INDEX

NAME: Arpan Sapkota STD.: MDS SEC.: I/II ROLL NO.: 07 SUB.: Statistical Methods for Data Science (MDS553)

S. No.	Date	Title	Page No.	Teacher's Sign / Remarks
		Statistical Methods for Data Science.	MDS553	
	1	Multinomial Distribution	4hr.	
	2.	Extreme Value Distribution.	4hrs.	
	3.	Generalized Power Series Distribution.	6Hrs	
	4.	Prior and Posterior Distribution	6hrs.	
	5.	Compand and Mixed Type Distribution.	6hr.	
	6.	Non-Parametric Tests	11 hrs.	
	7.	Testing of Hypotheses.	11 hrs.	
		Books:		
	1.	Inferential .... (Library)		
	2.	Fundamental of stat: SE Gupta, VK Kapur (Book)		

Unit 3

Generalized Power Series Distribution

Variable

1. Qualitative : Categorical, Response, Attributes.
2. Quantitative : Numerical.
   - 1. Discrete → Counting process.
   - 2. Continuous → Measuring process. → Interval.

Random Variable → Unbiased, fair, X = no. of passed student.
The variable which takes different values generated by random experiment is known as random variable. Random variable is denoted by x, y and z.
For Example:
If we toss a coin two times., then total no. of possible cases = $2^2 = 4$. (Sample Space)
$(\text{events})^{\text{no. of trails}}$ ⇒ Mutually Exclusive, collectively Exhaustive cases.

{HH, HT, TH, TT}	X	P(X)
Let X = no. of heads	0	1/4
	1	2/4
If implies X is a random variable.	2	1/4
		$\sum P(x) = 1$.

$f(x)$ → Probability density function (Pdf) $\int_{-\infty}^{\infty} f(x)\,dx = 1$, $0 \le f(x) \le 1$, C.
$P(x)$ → Probability Mass function. (Pmf) $\sum P(x) = 1$, $0 \le P(x) \le 1$, D

Mathematical Expectation (Avg)
Let X be a discrete random variable which takes the values $x_1, x_2, \ldots, x_n$ with respective probabilities $P(x_1), P(x_2), \ldots P(x_n)$ then the mathematical expectation of random variable X is given by,

$$E(X) = x_1 P(x_1) + \ldots + x_n P(x_n) = \sum x\, P(x) \quad \text{(Discrete)}$$

where,
P(x) = Probability Mass function.

Ily, $E(X) = \int_x x f(x)\,dx$, where $f(x)$ = probability Density function
(continuous)

{1 2 3 4 5 6}

x = values on the faces of the dice.

$E(X) = \bar{X} = \sum x P(x)$

$= \int x f(x)\,dx$

X	P(X)	xP(X)
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6

$E(X) = \sum xPx$

$= \frac{1}{6}[1+2+3+4+5+6]$

$= \frac{\sum x}{n} = \bar{X}$

Variance in terms of Mathematical Expectation

V(X) = Variance of X

$= E[X - E(X)]^2$

$\therefore V(X) = E(X^2) - [E(X)]^2$ D Where, $E(x^2) = \sum x^2 P(x)$

Ily, For Continuous.

$E(X^2) = \int_x x^2 f(x)\,dx$

Moments (Shape of distribution ? ⇒ Moments)

The various power of deviations taken from choosen value is known as moments. The choosen value may be any arbitary value (assumed mean).

If deviations are taken from mean then they are called central Moments If any arbitrary value then called raw moments.

Central Moments (μ_r).

The rth order central moment is denoted by μ_r and is given by,

$\mu_r = \frac{1}{n}\sum (x-\bar{x})^r$ [Individual series]

First order central Moment (μ_1) = $1/n \sum (x-\bar{x})^1 = 0$

Second order central Moment (μ_2) = $1/n \sum (x-\bar{x})^2$ = Variance

Raw Moments (μ_r')

The rth order raw moment about any arbitrary value (assumed mean) 'a' is given by,

$$\mu_r' = \frac{1}{n}\sum (x-a)^r$$

r^{th} raw moment about origin i.e. $a=0$

$$\mu_r' = \frac{1}{n}\sum (x-0)^r = \frac{1}{n}\sum x^r$$

$$\Rightarrow \mu_r' = E(x^r)$$

$r=1$, $\mu_1' = E(x)$ = Mean

$r=2$, $\mu_2' = E(x^2)$

* Note

μ_2 = Variance

$= E(x^2) - [E(x)]^2$

$= \mu_2' - (\mu_1')^2$

* Coeff. of Skewness $(\beta_1) = \frac{\mu_3^2}{\mu_2^3} \Rightarrow r_1 = \pm\sqrt{\beta_1}$

* Coeff. of kurtosis $(\beta_2) = \frac{\mu_4}{\mu_2^2}$

Moment Generating function.

The moment generating function of random variable X is given by,

$$M_x(t) = E(e^{tx}) = \sum_x e^{tx} p(x) \quad \text{(Discrete)}$$

$$= \int_x e^{tx} f(x)\,dx \quad \text{(continuous)}$$

* Note

$$M_x(t) = E(e^{tx}) \qquad \left[e^a = 1 + \frac{a}{1!} + \frac{a^2}{2!} + \dots\right]$$

$$= E\left[1 + \frac{tx}{1!} + \frac{t^2x^2}{2!} + \dots + \frac{t^r x^r}{r!} + \dots\right]$$

$$= 1 + \frac{t}{1!}E(x) + \frac{t^2}{2!}E(x^2) + \dots + \frac{t^r}{r!}E(x^r) + \dots$$

$$M_x(t) = 1 + \frac{t}{1!}\mu_1' + \frac{t^2}{2!}\mu_2' + \dots + \frac{t^r}{r!}\mu_r' + \dots$$

* Note:

$\mu_r' = \left.\frac{d^r}{dt^r} M_x(t)\right|_{t=0}$ and $\mu_2' = \left.\frac{d^2}{dt^2} M_x(t)\right|_{t=0}$

$\mu_1' = \left.\frac{d}{dt} M_x(t)\right|_{t=0}$

Properties

1. $M_{ax}(t) = M_x(at)$
2. If X and Y are two independent variables then the moment generating function of their sum $(x+y)$ is

$$M_{x+y}(t) = M_x(t)\, M_y(t)$$

$E(xy) = E(x)\,E(y)$ If x and y are independent

Characteristic function

$\Phi_x(t) = E[e^{itx}]$

$= \sum_x e^{itx} p(x)$ (Discrete)

$= \int_x e^{itx} f(x)\,dx$ (Continuous)

Covariance

Covariance between x and y.

$Cov(X,Y) = \frac{1}{n} \Sigma (x-\bar{x})(y-\bar{y})$

$= \frac{1}{n} \Sigma xy - \bar{x}\bar{y}$

$Cov(x,y) = E(xy) - E(x)\,E(y)$

* Note: If x and y are independ

$E(xy) = E(x)\,E(y)$

$\Rightarrow Cov(x,y) = 0$

Discrete Probability Distribution

1. Binomial
2. Poission
3. Negative Binomial

1. Binomial Distribution.

$P(X=x) = {}^{n}C_x p^x q^{n-x}$; $x = 0, 1, \ldots n$

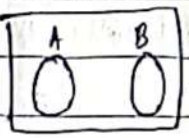
n = no. of trails (fixed)

p = prob. of success

$q = 1-p$ = prob. of failure

X = no. of successes.

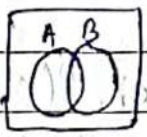
- ME Events ⇒ Disjoint Set



→ Coin Toss

→ Card: R & B

- NOT ME Events ⇒ Joint Set



→ Simultan. occur

→ cards: R & Queen

Generalized Power Series Distribution (GPSD)

A discrete random variable X is said to follow a general power series distribution (GPSD) if its probability mass function is given by,

$$P(X=x) = P(x) = \begin{cases} \frac{a_x \theta^x}{f(\theta)} & ; x = 0,1,2 \ldots ; a_x > 0 \\ 0 & ; \text{Otherwise} \end{cases}$$

where, $f(\theta) = \sum_{x \in S} a_x \theta^x$; $\theta > 0$ (power series)

S = sample.

So that $f(\theta)$ is positive, finite and differentiable and 'S' is a non-empty countable subset of non-negative integers.

Special Case of GPSD

1. Binomial Distribution.

In GPSD, take $\theta = \frac{p}{1-p}$, $f(\theta) = (1+\theta)^n$ and $S = 0, 1, 2, \ldots n$

then,

$$f(\theta) = (1+\theta)^n = \sum_{x=0}^{n} a_x \theta^x$$

Now, the binomial expansion of $(1+\theta)^n$ is $[a+b]^n = \sum_{x=0}^{n} {}^{n}C_x b^x a^{n-x}$

$$(1+\theta)^n = \sum_{x=0}^{n} {}^nC_x \theta^x (1)^{n-x}$$

$$= \sum_{x=0}^{n} {}^nC_x \theta^x$$

$$= \sum_{x=0}^{n} a_x \theta^x \qquad \therefore a_x = {}^nC_x$$

Since, $P(X=x) = \frac{a_x \theta^x}{f(\theta)} = \frac{{}^nC_x \left(\frac{p}{1-p}\right)^x}{\left(1+\frac{p}{1-p}\right)^n}$

$$= \frac{{}^nC_x \, p^x (1-p)^{-x}}{\left(\frac{1-p+p}{1-p}\right)^n} = {}^nC_x \, p^x (1-p)^{n-x}.$$

$\therefore P(X=x) = {}^nC_x p^x q^{n-x}$; $x = 0, 1, 2, \ldots n$

This is the probability mass function of Binomial distribution with parameters 'n' and 'p'.

Hence, the binomial distribution is the special case of GPSD.

Under the following condition Binomial distribution is applied.

1. No. of trails 'n' is fixed.
2. Events are mutually exclusive i.e. only two possible outcomes like head or tail, defective or non-defective, pass or fail, etc
3. Events are statistically independent.
4. Probability of succes 'p' remains constant in each trails.
5. $p+q=1$; q = probability of failure.

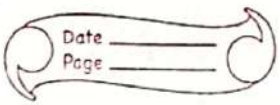
Definition:

If a discrete random variable X follow binomial distribution with parameter 'n' and 'p' then the probability mass function of the distribution is given by

$$P(X=x) = {}^nC_x \, p^x q^{n-x} \quad ; x = 0, 1, 2, \ldots n$$
$$p+q = 1$$

at least ($\geq$) : at least 2 $\Rightarrow$ $X \geq 2$

at most ($\leq$) : at most 2 $\Rightarrow$ $X \leq 2$



If gives the probability of getting 'x' success out of 'n' independent trails.
where,

p = prob. of success.

x = no. of successes

Note:

If $x \sim B(n,p)$ $\Rightarrow$ i.e. X follows Binomial distribution with parameters n & p.
Then,

Mean = np and

Variance of X = V(X) = npq.

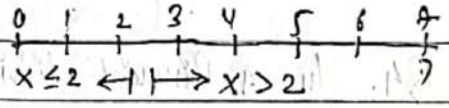
Problems on Binomial Distribution.

Q2. For a binomial distribution with n = 7 and p = 0.2, find.

