bar{x})^2 = 26.94, \quad \sum (y_i - \bar{y})^2 = 18.73.$

$$s^2 = \frac{1}{n_1+n_2-2}\left[\sum (x_i-\bar{x})^2 + \sum (y_i-\bar{y})^2\right]$$

$$= \frac{1}{9+7-2}\left[26.94 + 18.73\right] = 3.26$$

$s = \sqrt{3.26} = 1.81.$

Step 1: Null Hypothesis, $H_0 : \mu_1 = \mu_2$

Step 2: Alternate Hypothesis, $H_1 : \mu_1 \neq \mu_2$ (two-tailed test)

Step 3: For $\alpha = 0.05$

$df \quad \nu = n_1 + n_2 - 2 = 14$

For two-tailed test $t_{tab} = t_{\alpha/2} = t_{0.025}$

$t_{tab} = 2.15.$

Step 4: Test Statistic:

$$t_{cal} = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{|196.42 - 198.82|}{1.81\sqrt{\frac{1}{9}+\frac{1}{7}}} = |-2.63|$$

$t_{cal} = 2.63.$

Step 5: Conclusion:

$t_{cal} > t_{tab} \Rightarrow$ we reject H_0

∴ The sample cannot be considered to have been drawn from the same normal population.

Paired 't' test:

If (x_1, y_1) (x_2, y_2) ..... (x_n, y_n) be the pairs of sales data before and after the sales promotion in a business concern, then we apply paired t-test to examine the significance of the difference of the two situations.

Let $d_i = x_i - y_i$ (or) $y_i - x_i$ for $i = 1, 2, 3, \ldots, n$ then we get $d_1, d_2, \ldots, d_n$.

Step 1: Null Hypothesis: $H_0 : \mu_1 = \mu_2$, i.e., $\mu = 0$, μ be the mean of population of differences.

Step 2: Alternate hypothesis: $H_1 : \mu_1 \neq \mu_2$, $H_1 : \mu_1 > \mu_2$, $H_1 : \mu_1 < \mu_2$

Step 3: Based on level of significance α,

$t_{tab} = t_{\alpha/2}$ for two-tailed test

$t_{tab} = t_\alpha$ for left & right tailed test.

Degrees of freedom, $\nu = n - 1$.

Step 4: Test statistic, $t_{cal} = \dfrac{\bar{d}}{s/\sqrt{n}}$

where $\bar{d} = \dfrac{\sum d_i}{n}$

$$s^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}$$

n = sample size.

Step 5: Conclusion

If $t_{cal} < t_{tab}$ then we accept H_0, otherwise we reject H_0.

Q: Scores are obtained in shooting competition by 10 soldiers before and after intensive training are given below.

Before	67	24	57	55	63	54	56	68	33	43
After	70	38	58	58	56	67	68	75	42	38

Test whether intensive training is useful at 0.05 level of significance

Sol: Here we apply paired t-test.

$$\bar{d} = \frac{\sum d_i}{n} = \frac{-50}{10} = -5$$

$$s^2 = \frac{\sum (d_i - \bar{d})^2}{n-1} = \frac{482}{9} = 53.5$$

$$s = 7.32$$

x_i	y_i	$d_i = x_i - y_i$	$d_i - \bar{d} = d_i + 5$	$(d_i - \bar{d})^2$
67	70	-3	2	4
24	38	-14	-9	81
57	58	-1	4	16
55	58	-3	2	4
63	56	7	12	144
54	67	-13	-8	64
56	68	-12	-7	49
68	75	-7	-2	4
33	42	-9	-4	16
43	38	5	10	100
		$\Sigma d_i = -50$		$\Sigma (d_i - \bar{d})^2 = 482$

Step 1: Null Hypothesis $H_0: \mu_1 = \mu_2$ i.e., $\mu = 0$, there is no effect.

Step 2: Alternate Hypothesis $H_1: \mu_1 < \mu_2$ i.e., intensive training is useful.
(left-tailed test)

Step 3: For $\alpha = 0.05$,

df $\nu = n - 1 = 10 - 1 = 9$.

for left-tailed test, $t_{tab} = t_\alpha = t_{0.05} = 1.83$.