ⓐ. P(X=5) ⓑ. P(X>2) ⓒ P(X≥3) ⓓ P(X ≥ 4)

n = 7, p = 0.2

q = 1 - p = 0.8



If $x \sim B(n,p)$, then $P(X=x) = {}^nC_x\, p^x q^{n-x}$; x = 0, 1, 2, ... n,

ⓑ. $P(X>2) = 1 - P(X \leq 2)$

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

$= 1 - [{}^7C_0 (0.2)^0 (0.8)^{7-0} + {}^7C_1 (0.2)^1 (0.8)^{7-1} + {}^7C_2 (0.2)^2 (0.8)^{7-2}]$

$= 1 - [0.2097 + 0.367 + 0.271]$

$= 0.1480$

ⓓ. $P(X \geq 4) = P(X=4) + P(X=5) + P(X=6) + P(X=7)$

$= {}^7C_4 (0.2)^4 (0.8)^{7-4} + {}^7C_5 (0.2)^5 (0.8)^{7-5} + {}^7C_6 (0.2)^6 (0.8)^{7-6} + {}^7C_7 (0.2)^7 (0.8)^{7-7}$

$= 0.0333.$

$${}^nC_r = \frac{n!}{(n-r)!\,r!}$$

Q5. At a particular university it has been found that 20% of the student withdraw without completing the Business statistics course. Assume that 18 student have registered for the course this semester.

(a). What is the probability that none will withdraw?

(b). What is the probability that at least one will withdraw?

(c) What is the probability that at most 2 will withdraw?

X = no. of withdraw

p = prob. of withdraw = 0.20 $\Rightarrow q = 0.80$

n = 18

Here, $X \sim B(n, p)$ so, $P(X=x) = {}^nC_x\, p^x q^{n-x}$; $x = 0, 1, 2, \ldots n$

(a). $P(X=0) = {}^{18}C_0\,(0.20)^0 (0.18)^{18-0}$

(b). $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X=0) = 1 - 0.018 = 0.982$

(c) $P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$

$= 0.0184 + {}^{18}C_1\,(0.2)^1 (0.80)^{18-1} + {}^{18}C_2\,(0.2)^2 (0.80)^{18-2}$

Q4. In a Binomial distribution with 6 independent trails the probabilities of 3 and 4 successes are found to be 0.2457 and 0.089. Find the parameter p of the distribution. (p = 0.326)

n = 6

p = 0.326

$q = 1 - p = 1 - 0.326$

$P(X=3) = 0.2457$, $P(X=x) = {}^nC_x\, p^x q^{n-x}$

$P(X=4) = 0.089$

Now,

$$\frac{P(X=3)}{P(X=4)} = \frac{0.2457}{0.089} = \frac{{}^6C_3\, p^3 q^{6-3}}{{}^6C_4\, p^4 q^{6-4}} = 2.76$$

or, $\frac{20q}{15q} = 2.76 \Rightarrow \frac{1-p}{p} = \frac{2.76 \times 3}{4}$

or, $\frac{1}{p} - 1 = 2.07 \Rightarrow \frac{1}{p} - 1 = 2.07 \Rightarrow q = \frac{1}{3.07}$

Q13 A fair coin is toned ten times. Find the probability of obtaining

(a) Exactly 4 heads.

(b) No heads — $n = 10$

(c) At least one head — $p = 0.5$ (∵ since coin is fair)

(d) At most three heads — X = no. of heads

(e) More than 8 head — $q = 1 - p = 0.5$

(f) 3 heads and 7 tails.

no. H	10	9	8	.....	0
no. T	0	1	2	....	10

Here,

$X \sim B(n, p)$ so, $P(X = x) = {}^{n}C_x\, p^x q^{n-x}$, $x = 0, 1, 2 \ldots n$

(a) $P(X = 4) = {}^{10}C_4 (0.5)^4 (0.5)^{10-4} = 0.2051$

(b) $P(X = 0) = {}^{10}C_0 (0.5)^0 (0.5)^{10-0} = 0.00097$

(c) $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - 0.00097$

(d) $P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$

(e) $P(X > 8) = P(X = 9) + P(X = 10)$

(f) Prob [8 heads and 7 tails] = $P(X = 3)$

Q15. Prob. of being liberal = 0.30

prob. of being conservative = 0.55

prob. of middle of the road = 0.15

If $X \sim B(n, p)$ then $P(X = x) = {}^{n}C_x\, p^x q^{n-x}$; $x = 0, 1, 2, \ldots n$

$n = 10$. $p + q = 1$

(a) X = no. of liberal

$p = 0.30 \Rightarrow q = 0.70$

$P(X = 4) = {}^{10}C_7 (0.30)^4 (0.70)^{10-4}$

(b) X = no. of conservative

$p = 0.55$

$P(X = 0) =$

(c) X = no. of middle of the road.

$p = 0.11$

$P(X = 2)$

(d) X = no of liberal & $p = 0.30$

$P(X \geq 80) = P(X=8) + P(X=9) + P(X=10)$

Q17 $n=6$

p = prob. of illegal $= 0.45$ $\quad P(X=x) = {}^nC_x\, p^x q^{n-x}$; $x=0,1,\ldots n$.

$q = 1-p = 0.55$ $\quad p+q=1$.

X = no. of illegal.

(i). $P(X=2) =$

(ii) $P(X=0) =$

(iii). p = prob. of legal $= 0.55$

$q = 1-p = 0.45$

X = no. of legal

Q21. mean $= 0.4$ and s.d $= 0.6$

If $X \sim B(n,p)$ then mean $= np$ and variance $= npq$

and $P(X=x) = {}^nC_x\, p^x q^{n-x}$; $x=0,1,2,\ldots n$.

$\therefore np = 0.4$ and $npq = 0.6^2$

$0.4(q) = 0.6^2$

$\Rightarrow q = \frac{0.6^2}{0.4} = 0.9$

$\therefore p = 1-q = 0.1$

Further, $np = 0.4 \Rightarrow n = \frac{0.4}{0.1} = 4$

Prob. of at least one success

$P(X \geq 1) = 1 - P(X<1) = 1 - P(X=0) = 1 - {}^4C_0\,(0.1)^0 (0.4)^{4-0} = 0.3439$

Q18. $P(X \geq 2) = ?$

Q24. $n=5$

p = prob. of correct answer $= 1/4 = 0.25$

$q = 1-p = 0.75$

X = no. of correct answers.

Here, $X \sim B(n,p)$, so $P(X=x) = {}^nC_x\, p^x q^{n-x}$; $x = 0,1,2 \ldots n$

(a). $P(X=5) = {}^5C_5 (0.25)^5 (0.75)^{5-5} = 0.25^5 = 0.00098.$

(b). $P(X \geq 4) = P(X=4) + P(X=5) = {}^5C_4 (0.25)^4 (0.75)^{5-4} + 0.00098$

$= 0.0156.$

(events)$^{\text{no. of trails}}$.

Qes: $7 \to (6,1), (1,6), (5,2), (2,5), (4,3), (3,4)$

$11 \to (6,5), (5,6)$

$12 \to (6,6)$ $\quad n = 5.$

p = prob. that a player gets audited

$= \frac{9}{36} = \frac{1}{4}$ Favourable no. of cases for sum 7 or 11 or 12 $= 9$

$q = 1 - p = 0.75$ Total cases $= 36$

$n = 5$

X = no. of times a player gets audited.

Here, X follows $B(n,p)$ so, $P(X=x) = {}^{n}C_x\, p^x q^{n-x}$; $x = 0,1,2,\ldots n$.

$P(X \geq 1) = 1 - P(X=0)$

$= 1 - {}^{5}C_0\, (0.25)^0 (0.75)^{5-0} = 0.7627.$

Mean, Variance and Moment generating function of Binomial Distribution with parameters 'n' and 'p'.

Mean

If $X \sim B(n,p)$ then $P(X=x) = {}^{n}C_x\, p^x q^{n-x}$; $x = 0,1,\ldots n$.

$p+q = 1$

$$E(X) = \sum_{x=0}^{n} x\, P(x)$$

$$= \sum_{x=0}^{n} x \left\{ {}^{n}C_x\, p^x q^{n-x} \right\}$$

$$= \sum_{x=0}^{n} x \cdot \frac{n!}{(n-x)!\,x!}\, p^x q^{n-x} \quad \because {}^{n}C_r = \frac{n!}{(n-r)!\,r!}$$

$$= \sum_{x=0}^{n} x \cdot \frac{n(n-1)!}{(n-x)!\,x(x-1)!}\, p^{x-1} p\, q^{n-x}$$

$$\left[(a+b)^n = \sum_{x=0}^{n} {}^{n}C_x\, p^x q^{n-x} \right]$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(n-x)!\,(x-1)!}\, p^{x-1} q^{n-x}$$

$$= np \sum_{x=1}^{n} {}^{n-1}C_{x-1}\, p^{x-1} q^{(n-x)-(x-1)}$$

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

$\therefore \boxed{E(X) = np}$ $\quad [p+q=1]$

Variance

$\text{Var}(X) = E[X^2] - [E(X)]^2$

$= E(X^2) - (np)^2 \quad \text{— (1)}$

Now, $E(X^2) = \sum_{x=0}^{n} x^2 P(x)$

$= \sum_{x=0}^{n} [x(x-1)+x]\,P(x)$

$= \sum_{x=0}^{n} x(x-1)P(x) + \sum_{x=0}^{n} x P(x)$

$= \sum_{x=0}^{n} x(x-1)\,{}^{n}C_x\, p^x q^{n-x} + E(x)$

$= \sum_{x=0}^{n} x(x-1)\,\frac{n!}{(n-x)!\,x!}\, p^x q^{n-x} + np$

$= \sum_{x=0}^{n} x(x-1)\,\frac{n(n-1)(n-2)!}{(n-x)!\,x(x-1)(x-2)!}\, p^2 p^{x-2} q^{n-x} + np$

$= n(n-1)p^2 \sum_{x=2}^{n} {}^{n-2}C_{x-2}\, p^{x-2} q^{n-x} + np$

$= n(n-1)p^2\,[q^{n-2} + \cdots + p^{n-2}] + np$

$= n(n-1)p^2\,[q+p]^{n-2} + np$

$= n(n-1)p^2 + np \qquad [\because p+q=1]$

Now, eqn (1) becomes,

$\text{Var}(X) = n(n-1)p^2 + np - n^2q^2 = n^2p^2 - np^2 + np - n^2p^2$

$= np(1-p)$

$\therefore \boxed{\text{Var}(X) = npq}$

Moment Generating function (mgf)

The mgf of binomial r.v. X is given by,

$$M_X(t) = E[e^{tx}]$$

$$= \sum_{x=0}^{n} e^{tx} p(x)$$

$$= \sum_{x=0}^{n} e^{tx} \, {}^nC_x \, p^x q^{n-x}$$

$$= \sum_{x=0}^{n} {}^nC_x (pe^t) q^{n-x}$$

$$\boxed{M_X(t) = [q + pe^t]^n} \qquad \because (q+p)^n = \sum_{x=0}^{n} {}^nC_x \, p^x q^{n-x}$$

Mean and Variance using mgf.