Step 4: Test statistic, $t_{cal} = \frac{\bar{d}}{s/\sqrt{n}} = \frac{|-5|}{7.32/\sqrt{10}} = |-2.16|$

$$t_{cal} = 2.16.$$

Step 5: Conclusion $t_{cal} > t_{tab} \Rightarrow$ we reject H_0.

∴ Intensive training is useful.

Q: The blood pressure of 5 women before and after intake of a certain drug are given below.

Before	110	120	125	132	125
After	120	118	125	136	121

Test whether there is significant change in blood pressure at 1% level of significance.

Sol: Here we apply paired t-test.

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{10}{5} = 2.$$

$$s^2 = \frac{\Sigma (d_i - \bar{d})^2}{n-1} = \frac{1}{4}(120$$

x_i	y_i	$d_i = x_i - y_i$	$d_i - \bar{d} = d_i + 1.6$	$(d_i - \bar{d})^2$
110	120	-10	8.4	70.56
120	118	2	3.6	12.96
125	125	0	1.6	2.56
132	136	-4	-2.4	5.76
125	121	4	5.6	31.36
		$\Sigma d_i = -8$		$\Sigma (d_i - \bar{d})^2 = 123.2$

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{-8}{5} = 1.6$$

$$s^2 = \frac{\Sigma (d_i - \bar{d})^2}{n-1} = \frac{123.2}{5-1} = 30.8$$

$$s = 5.55$$

Step 1: Null Hypothesis, $H_0 : \mu_1 = \mu_2$

Step 2: Alternate Hypothesis, $H_1 : \mu_1 \neq \mu_2$ (two-tailed test)

Step 3: For $\alpha = 0.05$, 5%.

d.f, $\nu = n - 1 = 5 - 1 = 4$

For two-tailed test, $t_{tab} = t_{\alpha/2} = t_{0.025} = 2.776$

Step 4: Test Statistic, $t_{cal} = \frac{\bar{d}}{s/\sqrt{n}} = \frac{|-1.6|}{5.55/\sqrt{5}} = |-0.64|$

$$t_{cal} = 0.64$$

Step 5: Conclusion, $t_{cal} < t_{tab} \Rightarrow$ we accept H_0.

∴ No significant change in blood pressure.

Step 3: For $\alpha = 0.01$ 1%.

d.f $\nu = n - 1 = 5 - 1 = 4$

For two-tailed test, $t_{tab} = t_{\alpha/2} = t_{0.005} = 4.604$

Step 4: Test Statistic, $t_{cal} = \frac{\bar{d}}{s/\sqrt{n}} = \frac{|-1.6|}{5.55/\sqrt{5}} = |-0.64|$

$$t_{cal} = 0.64$$

Step 5: Conclusion, $t_{cal} < t_{tab} \Rightarrow$ we accept H_0

∴ No significant change in blood pressure

F-test for equality of two population variances:

Let two independent random samples of sizes n_1 and n_2 be drawn from two normal populations.

To test the hypothesis that the two population variances σ_1^2 and σ_2^2 are equal (or) not, we follow the steps below.

Step 1: Null Hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$.

Step 2: Alternate Hypothesis: $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Step 3: Level of significance:

At α level $F_{tab} = F_\alpha$

with degrees of freedom $(\vartheta_1, \vartheta_2) = (n_1 - 1, n_2 - 1)$.

Step 4: Test statistic:

$$F_{cal} = \frac{s_1^2}{s_2^2} \quad \text{if } s_1^2 > s_2^2 \qquad \text{where } s_1^2 = \frac{\Sigma (x_i - \bar{x})^2}{n_1 - 1}$$

$$= \frac{s_2^2}{s^2} \quad \text{if } s_2^2 > s_1^2 \qquad s_2^2 = \frac{\Sigma (y_i - \bar{y})^2}{n_2 - 1}$$

Step 5: Conclusion

If $F_{cal} < F_{tab}$ then we accept H_0 otherwise we reject H_0.

Q: The time taken by workers in performing a job by method I and method II is given below:

Method I	20	16	26	27	23	22	—
Method II	27	33	42	35	32	34	38

Do the data show that the variances of the time distribution from population from which these samples are drawn do not differ significantly?