The r^{th} raw moment abt origin is given by

$$\mu_r' = \left. \frac{d^r M_X(t)}{dt^r} \right|_{t=0}$$

$\mu_1' = E(X-0) = E(X)$

$\mu_2' = E(X-0)^2 = E(X^2)$

$V(X) = E(X^2) - [E(X)]^2$

We know,

$$M_X(t) = (q + pe^t)^n$$

$$\frac{d}{dt} M_X(t) = n(q+pe^t)^{n-1} pe^t$$

Now, $\mu_1' = E(X) = \left. \frac{d}{dt} M_X(t) \right|_{t=0} = n(q+pe^0)^{n-1} pe^0$

$$= n(q+p)^{n-1} p.$$

$$= np \qquad [\because q+p = 1]$$

Further,

$$\frac{d^2}{dt^2} M_X(t) = np \left[e^t (n-1)(q+pe^t)^{n-2} pe^t + (q+pe^t)^{n-1} e^t \right]$$

$$= np \left[e^t \frac{d}{dt}(q+pe^t)^{n-1} + (q+pe^t)^{n-1} \frac{de^t}{dt} \right]$$

$$\because \frac{d}{dt}(q+pe^t)^{n-1} = \frac{d}{d(q+pe^t)}(q+pe^t)^{n-1} \cdot \frac{d(q+pe^t)}{dt}$$

$$= (n-1)(q+pe^t)^{n-1-1} \cdot (0+pe^t) = (n-1)(q+pe^t)^{n-2} pe^t.$$

$= np\left[e^t(n-1)(q+pe^t)^{n-2}pe^t + (q+pe^t)^{n-1}e^t\right]$

$\therefore \mu_2' = \left.\frac{d^2M_x(t)}{dt^2}\right|_{t=0} = np\left[e^0(n-1)(q+pe^0)^{n-2}pe^0 + (q+pe^0)^{n-1}e^0\right]$

$= np[(n-1)p+1]$

$= n^2p^2 - np^2 + np$

$\therefore$ Variance $= E(x^2) - [E(x)]^2$

$= n^2p^2 - np^2 + np - n^2p^2$

$= np(1-p)$

$\boxed{\mu_2 = npq}$

0 —— x —— n Binomial

0 —— x —— ∞ Poisson

Poisson Distribution

If a discrete random variable x follows poisson distribution with λ, then the probability mass function of the distribution is given by,

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

Note:

If $X \sim P(\lambda)$ then mean $= \lambda$ and variance $= \lambda$.

where λ = average.

Example:

X = No. of calls in a day

λ = average no. of calls in a day.

Poisson Approximation to Binomial distribution or relation between binomial and poisson distribution.

The binomial distribution reduces to poisson distribution under following conditions:

(i) If no. of trails 'n' is very large i.e. $n \to \infty$ $(n > 20)$

(ii) If probability of success p is very small i.e. $p \to 0$ $(p \leq 0.05)$

$\Rightarrow \lambda = np$

$$\therefore P(X=x) = \frac{e^{-np}(np)^x}{x!}$$

Poisson distribution from GPSD.

$$P(x) = \frac{a_x \theta^x}{f(\theta)}$$

In GPSD, take $\theta = \lambda$, $f(\theta) = e^{\theta}$ and $S = \{0,1,2,\ldots\infty\}$ then,

where, $f(\theta) = \sum_{x \in S} a_x \theta^x$

$$f(\theta) = e^{\theta} = \sum_{x=0}^{\infty} a_x \theta^x$$

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots$$

Now, $e^{\theta} = 1 + \frac{\theta}{1!} + \frac{\theta^2}{2!} + \ldots = \sum_{x=0}^{\infty} \frac{1}{x!}\theta^x$

Now

$$\sum_{x=0}^{\infty} \frac{\theta^x}{x!} = \sum_{x=0}^{\infty} a_x \theta^x$$

$\Rightarrow$ $a_x = \frac{1}{x!}$

The proof of GPSD is given by,

$$P(x) = \frac{a_x \theta^x}{f(\theta)} = \frac{\left(\frac{1}{x!}\right)\lambda^x}{e^{\lambda}}$$

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

This is the proof of poission dishibution with parameter λ.
Hence, poission dishibution is the special case of GPSD.

Problems:

Q 1. $\lambda = 4.2$

(a). $P(X \geq 5) = 1 - P(X < 5)$

$= 1 - [P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)]$

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

$$P(X \geq 5) = 1 - \left[\frac{e^{-4.2}4.2^0}{0!} + \frac{e^{-4.2}4.2^1}{1!} + \frac{e^{-4.2}4.2^2}{2!} + \frac{e^{-4.2}4.2^3}{3!} + \frac{e^{-4.2}4.2^4}{4!}\right]$$

$$= 1 - e^{-4.2}[1 + 4.2 + 8.82 + 12.348 + 12.965]$$

$$= 0.41$$

Q. 3: x = no. of price taken in every 3 years.

λ = average no. of price taken in every 3 years.
$= 4.$

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

(a). Prob. of no. price hikes in a randomly selected period of 8 years is:

$$P(X=0) = \frac{e^{-4}4^0}{0!} = 0.0183$$

(ii) $P(X=2) = \frac{e^{-4}4^2}{2!} = 0.1464$

(iii) $P(X=4) = \frac{e^{-4}4^4}{4!} = 0.1952$

(iv) $P(X \geq 5) = 1 - [P(X<5)]$

$= 1 - [P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)]$

$= 1 - [0.0183 + \frac{e^{-4}4!}{1!} + 0.1464 + \frac{e^{-4}4^3}{3!} + 0.1952]$

Q4. Given, 5-days.

Average no. of sales in week-days = 1.6

(a). X = no of sales in two days

λ = average no. of sales in two days.

$= \frac{1.6}{5} \times 2$

$= 0.64$

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

$P(X=4) = \frac{e^{-0.64}\,0.64^4}{4!} = 0.0037$

(b). X = no. of sales in a day

λ = average no. of sales in a day.

$= \frac{1.6}{5}$

$= 0.32$

$P(X \geq 3) = 1 - P(X<3)$

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

Q8. X = no. of arrival in 1 minute

λ = average no. of arrival in a minute = 6.

(a) $P(X>3) = 1 - P(X \leq 3)$

$= 1 - [P(X=0) + P(X=1) + P(X=2) + P(X=3)]$

(b) $P(X = 3 \text{ or } 4) = P(X=3) + P(X=4)$ $\quad P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

(c) $P(X=1)$

(d) $P(X \geq 2) = 1 - P(X < 2)$

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

(e), (f) X = no. of arrivals in two minutes.

λ = average no. of arrivals in two minutes.

$= 6 \times 2 = 12$

(e) $P(X=1) =$

(f) $P(X \geq 2) = 1 - P(X < 2)$

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

10. λ = average no. of vehicle per cycle (45 seconds)

$= \frac{10}{60} \times 45$ $\qquad 60 - 10$

$= 7.5$ $\qquad 1 - \frac{10}{60}$

X = no. of vehicles per cycle. $\qquad 45 - \frac{10}{60} \times 45$

20. n = no. of dispute = 80

p = prob. of non-marital disputes. $\qquad p \leq 0.05$

$= 1 - 0.96$ $\qquad n > 20$

$p = 0.04$

Now, using poission approximation to binomial we get,

$\lambda = np = 80\ (0.04) = 3.2$

$\therefore P(X = x) = \dfrac{e^{-np}(np)^x}{x!}$

(a) $P(X=7) = \dfrac{e^{-3.2}\ 3.2^7}{7!} = 0.027$

(b) $P(X=0) = \dfrac{e^{-3.2}\ 3.2^0}{0!} = e^{-3.2} = 0.0407$

(c) $P(X \geq 2) = 1 - P(X < 2)$

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

$= 1 - \left[0.0407 + \dfrac{e^{-3.2}\ 3.2^1}{1!}\right]$

25. p = prob. of defective $= \dfrac{1}{400} = 0.0025$

$n = 100$

X = no. of defectives.

Using poisson approximation to binomial.

We get,

$\lambda = np = 100\ (0.0025) = 0.25$

(a) $P(X=0)$

(b) $P(X < 2)$

(c) $P(X \geq 1) = 1 - P(X=0)$

(d) $P(X \geq 3)$

Mean, Variance and Moment Generating function of Poisson Distribution with parameter λ.

If $X \sim P(\lambda)$ then the probability of the distribution is,

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0,1,2,\cdots \quad \lambda > 0 \qquad \left[e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots\right]$$

Mean:

$$E(X) = \sum x P(x) = \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{x \cdot \lambda\lambda^{x-1}}{x(x-1)!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

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

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

$$\therefore \boxed{E(X) = \lambda}$$

Variance:

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = E(X^2) - (\lambda)^2 \quad \leftarrow (1)$$

where, $E(X^2) = \sum_{x=0}^{\infty} x^2 P(x) = \sum_{x=0}^{\infty} [x(x-1) + x] P(x)$

$$= \sum_{x=0}^{\infty} x(x-1)P(x) + \sum_{x=0}^{\infty} x P(x) = \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} + E(X)$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{x(x-1)\lambda^2\lambda^{x-2}}{x(x-1)(x-2)!} + \lambda = \lambda^2 e^{-\lambda}\left[\sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!}\right] + \lambda$$

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

$$= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda$$

$$\therefore E(X^2) = \lambda^2 + \lambda$$

Now, (1) becomes,

$$\text{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

$$\boxed{\text{Var}(X) = \lambda}$$

Moment Generating function (mgf)

The mgf of discrete random variable X is

$$M_X(t) = E(e^{tx}) = \sum_{x=0}^{\infty} e^{tx} p(x) = \sum_{x=0}^{\infty} e^{tx} \left\{ \frac{e^{-\lambda} \lambda^x}{x!} \right\}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{-\lambda} \left[1 + \frac{\lambda e^t}{1!} + \frac{(\lambda e^t)^2}{2!} + \dots \right]$$

$$= e^{-\lambda} e^{\lambda e^t} = e^{\lambda e^t - \lambda} = e^{\lambda(e^t - 1)}$$

$$\therefore \boxed{M_X(t) = e^{\lambda(e^t - 1)}}$$

Mean and Variance using mgf

The rth order raw moment about origin is,

$$\mu_r' = \left. \frac{d^r M_X(t)}{dt^r} \right|_{t=0}$$

$$\mu_1' = E(X) = \left. \frac{d M_X(t)}{dt} \right|_{t=0}$$

Now,

$$\frac{d M_X(t)}{dt} = \frac{d}{dt} e^{\lambda(e^t - 1)} = \frac{d e^{\lambda(e^t-1)}}{d(\lambda(e^t-1))} \cdot \frac{d}{dt} \{\lambda(e^t - 1)\}$$

$$= e^{\lambda(e^t-1)} \lambda e^t$$

$$\therefore E(X) = e^{\lambda(e^0 - 1)} \lambda e^0 = \lambda$$

$$\mu_2' = E(X^2) = \left. \frac{d^2 M_X(t)}{dt^2} \right|_{t=0}$$

Now,

$$\frac{d^2 M_X(t)}{dt^2} = \frac{d}{dt} \{ e^{\lambda(e^t-1)} \lambda e^t \} = \lambda \left[e^t \{ e^{\lambda(e^t-1)} \lambda e^t \} + e^t e^{\lambda(e^t-1)} \right]$$

$$E(X^2) = \lambda \left[e^0 e^{\lambda(e^0-1)} \lambda e^0 + e^0 e^{\lambda(e^0-1)} \right]$$

$$= \lambda[\lambda + 1] = \lambda^2 + \lambda$$

$$\therefore \text{Variance} = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2$$

$$\boxed{\text{Var}(X) = \lambda}$$

$\therefore$ If $X \sim P(\lambda)$ then mean of X = Variance of $X = \lambda$

Negative Binomial Distribution (NBD)

p = prob. of success in each trail
$q = 1-p$ = prob. of failure in ...
k = fixed no. of success
x = no. of failures (r.v.)
$x+k$ = total no. of trails.

NA ME
Dep. Indep.
$P(A \cap B)$ $P(A \cap B) = P(A)P(B)$

If a trail is repeated independently till fixed no. of success occurs with prob. 'p' which is constant for each trail, then such experiment follows negative Binomial Distribution.