Sol: Given, $n_1 = 6$, $n_2 = 7$.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
20	-2.3	5.29	27	-7.4	54.76
16	-6.3	39.69	33	-1.4	1.96
26	3.7	13.69	42	7.6	57.76
27	4.7	22.09	35	0.6	0.36
23	0.7	0.49	32	-2.4	5.76
22	-0.3	0.09	34	-0.4	0.16
			38	3.6	12.96
Σ 134		81.34	241		133.72

$$\bar{x} = \frac{\Sigma x_i}{n_1} = \frac{134}{6} = 22.3$$

$$\bar{y} = \frac{\Sigma y_i}{n_2} = \frac{241}{7} = 34.4$$

$$s_1^2 = \frac{\Sigma (x_i - \bar{x})^2}{n_1 - 1} = \frac{81.34}{6 - 1} = 16.26.$$

$$s_2^2 = \frac{\Sigma (y_i - \bar{y})^2}{n_2 - 1} = \frac{133.72}{7 - 1} = 22.29.$$

Here $s_2^2 > s_1^2$.

Step1: Null Hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$.

Step2: Alternate Hypothesis: $H_1: \sigma_1^2 \neq \sigma_2^2$.

Step3: Level of significance;

$\alpha = 0.05 \quad (\vartheta_1, \vartheta_2) = (n_1 - 1, n_2 - 1) = (6-1, 7-1) = (5, 6)$

$F_{tab} = F_\alpha = F_{0.05}$

$F_{tab} = 4.39$.

Step4: Test statistic

$$F_{cal} = \frac{s_2^2}{s_1^2} \quad [\because s_2^2 > s_1^2]$$

$$F_{cal} = \frac{22.29}{16.26} = 1.37.$$

Step5: Conclusion:

$F_{cal} < F_{tab} \Rightarrow$ we accept H_0.

$\therefore$ The variances of time distributions from population from which those samples are drawn do not differ significantly.

Q: In one sample of 8 observations from a normal population, the sum of the squares of deviations of the sample values from the sample mean is 84.4 and in another sample of 10 observations it was 102.6. Test at 5% level whether the populations have the same variance.

Sol: Given,

$n_1 = 8, \quad n_2 = 10$

$\sum (x_i - \bar{x})^2 = 84.4, \quad \sum (y_i - \bar{y})^2 = 102.6.$

$$s_1^2 = \frac{\sum (x_i - \bar{x})^2}{n_1 - 1} = \frac{84.4}{8-1} = 12.057.$$

$$s_2^2 = \frac{\sum (y_i - \bar{y})^2}{n_2 - 1} = \frac{102.6}{10-1} = 11.4.$$

Here $s_1^2 > s_2^2$.

Step1; Null Hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$

Step2: Alternate Hypothesis: $H_1: \sigma_1^2 \neq \sigma_2^2$.

Step3; level of significance;

$\alpha = 0.05, \quad d.f, (\vartheta_1, \vartheta_2) = (n_1 - 1, n_2 - 1) = (7, 9).$

$F_{tab} = F_\alpha = F_{0.05} = 3.29.$

Step4: Test statistic, $F_{cal} = \frac{s_1^2}{s_2^2} \quad \because s_1^2 > s_2^2$.

$$F_{cal} = \frac{12.057}{11.4} = 1.057.$$

Step5: Conclusion, $F_{cal} < F_{tab} \Rightarrow$ we accept H_0.

$\therefore$ the populations have same variance.

χ^2 (Chi-square) test for goodness of fit:

We use this test to decide whether the discrepancy between theory and experiment is significant or not i.e., to test whether the difference between the theoretical (E_i) and observed values (O_i) can be attributed to chance (or) not

Step 1: Null Hypothesis: $H_0: O_i = E_i$

Step 2: Alternate Hypothesis: $H_1: O_i \neq E_i$.

Step 3: Level of Significance:

At α level $\chi^2_\alpha = \chi^2_{tab}$.

Degrees of freedom, $\nu = n-1$.

Step 4: Test Statistic:

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i = Observed frequency (or) Experimental Frequency

E_i = Expected frequency (or) Theoretical Frequency.

Step 5: Conclusion:

If $\chi^2_{cal} < \chi^2_{tab}$ then we accept H_0 otherwise we reject H_0.

Q: The following figures show the distribution of digits in numbers chosen at random from a telephone dictionary.

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	1026	1107	997	966	1075	933	1107	972	964	853

Test whether the digits may be taken to occur equally frequently in the directory.

Sol: Given $n = 10$

Digits	O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
0	1026	1000	676	0.676
1	1107	1000	11449	11.449
2	997	1000	9	0.009
3	966	1000	1156	1.156
4	1075	1000	5625	5.625
5	933	1000	4489	4.489
6	1107	1000	11449	11.449
7	972	1000	784	0.784
8	964	1000	1296	1.296
9	853	1000	21609	21.609
Total	10000			58.542

$$E_i = \frac{10,000}{10}$$

$$E_i = 1000$$

$$\sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

$$= 58.542$$

Step 1: Null Hypothesis: $H_0: O_i = E_i$

Step 2: Alternate Hypothesis: $H_1: O_i \neq E_i$.

Step 3: Level of significance:

$\alpha = 0.05$, d.f, $\vartheta = n-1 = 10-1 = 9$.

$\chi^2_{tab} = \chi^2_{\alpha} = \chi^2_{0.05}$

$\chi^2_{tab} = 16.919$.

Step 4: Test statistic, $\chi^2_{cal} = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] = 58.542$

Step 5: Conclusion, $\chi^2_{cal} > \chi^2_{tab} \Rightarrow$ We reject H_0.

∴ The observed values and expected values are not equal.

Q: The number of automobile accidents per week in a certain community are as follows: 12, 8, 20, 2, 14, 10, 15, 6, 9, 4. Are these frequencies in agreement with the belief that accident conditions were the same during this 10 week period.

Sol: Given, $n = 10$.

O_i	E_i	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
12	10	2	0.4
8	10	-2	0.4
20	10	10	10
2	10	-8	6.4
14	10	4	1.6
10	10	0	0
15	10	5	2.5
6	10	-4	1.6
9	10	-1	0.1
4	10	-6	3.6
Total 100			26.6

$E_i = \frac{100}{10} = 10$.

Step 1: Null Hypothesis: $H_0: O_i = E_i$

Step 2: Alternative Hypothesis: $H_1: O_i \neq E_i$.

Step 3: Level of significance:

$\alpha = 0.05$, d.f, $\vartheta = n-1 = 10-1 = 9$

$\chi^2_{tab} = \chi^2_{\alpha} = \chi^2_{0.05} = 16.9$.

Step 4: Test statistic:

$\chi^2_{cal} = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] = 26.6$.

Step 5: Conclusion:

$\chi^2_{cal} > \chi^2_{tab} \Rightarrow$ we reject H_0.

∴ Observed and Expected values are not equal.

Chi-Square Test for Independence of Attributes:

We have to test whether two attributes are independent or related (associated).

Let the observations be classified according to two attributes and the frequencies O_i in the different categories be shown in a two-way table called contingency table.

O_i			Total
Attribute I	a	b	a+b
Attribute II	c	d	c+d
Total	a+c	b+d	a+b+c+d

The expected frequencies are given by.

$$E_i^o = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}.$$

i.e.,

I	$E_a = \frac{(a+c)(a+b)}{a+b+c+d}$	$E_b = \frac{(a+b)(b+d)}{a+b+c+d}$
II	$E_c = \frac{(a+c)(c+d)}{a+b+c+d}$	$E_d = \frac{(b+d)(c+d)}{a+b+c+d}$

Step 1: Null hypothesis: The attributes are independent. i.e., there is no association.

Step 2: Alternate hypothesis: The attributes are associated or related with each other i.e., they are not independent.

Step 3: Level of significance.

At α level $\chi^2_{tab} = \chi^2_{\alpha}$

with degrees of freedom, $\nu = (\text{No. of rows} - 1) \times (\text{No. of columns} - 1)$.