Derivation of NBD

Consider a binomial experiment consists of 'n' trails and the prob. of success 'p' is constant for each trail. Let there be 'x' failures preceding the k^{th} success in $(x+k)$ trails. Here, $n = x+k$ trails are required to produce 'k' success (which is fixed) and x is a random variable.

Now, we need to find the probability that k^{th} success occurs in $(x+k)$ trails, which is the probability that exactly 'x' failures preceding the k^{th} success in $x+k$ trails.

Here, the last trails must be success with prob. 'p' and in the remaining $x+k-1$ trails, there are $(k-1)$ success and where prob. is given by,

$$^{x+k-1}C_{k-1}\, p^{k-1} q^x ; \quad p+q=1$$

Now,

By Multiplication Theorem of Probability, the prob. of x failures preceding the k^{th} success in $x+k$ trails is given by,

$$(^{x+k-1}C_{k-1}\, p^{k-1} q^x) \cdot p$$
$$= {}^{x+k-1}C_{k-1}\, p^k q^x$$

This is the prob. (prob. mass function) of NBD with parameter 'p' & 'k'

Note: If $X \sim NB(k,p)$ then,

* $P(X=x) = {}^{x+k-1}C_{k-1}\, p^k q^x$; $x = 0,1,2,\ldots$, $k \geq 1$

$= {}^{x+k-1}C_x\, p^k q^x$

* ${}^{x+k-1}C_{k-1} = {}^{x+k-1}C_x$

* $(1-a)^{-k} \sum_{x=0}^{\infty} {}^{x+k-1}C_x\, a^x$

Negative Binomial Distribution from GPSD.

In GPSD, take, $\theta = \frac{q/p}{1+q/p}$ and $f(\theta) = (1-\theta)^{-k}$

and, $S = \{0,1,2,\ldots\infty\}$

Now, $f(\theta) = \sum_{x=0}^{\infty} a_x \theta^x = (1-\theta)^{-k}$ — (i)

Now, $(1-\theta)^{-k} = \sum_{x=0}^{\infty} {}^{x+k-1}C_x\, \theta^x$ — (ii)

From (i) & (ii), we get.

$a_x = {}^{x+k-1}C_x$

From GPSD,

$$P(x) = \frac{a_x \theta^x}{f(\theta)} = \frac{{}^{x+k-1}C_x \left(\frac{q/p}{1+q/p}\right)^x}{\left(1-\frac{q/p}{1+q/p}\right)^{-k}}$$

$$\boxed{P(x) = {}^{x+k-1}C_x\, q^x p^k} \; ; \; x = 0,1,2,\ldots, \; k \geq 1$$

Mean, Variance and Moment Generating f NBD.

If a d.r.v $X \sim NB(k,p)$ then the prob. f the distribution is given by.

$P(X=x) = {}^{x+k-1}C_{k-1}\, p^k q^x$; $x = 0,1,2,\ldots$, $k > 0$

$= {}^{x+k-1}C_x\, p^k q^x \Big| (1-a)^{-n} = \sum_{n=0}^{\infty} {}^{x+n-1}C_x\, a^x$

Mean:

$$E(X) = \sum_{x=0}^{\infty} x\, P(X=x) = \sum_{x=0}^{\infty} x\, {}^{x+k-1}C_x\, p^k q^x$$

$$= p^k \sum_{x=0}^{\infty} \frac{(x+k-1)!}{(x+k-1-x)!\, x!}\, q^x$$

$$= p^k \sum_{x=0}^{\infty} \frac{x\,(x+k-1)!}{(k-1)!\, x(x-1)!}\, q^{x-1} \cdot q$$

$$= qp^k k \sum_{x=1}^{\infty} \frac{(x+k-1)!}{k!\,(x-1)!}\, q^{x-1}$$

$$= qp^k k \sum_{x=1}^{\infty} {}^{x+k-1}C_{x-1}\, q^{x-1}$$

$$= qp^k k \sum_{x=1}^{\infty} {}^{(x-1)+(k+1)-1}C_{x-1}\, q^{x-1}$$

$$= kqp^k (1-q)^{-(k+1)} \quad \because \text{↗}$$

$$= kqp^k p^{-(k+1)} \quad \because p+q=1$$

$$\therefore \boxed{E(X) = \frac{kq}{p}}$$

Variance:

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = E(X^2) - \left(\frac{kq}{p}\right)^2 \quad \text{— (1)}$$

Now, $E(X^2) = \sum_{x=0}^{\infty} x^2 P(x) = \sum_{x=0}^{\infty} [x(x-1)+x]\, P(x)$

$$= \sum_{x=0}^{\infty} x(x-1)\, P(x) + \sum_{x=0}^{\infty} x\, P(x)$$

$$= \sum_{x=0}^{\infty} x(x-1)\, {}^{x+k-1}C_x\, p^k q^x + E(X)$$

$$= p^k \sum_{x=0}^{\infty} \frac{x(x-1)(x+k-1)!}{(x+k-1-x)!\, x!}\, q^x + E(X)$$

$$= p^k q^2 \sum_{x=1}^{\infty} \frac{x(x-1)(x+k-1)!}{(k-1)!\, x(x-1)(x-2)!}\, q^{x-2} + E(X)$$

$$= p^k q^2 k(k+1) \sum_{x=2}^{\infty} \frac{(x+k-1)!}{(k+1)k(k-1)!(x-2)!} q^{x-2} + E(x)$$

$$= p^k q^2 k(k+1) \sum_{x=2}^{\infty} \frac{(x+k-1)!}{(k+1)!(x-2)!} q^{x-2} + E(x)$$

$$= p^k q^2 k(k+1) \sum_{x=2}^{\infty} {}^{x+k-1}C_{x-2}\, q^{x-2} + E(x).$$

$${}^{x+k-1}C_{x-2} = \frac{(x+k-1)!}{(x+k-1-x+2)!\,(x-2)!} = \frac{(x+k-1)!}{(k+1)!\,(x-2)!}$$

$$= p^k q^2 k(k+1) \sum_{x=2}^{\infty} {}^{(x-2)+(k+2)-1}C_{x-2}\, q^{x-2} + E(x)$$

$$= p^k q^2 k(k+1) (1-q)^{-(k+2)} + \frac{kq}{p}$$

$$\because (1-q)^{-n} = \sum_{x=0}^{\infty} {}^{x+n-1}C_x\, q^x$$

$$= p^k q^2 k(k+1) p^{-(k+2)} + \frac{kq}{p}$$

$$= \frac{q^2 k(k+1)}{p^2} + \frac{q^2 k}{p^2} + \frac{kq}{p} = \frac{q^2 k^2}{p^2} + \frac{q^2 k}{p^2} + \frac{kq}{p}$$

Now, Eq ① becomes,

$$\text{Var}(x) = \frac{q^2k^2}{p^2} + \frac{q^2k}{p^2} + \frac{kq}{p} - \frac{k^2q^2}{p^2}$$

$$= \frac{kq}{p}\left(\frac{q}{p} + 1\right)$$

$$= \frac{kq}{p}\left(\frac{q+p}{p}\right)$$

$$\boxed{\text{Var}(X) = \frac{kq}{p^2}}$$

Moment Generating function

Now, the mgf of X is

$$M_x(t) = E(e^{tx}) = \sum_{x=0}^{\infty} e^{tx} p(x) = \sum_{x=0}^{\infty} e^{tx}\, {}^{x+k-1}C_x\, p^k q^x$$

$$= p^k \sum_{x=0}^{\infty} {}^{x+k-1}C_x\, (qe^t)^x = p^k (1-qe^t)^{-k}$$

$$\boxed{M_x(t) = \frac{p^k}{(1-qe^t)^k}}$$

Mean and Variance using MGF

If $X \sim NB(k, p)$ then,

$$M_X(t) = \frac{p^k}{(1-qe^t)^k}$$

The rth raw moment about origin,

$$\mu_1' = \left.\frac{d^r M_X(t)}{dt^r}\right|_{t=0}$$

Now, Mean:

$$\mu_1' = E(X) = \left.\frac{dM_X(t)}{dt}\right|_{t=0}$$

Now $\frac{dM_X(t)}{dt} = \frac{d}{dt}\{p^k(1-qe^t)^{-k}\} = p^k \frac{d}{d(1-qe^t)}(1-qe^t)^{-k} \cdot \frac{d(1-qe^t)}{dt}$

$= p^k(-k)(1-qe^t)^{-k-1}(-q)e^t$

$= p^k kqe^t(1-qe^t)^{-k-1}$

$\therefore \mu_1' = p^k kqe^0(1-qe^0)^{-k-1} = p^k kq(1-q)^{-k-1} = p^k kq\, p^{-k-1}$

$$\boxed{E(X) = \frac{kq}{p}}$$

$$\mu_2' = \left.\frac{d^2 M_X(t)}{dt^2}\right|_{t=0}$$

$$\frac{d^2 M_X(t)}{dt^2} = p^k kq \frac{d}{dt}\{e^t(1-qe^t)^{-k-1}\}$$

$= p^k kq\left[(1-qe^t)^{-k-1}e^t + e^t(-k-1)(1-qe^t)^{-k-1-1}(-qe^t)\right]$

$\therefore \mu_2' = p^k kq\left[(1-qe^0)^{-k-1}e^0 + e^0(k+1)(1-qe^0)^{-k-2}qe^0\right]$

$= p^k kq\left[(1-q)^{-k-1} + (k+1)(1-q)^{-k-2}q\right]$

$= p^k kq\left[p^{-k-1} + (k+1)p^{-k-2}q\right]$

$= p^k kq\, p^{-k-1}\left[1 + (k+1)p^{-1}q\right]$

$= \frac{kq}{p}\left[1 + \frac{(k+1)q}{p}\right]$

$$= \frac{kq}{p} + \frac{k^2q^2}{p^2} + \frac{kq^2}{p^2}$$

Variance:

$$Var(x) = E(x^2) - [E(x)]^2$$

$$= \frac{kq}{p} + \frac{k^2p^2}{p^2} + \frac{kq^2}{p^2} - \frac{k^2q^2}{p^2}$$

$$= \frac{kq}{p}(1 + q/p) = \frac{kq}{p^2}(p+q)$$

$$\therefore \boxed{Var(X) = \frac{kq}{p^2}}$$

Q. If the prob. that an applicant will pass the driving license test on any given trail is 0.75. Find the prob. that an applicant will finally pass the test in the fourth try.

$x+k=4$

$P(X=3) = ?$

$x = 3$

$k = 1$

$p = 0.75$

$q = 0.25$

$^{x+k-1}C_x\, p^x q^x.$

p = prob. of an applicant will pass the test = 0.75

$q = 1-p = 0.25$

Here, X = no of failures.

Here $X \sim NB(k, p)$ so, $P(X=x) = {}^{x+k-1}C_x\, p^k q^x$; $x = 0, 1, \ldots$; $k > 0$

We have, $x+k=4$; $k=1$

$\therefore x = 3$

Now, the prob. that applicant will pass the test on fourth try is,

$P(X=3) = {}^{3+1-1}C_3 (0.75)^1 (0.25)^3 = 0.01$

Q. The probability of hitting the target at any trail is 0.2. If a shooter aims at a target. Find the prob. that the fifth fire is second hit.

$p = 0.2$, $q = 0.8$

$x+k=5$, $k=2$

Q3. A boy is throwing stones at a target. If the prob. of hitting the target at any point is 1/2, what is the prob. that his 10th throw is 5th hit?

Soln,

p = prob. of hitting a target at any trail = 0.5

q = 0.5

k = 5

x+k = 10

$P(X=7) = {}^{x+k-1}C_x\, p^k q^k$

Now, the prob. that his 10th throw is 5th hit is given by,

$P(X=5) = {}^{5+5-1}C_5\, (0.5)^5\, (0.5)^5 = 0.123$

The prob. that a striker will make a goal in a football match is 0.2. What is the prob that he will do fourth goal in fifth shot (0.0051)

Unit 1

Multinomial Distribution

An experiment is said to be multinomial experiment, if:

(i). The experiment consists of n fixed trails.
(ii). For each trail there are k (≥2 i.e. 3 or more) possible outcomes.
(iii) The trails are independent.
(iv) The prob for each outcome remains the same from trail to trail.

The examples of the multinomial experiments are,

(i). No. of throw of a fair dice in which each throw can result six different outcomes.
(ii). A number of selection or drawing's of balls at random with replacement from a box containing 30 balls of which 10 are white, 15 are black and 5 are red.

Definition:

The random variable $X = X_1, X_2, \ldots X_k$ denoting the outcome of 'n' trails, where X_i = freq^n or number of outcome E_i with respective probability P_i $(i = 1, 2, \ldots k)$, is said to have multinomial distribution with parameters $(n, P_1, P_2, \ldots P_k)$ if its prob. is given by,

$$P(X_1 = x_1, X_2 = x_2, \ldots X_k = x_k) = n! \frac{P_1^{x_1} P_2^{x_2} \ldots P_k^{x_k}}{x_1! x_2! \ldots x_k!}$$

where, $\sum_{i=1}^{k} x_i = n$ and $\sum_{i=1}^{k} P_i = 1$.

Q. An urn contains 8 red balls, 3 yellow ball's and 9 white balls. 6 balls are drawn at random with replacement. what is the prob. that 2 are red, 1 is yellow and 3 are white?

Soln:

Since, 6 balls are randomly selected with replacement so, there are 6 independent trails of a multinomial experiment.

P_1 = Prob. of getting a red ball in any trail = 8/20
P_2 = Prob. of getting a yellow balls in any trail = 3/20
P_3 = Prob of getting a white ball in any trail = 9/20.

The prob. of multinomial distn with parameters $(n, p_1, p_2, \ldots p_k)$ is

$$P(X_1 = x_1, X_2 = x_2, \ldots X_k = x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

The prob. of getting 2R, 1Y and 3W is given by,

$$P(X_1 = 2, X_2 = 1, X_3 = 3) = \frac{6!}{2!\,1!\,3!} \left(\frac{8}{20}\right)^2 \left(\frac{3}{20}\right)^1 \left(\frac{9}{20}\right)^3$$

$$= 0.131$$

Q. The painted light balls produced by a company are 50% red, 30% blue and 20% green. In a sample of 5 balls, find the prob that 2 red, 1 is green and 2 are blue.

Soln,

p_1 = Red = $(1/2)$

p_2 = Blue = $(3/5)$

p_3 = Green = $(2/5)$

$$P(X_1 = 2, X_2 = 1, X_3 = 2) = \frac{5!}{2!\,1!\,2!} (0.5)^2 (0.2)^1 (0.3)^2$$

$$= 0.135.$$

Reduction of Multinomial distributions into binomial distribution.

When $k = 2$, then the multinomial distribution has prob. mass function.

$$P(X_1 = x_1, X_2 = x_2) = \frac{n!}{x_1! x_2!} p_1^{x_1} p_2^{x_2}, \quad x_1 + x_2 = n$$

$$p_1 + p_2 = 1$$

$$x_2 = n - x_1 \quad \& \quad p_2 = 1 - p_1$$

$$P(x_1, n - x_1) = \frac{n!}{x_1!\,(n - x_1)!} p_1^{x_1} (1 - p_1)^{n - x_1}$$

If $p_1 = p$ and $x_1 = x$ $\therefore P(x) = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$.

$$\Rightarrow \boxed{P(x) = {}^nC_x\, p^x q^{n-x}}$$

Moment Generating function of Multinomial distribution.

If $X = (x_1, x_2, \ldots x_k) \sim MD(n, p_1, p_2, \ldots p_k)$

Then,

$$P(X) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

Now, the moment generating function of X is given by,

$$M_X(t) = M_{x_1 x_2 \ldots x_k}(t_1 t_2 \ldots t_k)$$

$$= E\left[e^{t_1 x_1 + t_2 x_2 + \ldots + t_k x_k}\right]$$

$$= \sum_x \left\{ e^{t_1 x_1 + t_2 x_2 + \ldots + t_k x_k} \right\} \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

$$= \sum_x \frac{n!}{x_1! x_2! \ldots x_k!} (p_1 e^{t_1})^{x_1} (p_2 e^{t_2})^{x_2} \ldots (p_k e^{t_k})^{x_k}.$$

$$\therefore M_X(t) = (p_1 e^{t_1} + p_2 e^{t_2} + \ldots + p_k e^{t_k})^n$$

This is the moment generating function of multinomial distribution.

If $t_1 \neq 0$ and $t_i = 0$ for all $i = 2, 3, \ldots k$

$$M_{X_1}(t_1) = M_{X_1, X_2, X_3 \ldots X_k}(t_1, 0, \ldots 0)$$

$$= (p_1 e^{t_1} + p_2 e^{0} + \ldots + p_k e^{0})^n$$

$$= (p_1 e^{t_1} + p_2 + \ldots + p_k)^n \quad \because p_1 + p_2 + \ldots + p_k = 1$$

$$= [p_1 e^{t_1} + (1 - p_1)]^n \quad p_2 + \ldots p_k = 1 - p_1$$

$$\therefore M_{X_1}(t_1) = (p_1 e^{t_1} + q_1)^n$$

This is the moment generating function of binomial distribution with parameters n and p_1.

Hence, by Consequence property of mgf $X_1 \sim B(n, p_1)$

$\Rightarrow$ for all $i = 1, 2, \ldots k$

$X_i \sim B(n, p_i)$

$\therefore E(X_i) = np_i$ and $V(X_i) = np_i q_i$

$P(X) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$

$\sum P(X) = \sum \left[\frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} \right]$

$= (p_1 + p_2 + \ldots + p_k)^n$

$= 1$

$e^{t x_1} e^{t x_2}$.

Characteristic function of Multinomial Distribution

$\phi_X(t) = E[e^{itx}]$

$= \sum e^{itx} p(x)$ (D)

$= \int_{-\infty}^{\infty} e^{itx} f(x)dx$ (C)

$-\infty < x < \infty$

If $X = (x_1, x_2, \ldots x_k) \sim M(n, p_1, p_2, \ldots p_k)$ then the pmb. of the distribution is given by,

$$P(X_1 = x_1, X_2 = x_2, \ldots X_k = x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

Now, the characteristic function of x is given by,

$$\phi_X(t) = \phi_{X_1 X_2 \ldots X_k}(t_1, t_2, \ldots, t_k)$$

$$= E[e^{it_1 x_1} e^{it_2 x_2} \ldots e^{it_k x_k}]$$

$$= \sum_x [e^{it_1 x_1} e^{it_2 x_2} \ldots e^{it_k x_k}] \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

$$= \sum_x \frac{n!}{x_1! x_2! \ldots x_k!} (p_1 e^{it_1})^{x_1} (p_2 e^{it_2})^{x_2} \ldots (p_k e^{it_k})^{x_k}$$

$$\phi_X(t) = (p_1 e^{it_1} + p_2 e^{it_2} + \ldots + p_k e^{it_k})^n.$$

Covariance between x_i and x_j in Multinomial distribution.

$(i \neq j = 1, 2, \ldots k)$

$\text{Cov}(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$

Covariance between x_i and x_j,

$\text{Cov}(x_i, x_j) = E(x_i x_j) - E(x_i) E(x_j)$

$= E(x_i x_j) - (np_i)(np_j)$ — (1)

$= E[\{x - E(x)\}\{y - E(y)\}]$

$= \frac{1}{n} \Sigma xy - \bar{x}.\bar{y}$

$= E(xy) - E(x).E(y)$

Now, $E(x_i x_j) = \left.\frac{d^2 M_x(t)}{dt_i dt_j}\right|_{t=0}$

$= \left.\frac{d^2}{dt_i dt_j} [p_1 e^{t_1} + p_2 e^{t_2} + \ldots + p_k e^{t_k}]^n\right|_{t=0}$

$= \left.\frac{d}{dt_i}\left[\frac{d}{dt_j}(p_1 e^{t_1} + \ldots + p_k e^{t_k})^n\right]\right|_{t=0}$

$= \left.\frac{d}{dt_i}\left[n(p_1 e^{t_1} + \ldots + p_k e^{t_k})^{n-1} p_j e^{t_j}\right]\right|_{t=0}$

$= \left. n p_j e^{t_j} \frac{d}{dt_i}\{(p_1 e^{t_1} + \ldots + p_k e^{t_k})^{n-1}\right|_{t=0}$

$$E(X_i X_j) = n p_j e^{t_j} (n-1)(p_1 e^{t_1} + \dots + p_k e^{t_k})^{n-2} p_i e^{t_i} \Big|_{t=0}$$

$$= n p_j (n-1)(p_1 + \dots + p_k)^{n-2} p_i e^{t_i} \Big|_{t=0}$$

$$= n p_i p_j (n-1)(1)^{n-2} \quad \left[\because \sum_{i=1}^{k} p_i = 1\right]$$

$$\left[p_1 e^{t_1} + \dots p_i e^{t_i} + \dots + p_j e^{t_j} + \dots + p_k e^{t_k}\right]^n$$

Now, eqn (1) becomes,

$$Cov(X_i, X_j) = n p_i p_j (n-1) - (n p_i)(n p_j).$$

$$= n^2 p_i p_j - n p_i p_j - n^2 p_i p_j.$$

$$\therefore Cov(X_i, X_j) = -n p_i p_j$$

It means, the multinomial distribution is negatively correlated

Correlation between X_i & X_j.

$$\gamma = \frac{Cov(X_i X_j)}{SD(X_i) \cdot SD(X_j)}$$

$$V(X_i) = n p_i q_i$$
$$= n p_i (1 - p_i)$$

$$= \frac{-n p_i p_j}{\sqrt{n p_i (1-p_i)}\sqrt{n p_j (1-p_j)}}$$

$$= \frac{-\sqrt{p_i}\sqrt{p_j}}{\sqrt{(1-p_i)}\sqrt{(1-p_j)}}$$

$$\therefore \gamma = -\left(\frac{p_i p_j}{(1-p_i)(1-p_j)}\right)^{1/2}$$

A fair dice is thrown five times,

(i) What is the prob. of getting one 6, one 5, two 4, one 3 and no others.

(ii) What is the prob. that the sum of the scores is exactly 8?