Step 4: Test statistic,

$$\chi^2_{cal} = \sum_{i=1}^{n} \left[\frac{(O_i^o - E_i^o)^2}{E_i^o} \right]$$

Step 5: Conclusion.

If $\chi^2_{cal} < \chi^2_{tab}$ then we accept H_0 otherwise we reject H_0.

Q: On the basis of information given below about the treatment of 200 patients suffering from a disease, state whether the new treatment is comparatively superior to the conventional treatment.

	Favourable	Not favourable	Total
New	60	30	90
Conventional	40	70	110.

Sol: Given, O_i's are:

O_i	Fav.	Not Fav.	Total
New	60	30	90
Conventional	40	70	110
Total	100	100	200.

$$E_i = \frac{\text{Row total} \times \text{Column Total}}{\text{Grand Total}}.$$

~~E_i Fav. Not Fav.~~

~~New $\frac{90 \times 100}{200}$~~

~~Conventional~~

E_i	Favourable	Not. Favourable
New	$\frac{90 \times 100}{200} = 45$	$\frac{90 \times 100}{200} = 45$
Conventional	$\frac{100 \times 110}{200} = 55$	$\frac{100 \times 110}{200} = 55$

O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
60	45	225	5
30	45	225	5
40	55	225	4.09
70	55	225	4.09
			18.18
Total: 200	200		

$$= \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

$= 18.18.$

Step 1: Null Hypothesis: New and conventional treatment are independent.

Step 2: Alternate Hypothesis: New treatment is comparitively superior to the conventional treatment.

Step 3: Level of significance:

$$\alpha = 0.05$$

$$d.f, \nu = (\text{Rows} - 1)(\text{columns} - 1)$$

$$\nu = (2-1)(2-1) = 1.$$

$$\chi^2_{tab} = \chi^2_{\alpha} = \chi^2_{0.05} = 3.841.$$

Step 4: Test statistic:

$$\chi^2_{cal} = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

$$\chi^2_{cal} = 18.18.$$

Step 5: Conclusion:

$$\chi^2_{cal} > \chi^2_{tab} \Rightarrow \text{we reject } H_0.$$

∴ New treatment is comparatively superior to conventional treatment.

Chi-square test for testing variance of a normal distribution:

Suppose that a random sample $x_1, x_2, \ldots, x_n$ is drawn from a normal population with mean μ and variance σ^2. To test the hypothesis that the population variance σ^2 has a specified value σ_0^2.

Step 1: Null Hypothesis: $H_0: \sigma^2 = \sigma_0^2$

Step 2: Alternate Hypothesis: $H_1: \sigma^2 \neq \sigma_0^2$

$H_1: \sigma^2 > \sigma_0^2$

$H_1: \sigma^2 < \sigma_0^2$

Step 3: Level of significance:

At α level, $\chi^2_{tab} = \chi^2_\alpha$

with degrees of freedom, $\nu = n-1$.

Step 4: Test statistic:

$$\chi^2_{cal} = \frac{ns^2}{\sigma_0^2}.$$

where s^2 = sample variance.

Step 5: Conclusion:

If $\chi^2_{cal} < \chi^2_{tab}$ then we accept H_0 otherwise we reject H_0.

Q: A random sample of size 20 from a normal population gives a mean of 42 and a variance of 25. Test the hypothesis that the population standard deviation is 8 at 5% level of significance.

Sol: Given, $n = 20$, $\bar{x} = 42$, $s^2 = 25$, $s = 5$, $\alpha = 0.05$.

Step 1: Null Hypothesis: $H_0: \sigma = \sigma_0 = 8 \Rightarrow H_0: \sigma = 8$.

Step 2: Alternate Hypothesis: $H_1: \sigma \neq 8$.

Step 3: For $\alpha = 0.05$, d.f, $\nu = n-1 = 19$, $\chi^2_{tab} = \chi^2_\alpha = 30.14$.

Step 4: Test Statistic, $\chi^2_{cal} = \frac{ns^2}{\sigma_0^2} = \frac{20(25)}{(8)^2} = 7.8125$.

Step 5: Conclusion: $\chi^2_{cal} < \chi^2_{tab} \Rightarrow$ we accept H_0.

∴ The population standard deviation is 8 at 5% level.