Trails:	1	2	3	4	5	Sum
Faces:	1	1	2	2	2	8
:	1	1	1	2	3	8
:	1	1	1	1	4	8

$$P(\text{Sum } 8) = p(1,1,2,2,2) + P(1,1,1,2,3) + p(1,1,1,1,4)$$

Soln:

Since, a fair dice is thrown five times so we have 5 independent multinomial experiment and at each trail we have 6 possible outcomes i.e. 1, 2, 3, 4, 5 and 6 each with prb. 1/6 $\forall$ i

$\therefore n = 5, \quad X_i = 1, 2, 3, 4, 5, 6 \quad , f(x_i) = 1/6$

(i) Now, the prb. of getting one 6, one 5, two 4, one 3 and no ones (no others) is given by,

$$P(X_1=0, X_2=0, X_3=1, X_4=2, X_5=1, X_6=1)$$

$$\left[\frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}\right]$$

$$= \frac{5!}{0!\,0!\,1!\,2!\,1!\,1!} (1/6)^0 (1/6)^0 (1/6)^1 (1/6)^2 (1/6)^1 (1/6)^1$$

$$= 0.0077.$$

(ii) A sum of 8 can be obtained from any three mutually exclusive ways (1,1,2,2,2) or (1,1,1,2,3) or (1,1,1,1,4) [In any Order]

$$\therefore P(\text{sum } 8) = P(X_1=2, X_2=3) + P(X_1=3, X_2=1, X_3=1) + P(X_1=4, X_4=1)$$

$$= \frac{5!}{2!\,3!} (1/6)^2 (1/6)^3 + \frac{5!}{3!\,1!\,1!} (1/6)^3 (1/6)^1 (1/6)^1 + \frac{5!}{4!\,1!} (1/6)^4 (1/6)^1$$

$$= 0.0045$$

HW (1)

Type	O	A	B	AB
Prb.	0.44	0.42	0.10	0.04

In a random sample of 10 Americans what is the prb. that 6 have O, 2 have A, 1 has B & 1 has AB?

(2) The prb. that player A will win any game is 30%, the prb that player B will win is 20% and the prob. that player C will is 50%. If they play 7 games, what is the prb. that A will win 2 games, player B will win 2 games and player C will win 3 games?

(3) A fair dice is thrown 15 times. what is the prb. of getting five 6s, four 5s three 4s two 3s, one 2 and no 1?

Unit 6

Non-Parametric Tests → No Distribution, No Assumption on Data.

Hypothesis Testing
- Parametric
- Non-parametric

parameter → population → μ, σ, σ^2
Statistics → Sample → $\bar{x}, r, s^2$

Parameter
Statistical Measures which are obtained from population (constant)

Statistics
Statistical Measures which are obtained from Sample. (variable)

Types of Hypothesis

1. Null Hypothesis (H_0) : $\mu = \mu_0$.
   - → No difference between sample & population parameters.
   - → True − Expected = 0

2. Alternative Hypothesis (H_1)
   - → Complementary of Null hypothesis.
   - → Mutually exclusive

$H_1: \mu \neq \mu_0$ (TTT) → No direction , $H_1: <, >, =$ (OTT)

Types of Error:
- → Reject H_0 when it is true
- → Accept H_0 when it is False.

Type-I : α = Prob = level of significance
Type-II : β

Situation / Decision	H_0 True	H_0 False
Reject H_0	WD: Type-I	Correct Decision
Accept H_0	Correct Decision	WD: Type-II

Type-I Error → ↓ Risk
α
Producers Risk

Type-II Error → ↑ Risk
β
Consumers Risk.

Critical Value (Significant Value)

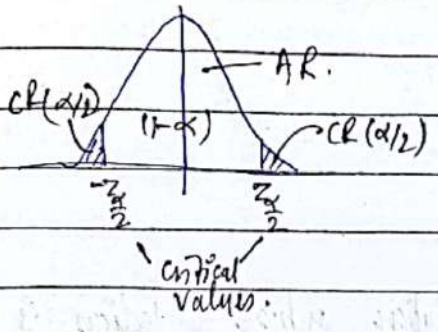
→ Value that seperates the acceptance and rejection region.

→ It depends on:

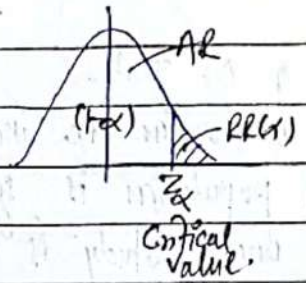
(i) Size of the level of significance used.

(ii) Type of H_1 used. (TTT or OTT)

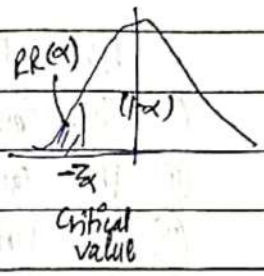
TTT



OTT (RTT)



(LTT)



p-value $\geq \alpha$ insignificant $\Rightarrow$ H_0 is true (Do not reject H_0)

p-value $\leq \alpha$ significant $\Rightarrow$ H_0 is false (H_1 is true) (Reject H_0)

Degree of freedom : (df)

(1). One Sample $\rightarrow df = n-1$

Identification of One tailed test and two tailed test.

One Tailed Test (OTT)

If the direction of difference like at least, at most, increase, decrease, minimum, maximum, superior, inferior, less than, more than, etc. are involved in the statement of hypothesis then we use OTT otherwise Two Tailed Test. (TTT)

Non-Parametric Test

Non-parametric (NP) tests are defined as those statistical tests of hypothesis in which no parameter is involved and which are based on other statistic.

In other words, those test of hypothesis are said to be non-parametric tests if the hypothesis does not involve any parameter of the population and if the measurements are on the nominal or ordinal scale.

Assumptions of NP Tests.

(i) The sample observation are independent
(ii) The sampled population is symmetric
(iii) The variable under study is continuous.
(iv) A random sample is drawn from a population whose median is unknown.

Median Test

$n \geq 20 \rightarrow$ Large Sample.

Individual Series

Median (Md) = Value of $\left(\frac{n+1}{2}\right)^{th}$ term.

Median Test is non-parametric test used to test the difference in medians of two independent distributions. In other words this test is applied to test whether two independent random samples have been drawn from two populations with same median or not. Also it is used to test wheather the two treatments applied in an experiment are equally effective or not.

Procedure

H_0 : $Md_1 = Md_2$ ie. There is no significant difference between two medians.
H_1 : $Md_1 \neq Md_2$ ie. There is significant difference between two median (TTT)
or, H_1 : $Md_1 > Md_2$ (Right) } OTT
or, H_1 : $Md_1 < Md_2$ (Left) }

Small Sample Size

$n_1 \leq 10$ and $n_2 \leq 10$

Test - Statistic

The test-statistic is obtained from following procedure :-

(i) Combine all the observations of both samples in ascending order of magnitude and find the value of median.

(ii) Count no. of observations less or equal to median in first sample and denote it by 'a'

(iii) The test-statistic is 'a'.

Critical Region

We can obtain the p-value (P_0), the probability associated with the value as extreme as the observed 'a' for n_1 and n_2.

i.e. $P_0 = P(A \leq a)$

where, $P_0 = \dfrac{{}^{n_1}C_a \; {}^{n_2}C_{k-a}}{{}^{n_1+n_2}C_k}$; $k = \dfrac{n_1+n_2}{2}$

$a = 0, 1, 2, \ldots, \min(n_1, k)$

Decision

For TTT:

If p-value $= 2P_0 > \alpha$ then we do not reject H_0 otherwise reject H_0.

For OTT

If p-value $= P_0 > \alpha$ then we do not reject H_0. otherwise reject H_0.

The following are the yield data of 10 plots under two treatment x & y.

x :	46	45	32	42	39	48	49	30	51	34
y :	44	40	59	47	55	47	50	71	43	55

Use median test to test the effectiveness of two treatments.

$\alpha = 0.05$, if NA given.

Soln,

H_0: Both the treatments x and y are equally effective.

H_1: Both the treatments x and y are not equally effective (TTT)

Test -static

Under H_0, the test statistic is 'a'

Calculation

Combining both samples in ascending order of magnitude

30, 32, 34, 39, 40, 42, 43, 44, 45, (46, 47) 47, 48, 49, 50, 51, 55, 55, 59, 71

Md = Value of $\left(\frac{n+1}{2}\right)^{th}$ term, $n = 20$

= value of 10.5^{th} term.

$= \frac{46+47}{2} = 46.5$

Now, no. of observation less or equal to Median in the first sample = 7

$\therefore a = 7$

Calculation of p-value.

p-value (p_0) $= P(A \geq a)$

$= P(A \geq 7)$

$= \sum_{a=7}^{10} \frac{{}^{n_1}C_a \; {}^{n_2}C_{k-a}}{{}^{n_1+n_2}C_k}$

$n_1 = 10$

$n_2 = 10$

$k = \frac{n_1+n_2}{2} = 10$

$\frac{{}^{n_1}C_a \; {}^{n_2}C_{k-a}}{{}^{n_1+n_2}C_k}$

$= \sum_{a=7}^{10} \frac{{}^{10}C_a \; {}^{10}C_{10-a}}{{}^{20}C_{10}}$

$= \frac{1}{{}^{20}C_{10}} \left[{}^{10}C_7 \; {}^{10}C_{10-7} + {}^{10}C_8 \; {}^{10}C_{10-8} + {}^{10}C_9 \; {}^{10}C_{10-9} + {}^{10}C_{10} \; {}^{10}C_{10-10} \right]$

$= 0.089$

Decision:

p-value $= 2P_0$ [$\because$ Problem is TTT].

$= 2(0.089)$

$=$

$\alpha = 0.05$

Since p-value $> \alpha$ so we do not reject H_0.

Hence, both the treatment x and y are equally effective.

Median Test Small Sample size $n_1, n_2 \leq 10$.

Q. A quality controller wishes to determine whether there is a difference in outcome between two different tools of software I and II. The following data shows the outcome of two different tools. Can the controller conclude that a difference exists? Use median test at 5% level of significance.

Software I	24.0	16.7	22.8	19.8	18.9	
software II	23.2	19.8	18.1	17.6	20.2	17.8

Soln,

H_0: There is no significant difference between the outcomes of two different tools.

H_1: There is significant difference between the outcomes of two different tools.

(TTT)

Combining both samples in ascending order of magnitude.

16.7 17.6 17.8 18.1 18.9 19.8 19.8 20.2 22.8 23.2 24.0

Md = Value of $\left(\frac{n+1}{2}\right)^{th}$ term ; $n=11$ [$\because n_1 = 5$ and $n_2 = 6$]

= value of 6^{th} term.

= 19.8

No. of observations less or equal to median is first sample (a) = 3.

Test statistic : Under H_0, $a=3$

Calculation of p-value:

We have, $n_1 = 5$, $n_2 = 6$

$\therefore k = \frac{n_1+n_2}{2} = 5.5 \approx 6$

$P_0 = P(A \geq a)$

$= \sum_{a=3}^{5} \frac{{}^{n_1}C_a \; {}^{n_2}C_{k-a}}{{}^{n_1+n_2}C_k}$

$= \frac{{}^5C_3 \; {}^6C_{6-3} + {}^5C_4 \; {}^6C_{6-4} + {}^5C_5 \; {}^6C_{6-5}}{{}^{11}C_6}$

$= 0.608$

Since, the problem is two tailed test so,

P-value $= 2P_0 = 2(0.608) \simeq 1$

Decision:

We have, $\alpha = 0.05$ and p-value 1.

Since p-value $> \alpha$ so we do not reject H_0.

Hence, there is no significant difference between the outcomes of two different tools.

Large Sample Size i.e. $n_1 > 10$ or $n_2 > 10$. {Chi-Square Test / χ^2-test}

	No. of obs $\leq M_d$	No. of obs $> M_d$	Total
First	a	c	a+c
Second	b	d	b+d
Total	(a+b)	(c+d)	n = a+b+c+d

Test Statistic: Under H_0.

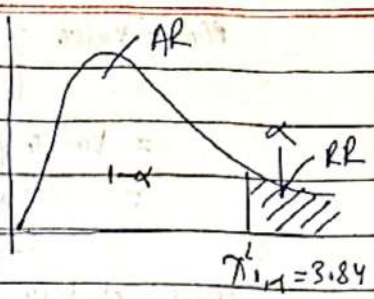
$\chi^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \sim \chi^2_{(1)}$

d.f

$df = (r-1)(c-1)$

$= (2-1)(2-1) = 1$

Decision :-



If test statistics value falls in Rejection Region(RR) then we reject H_0 otherwise don't reject H_0.

If $\chi^2_{cal} > \chi^2_{tab}$, $\Rightarrow$ Reject H_0 else don't reject H_0.

Q. The same C programming papers were marked by two A and B. The final marks were recorded as follows:

Teacher A	73	89	82	43	80	73	66	45	93	36	77	60
Teacher B	88	78	91	48	85	74	77	31	78	62	76	77

Using median test at 5% level of significance to determine if the marks distribution of two teachers differ significantly.

Stem (10)	leaf (1)
3	6 1
4	3 5 8
5	
6	6 0 2
7	3 3 7 8 4 7 8 6 7
8	9 2 0 8 5
9	3 1

Soln,

H_0 : The marks distribution of two teachers do not differ significantly.

H_1 : The marks distribution of two teachers differ significantly.

Calculation.

Arranging given data in ascending order of magnitude.

31, 36, 43, 45, 48, 60, 62, 66, 73, 74, 76, 72, 77, 77, 78, 78, 80, 82, 85, 88, 89 91, 93.

Med = Value of $\left(\frac{n+1}{2}\right)^{th}$ item , $n = 24$

= Value of 12.5^{th} term

= $\frac{76+77}{2} = 76.5$

Test-statistics: For large sample, the test-statistic is given by,

$$\chi^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

$$= \frac{24(49-25)^2}{12.12.12.12}$$

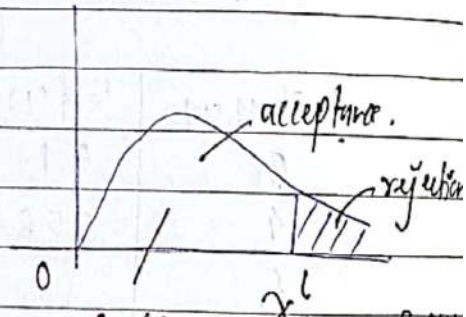
$$= 0.667$$

	$x_i \le$ Med	$x_i >$ Med	Total
First	$a=7$	$c=5$	12
Second	$b=5$	$d=7$	12
Total	12	12	$n=24$

Critical Value:

We have $\alpha = 0.05$ and $df = 1$

$$\chi^2_{tab} = \chi^2_{1,0.05} = 3.841$$



Decision:

Since, the test-statistic value falls in the acceptance region. So do not reject H_0.

Since $\chi^2_{cal} < \chi^2_{tab}$ so we do not reject H_0.

Hence, the marks distribution of two teachers do not differ significantly.

Mann Whitney 'U' test.

This test is most powerful non-parametric test for testing the hypothesis of difference between two independent locations / two independent samples samples. This is non-parametric alternative test of t-test.

Procedure

H_0 : $Md_1 = Md_2$ i.e. There is no significant difference between two locations or two Medians.

H_1 : $Md_1 \neq Md_2$ i.e. There is significant difference between two locations or two Md.

or, H_1 : $Md_1 > Md_2$ (Right) } OTT

or H_1 : $Md_1 < Md_2$ (Left) }

Small Sample Size. i.e. n_1 or $n_2 \leq 10$

Test-statistic: Under H_0.

U_0 = minimum of $\{U_1$ and $U_2\}$

Where, $U_1 = n_1 \cdot n_2 + \frac{n_1(n_1+1)}{2} - R_1$

$U_2 = n_1 \cdot n_2 + \frac{n_2(n_2+1)}{2} - R_2$

where, R_1 = Sums of the combined ranks given to the first sample.

R_2 = Sum of the combined ranks given to the second sample.

$U_1 + U_2 = n_1 \cdot n_2$

Critical value

For two tailed test $U_{tab} = U_{\alpha, (n_1, n_2)}$

For One tailed test, $U_{tab} = U_{\alpha/2, (n_1, n_2)}$

Decision

If $U_0 > U_{tab}$ then we do not reject H_0 Otherwise reject H_0.

Q. The nicotine contents of two brands of cigarettes, measured in milligrams was found to be as follows.

Brand A : 2.1 4.0 6.3 5.4 4.8 3.7 6.1 3.3

Brand B : 4.1 0.6 3.1 2.5 4.0 6.2 1.6 2.2 1.9 5.4.

Is there any significant difference between two brands of cigarettes. Use Mann-Whitney U-test at $\alpha = 0.10$?

Soln

H_0: There is no significant difference between two brands of cigarettes.

H_1: There is significant " " " " " "

Test statistic: Under H_0.

U_0 = minimum of $\{U_1$ and $U_2\}$

Check

$U_1 + U_2 = n_1 n_2$

$23 + 57 = 8(10)$

Calculation:

A	B	Combined Rank of A	Combined Rank of B
2.1	4.1	4	12
4.0	0.6	10.5	1
6.3	3.1	18	7
5.4	2.5	14.5	6
4.8	4.0	13	10.5
3.7	6.2	9	17
6.1	1.6	16	2
3.3	2.2	8	5
	1.9		3.
	5.4		14.5
		$R_1 = 93$	$R_2 = 78$

4.0 → 10

4.0 → 11

21/2

= 10.5

Now,

$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$

$= 8(10) + \frac{8(8+1)}{2} - 93$

$= 23$

and

$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$

$= 8(10) + \frac{10(10+1)}{2} - 78$

$= 57$

Critical value:

We have $\alpha = 0.10$ and $(n_1, n_2) = (8, 10)$

$\therefore U_{\alpha, (n_1, n_2)} = U_{0.10, (8,10)}$ (TTT)

$= 20$.

Decision:

Since, $U_0 = 23 > U_{0.10, (8,10)} = 20$, we do not reject H_0.

Hence, there is no significant difference between.....

Large Sample Size i.e. $n_1 > 10$ or $n_2 > 10$.

For Large sample size the distribution of U_0 is normal with mean $= \frac{n_1 n_2}{2}$ and s.d $= \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$

$$\therefore Z = \frac{U_0 - \mu_{U_0}}{\sigma_{U_0}}$$

$$= \frac{U_0 - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}}$$

$X \sim N(\mu, \sigma^2)$
i.e. X follows N.D. with mean μ & stan. σ^2.
$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
Std. Normal.

Critical Value:

The critical value of the test-statistic is obtained from Z-table at α level of significance for one tailed and two tailed test.

$\therefore Z_{tab} = Z_\alpha$.

$\alpha = 0.05$

One	Two
1.645	1.96

Decision:

If $|Z_{cal}| > Z_{tab}$ then we reject H_0 otherwise do not reject H_0.

Note:

In case of tied rank

$$\sigma_{U_0} = \sqrt{\frac{n_1 n_2}{n(n-1)} \left[\frac{n^3 - n}{12} - \sum T_i\right]}$$

Where $T_i = \frac{\sum (t_i^3 - t_i)}{12}$

t_i = no. of times ith rank is repeated.

Q. A farmer wishes to determine whether there is a difference in yields between two different varieties of wheat I and II. The following data shows the production of wheat per unit area using the two varieties. Can the farmer conclude at significance level 0.01 that a difference exists?

Wheat I : 15.9 15.3 16.4 14.9 15.3 16.0 14.6 15.3 14.5 16.6 16.0

Wheat II : 16.3 16.8 17.1 16.9 18.0 15.6 18.1 17.2 15.4

$n_1 = 11$ and $n_2 = 9$ i.e. Large Sample Size (z-test)

H_0: There is no significant difference in the yields of wheat I and II or the difference does not exist.

H_1: There is significant diff. " " " " " or the difference exists. (IT

Calculation:

Wheat I	Wheat II	Combined Rank I	Rank II
15.9	16.3	9	12.5
15.3	16.8	5	15
16.4	17.1	12.5	17
14.9	16.9	3	16
15.3	16.0	5	14
16.0	15.6	10.5	8
14.6	18.1	2	20
15.3	17.2	5	18
14.5	15.4	1	7
16.6		14	
16.0		10.5	

$R_1 = 77.5$ $R_2 = 132.5$

$t_1 = 3$, $t_2 = 2$ $t_3 = 2$

Now,

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$= 11(9) + \frac{11(11+1)}{2} - 77.5$$

$$= 87.5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$= 11(9) + \frac{9(9+1)}{2} - 132.5$$

$$= 11.5$$

$[U_1 + U_2 = n_1 n_2]$

$\Rightarrow 87.5 + 11.5 = 11(9)$

$$E(U_0) = \frac{n_1 n_2}{2} = \frac{99}{2} = 49.5$$

$$S.d(U_0) = \sqrt{\frac{n_1 n_2}{n(n-1)} \left\{ \frac{n^3 - n}{12} - \Sigma T_i \right\}}$$

Where, $\Sigma T_i = \frac{\Sigma(t_i^3 - t_i)}{12} = \frac{3^3-3}{12} + \frac{2^3-2}{12} + \frac{2^3-2}{12} = 3.$

$\therefore Sd(U_0) = \sigma_{U_0} = \sqrt{\frac{11(9)}{20(20-1)}\left\{\frac{20^3-20}{12} - 3\right\}}$

$= 13.13.$

Test statistic: Under H_0.

$Z = \frac{U_0 - E(U_0)}{Sd(U_0)}$ $\quad [U_0 = \text{minimum}\{U_1, U_2\}]$

$= \frac{11.5 - 49.5}{13.13}$

$= -2.89$

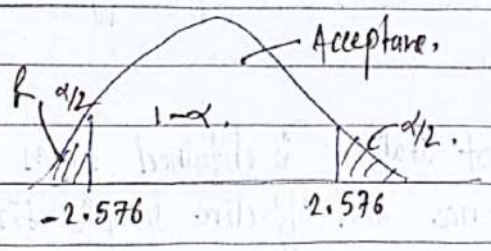
$\therefore |Z_{cal}| = 2.89$

Critical Value.

We have, $\alpha = 0.01$

$\therefore Z_{tab} = Z_{0.01}$ (TTT)

$= 2.576$



Calculated > tabulated.
Then we reject H_0 otherwise do not reject H_0.

Decision:
Since $|Z_{cal}| > Z_{tab}$ so we reject H_0.
Hence, there is significant difference in the yields of wheat I and II.

Paired Test (For Dependent Samples)

Wilcoxon Matched Pairs Signed Rank Test. [small sample size $n \leq 25$]

This test is applied to test the hypothesis concerning the difference between two treatments used on the two random samples. It can also be applied to test the hypothesis concerning the effectiveness of some activity (treatment)

Procedure:

H_0 : $Md_1 = Md_2$ i.e. no. sig. diff before/after } Md_1 = Median score before Treatm.
H_1 : $Md_1 \neq Md_2$ i.e. sig. diff before/after } Md_2 = Median score after Treatment
or, H_1 : $Md_1 > Md_2$ (Right) } OTT.
or H_1 : $Md_1 < Md_2$ (Left) }

Test statistic : Under H_0.

T = minimum of $\{S(+)$ and $S(-)\}$;

$S(+)$ = sum of the ranks with '+' sign

$S(-)$ = sum of the ranks with '-' sign.

[Ranks are given to the absolute value of difference i.e. d]

Critical Value.

The critical value of the test-statistic is obtained from wilcoxon table at α level of significance and effective sample size $n_e(n)$

i.e. $T_{tab} = T_{\alpha, n_e}$

$n_e = n -$ no. of $[d_i = 0]$

Decision

If $T \leq T_{\alpha, n_e}$ then we reject H_0. Otherwise do not reject H_0.

Seven prospective graduate student took a test twice with the following scores:

First Attempt : 4.70 5.30 6.10 4.40 6.00 5.90 5.60

Second Attempt : 5.10 5.50 6.80 4.90 5.85 6.20 5.98

Test wheather there is significant difference between scores in first attempt and second attempt using wilcoxon Matched pair signed rank test.

$n=7$

H_0: There is no significant difference between the scores of first & second attempts.

H_1: There is significant " " " " of " (TTT)

Test statistic : Under H_0.

T = minimum of $\{S(+), S(-)\}$

$\therefore T = 3$

Calculation

First attempt (X)	Second attempt (Y)	$d = X - Y$	Rank of $\lvert d \rvert$	Rank with + sign	Rank with − sign
4.70	5.10	-.40	6		6
5.30	5.50	-.20	4		4
6.10	6.80	.10	1	1	
4.40	4.90	-.50	7		7
6.00	5.85	.15	2	2	
5.90	6.20	-.30	5		5
5.80	5.98	-.18	3		3
				$S(+)=3$	$S(-)=25$

Critical value

We have, $\alpha = 0.05$ and $n = n_e = 7$.

$\therefore T_{7, 0.05}\ (TTT) = 2$.

Decision :

Since $T(=3) > T_{7, 0.05}(=2)$. So, we do not reject H_0. Hence there is no significant difference between the scores of first and second attempts.

ignore 'D'

Q. To evaluate a speed reading course, a group of 10 subject were asked to read two comparable articles one before the course and one after the course. Their scores on reading test are as follows.

Before course(X): 57 40 64 70 90 59 76 98 70 83

After course(Y): 60 90 62 70 95 58 80 99 75 94

Test whether the course is beneficial using the Wilcoxon matched pairs signed rank test at 5% level of significance.

H_0: $M_{dx} = M_{dy}$ i.e. the course is not beneficial

H_1: $M_{dx} < M_{dy}$ i.e. " " " beneficial (OTT)

Test statistic: Under H_0,

T = minimum of $\{S(+), S(-)\}$

Calculation:

X	Y	d	Rank of |d|	Rank with +sign	Rank with −sign
57	60	−3	4		4
40	90	−50	9		9
64	62	2	3	3	
70	70	0	—		
90	95	−5	6.5		6.5
59	58	1	1.5	1.5	
76	80	−4	5		5
98	99	−1	1.5		1.5
70	95	−5	6.5		6.5
83	94	−11	8		8
				$S(+) = 4.5$	$S(-) = 40.5$

$\therefore T = 4.5$

Critical Value:

We have, $\alpha = 0.05$ & $n_e = n - 1 = 10 - 1 = 9$ $\therefore T_{9,0.05\,(OTT)} = 8$

Decision: Since, $T(=4.5) < T_{9,0.05\,(OTT)} = 8$. Hence the course is beneficial. We reject H_0.

Kolmogorov Smirnov Test (K-S Test) – One Sample.

It is used to test the difference between observed and expected frequencies.

H_0: $F(x) = F(y)$ i.e. There is no significant difference between observed and expected frequencies or frequencies are uniformly distributed.

H_1: $F(x) \neq F(y)$ i.e. There is significant " " " " are not ... (TTT)

or, H_1: $F(x) > F(y)$ } OTT

or H_1: $F(x) < F(y)$ }

Test-statistic: Under H_0,

D_0 = maximum of $|f_e - f_o|$

where,

f_e = Relative Cumulative frequency = $\frac{Cf_e}{n}$

$f_o = \frac{Cf_o}{n}$

Cf_e = Cumulative frequency,

o = Observed frequency

e = expected frequency.

Critical value or tabulated value:

The critical value of the test statistic is obtained from kolmogorov Smirnov test table at α level of significance and n no. of observations for one tailed and two tailed tests.

Decision:

If $D_0 \geq D_{n,\alpha}$ then we reject H_0 otherwise do not reject H_0.

Q. The number of laptop in 10 different departments are given below. Test whether the laptop are uniformly distributed over the entire office use ks test.

Department No.:	1	2	3	4	5	6	7	8	9	10
No. of laptop:	8	10	9	12	15	7	5	12	13	9

H_0: The laptop are uniformly distributed over the entire office.
H_1: The laptop are not " " " " " "

Test-statistic: Under H_0:
D_0 = maximum of $|Fe - Fo|$: expected freq^n = NP

Calculation:

Dept.	Observed freq (O)	Expected freq (e)	cfo	$Fo = \frac{cfo}{n}$	cfe	$fe = \frac{cfe}{n}$	$\|Fe-Fo\|$
1	8	10	8	8/100	10	10/100	2/100
2	10	10	18	18/100	20	20/100	2/100
3	9	10	27	29/100	30	30/100	3/100
4	12	10	39	39/100	40	40/100	1/100
5	15	10	54	54/100	50	50/100	4/100
6	7	10	61	61/100	60	60/100	1/100
7	5	10	66	66/100	70	70/100	4/100
8	12	10	78	78/100	80	80/100	2/100
9	13	10	91	91/100	90	91/100	1/100
10	9	10	100	100/100	100	100/100	0

$\Sigma O = 100$.

$\therefore D_0 = 4/100 = 0.04$.

Critical value,
we have, $\alpha = 0.05$, & $n = 100$.
$\therefore D_{n,\alpha} = D_{100,0.05}$ (TTT) $= \frac{1.22}{\sqrt{100}} = 0.122$.

Decision:
Since $D_0 < D_{n,\alpha}$, so we do not reject H_0.
Hence, Laptops are uniformly distributed over the entire offices.

Q. In a certain computer hardware manufacturing industry six different types of machines are working to cut pieces of wires. The number of wires of unequal length recorded in a day is as follows:

Machine	1	2	3	4	5	6
No. of wires	2	0	4	8	5	11

Do these Data provide sufficient evidence that the six machines equally cut the wires of unequal length? Apply Kolmogorov Smirnov test at 5% level of significance.

Soln,

H_0: The data provide sufficient evidence that the six machines equally cut the wires of unequal length.

H_1: The data do not provide ..... (TTT)

Test-Statistic: Under H_0, $\rightarrow e = \Sigma o / no.\text{ of groups}$

Machine	Observed(o)	Expected(e)	cfo	$Fo = cfo/n$	cfe	$Fe = cfe/n$	$\lvert Fe - Fo \rvert$
1	2	5	2	2/30	5	5/30	3/30
2	0	5	2	2/30	10	10/30	8/30
3	4	5	6	6/30	15	15/30	9/30
4	8	5	14	14/30	20	20/30	6/30
5	5	5	19	19/30	25	25/30	0/30
6	11	5	30	30/30	30	30/30	0

$\therefore D_0 = 9/30 = 0.3.$

Critical Value: or Tabulated Value,

$D_{n,\alpha} = D_{30,0.05}$ (TTT) $= 0.242$

Decision: Since, $D_0 > D_{n,\alpha}$ so we reject H_0.

Hence, the data do not provide sufficient evidence that the six machines....

Kolmogorov Smirnov Test for two sample:

It is non-parametric test used to test whether two independent samples are from same population or not.

Procedure:

H_0: $F(x) = F(y)$ i.e. the two independent samples are from same population or same distribution

H_1: $F(x) \neq F(y)$ i.e. The two independent samples are not from ... (TTT)

or, H_1: $F(x) > F(y)$ } OTT

or H_1: $F(x) < F(y)$ }

Small Sample Size

(i). $n_1 = n_2 \leq 40$

(ii) if $n_1 \neq n_2 \leq 20$

Test-statistic: Under H_0

D_0 = maximum of $|F(x) - F(y)|$

where, $F(x) = \frac{cf_x}{n_1}$ and, $F(y) = \frac{cf_y}{n_2}$

Critical Value:

$D_{tab} = D_{(n_1, n_2), \alpha}$

Decision:

If $D_0 \geq D_{(n_1, n_2), \alpha}$ then we reject H_0, otherwise do not reject H_0.

Large Sample size:

$n_1 = n_2 > 40$ and $n_1 \neq n_2 > 20$

Test-statistic

For two tailed test,

D_0 = maximum of $|F(x) - f(y)|$

For one tailed test,

$$\chi^2 = 4 D_0^2 \left(\frac{n_1 n_2}{n_1 + n_2}\right) \quad \therefore \text{OTT}, \chi^2 \Rightarrow df = 2$$

Critical value

$$\chi^2_{tab} = \chi^2_{2,\alpha} \rightarrow \boxed{df = 2}$$

Decision

If calculated $\geq$ tabulated then we reject H_0. otherwise do not reject H_0.

Q. Life in years of two types cells used in laptop are given below:

cell x:	4	6	5	6	3	4	5	3	5
cell y:	2	5	4	3	4	2	4	3	5.

Test whether. life of two brand of cells are same or not? Use ks test at $\alpha = 0.05$.

Soln

H_0: The life of two brands of cells are same.

H_1: " " " " " are not same (TTT).

D_0 = maximum of $|F(x) - F(y)|$.

Calculation:

Life in years	Freq of x	Freq of y	cfx	cfy	F(x)	F(y)	$\lvert F(x) - F(y) \rvert$
2	0	2	0	2	0/9	2/9	2/9
3	2	2	2	4	2/9	4/9	2/9
4	2	3	4	7	4/9	7/9	3/9 = D_0.
5	3	2	7	9	7/9	9/9	2/9
6	2	0	9	9	9/9	9/9	0
	$n_1 = 9$		$n_2 = 9$.				

$\therefore D_0 = 3/9$.

Critical value: We have $n_1 = n_2 = 9$ and $\alpha = 0.05$

$\therefore D_{9,0.05}\ (TTT) = 5/4$.

more → right
less → left.

Decision:

Since $D_0 < D_{9,0.05}$ so we do not reject H_0.

Hence the life of two brands of cells are same.

3. Given below represent monthly income distn of employees in a hardware company and software company.

Income (000 Rs)	No. of employees in Hw Company	No. of emp. in Sw Company
20-30	6	12
30-40	10	18
41-50	11	16
50-60	13	12
60-70	25	10
70-80	15	12
80-90	10	10
	$n_1 = 90$	$n_2 = 10$

Do the income distribution support that income of employees in hardware company is more than income of employees in software company? Use ks test.

$Md_1 = Md_2$

$Md_1 \neq Md_2$.

H_0: The income of employees of Hardware company is not more than that of software company

H_1: The income " " " " " " " is more " "

Test statistics: Under H_0,

$$\chi^2 = 4D_0^2 \left(\frac{n_1 n_2}{n_1 + n_2} \right) \sim \chi^2_{(2)}.$$

Income (000 Rs)	Hw Comp. fx	Sw. Company fy	$F(x) = \frac{cfx}{90}$	$F(y) = \frac{cfy}{90}$	$\lvert F(x) - F(y) \rvert$
20-30	6	12	6/90	12/90	
30-40	10	18	16/90	30/90	
40-50	11	16	27/90	46/90	
50-60	13	12	40/90	58/90	
60-70	25	10	65/90	68/90	
70-80	15	12	80/90	80/90	
80-90	10	10	90/90	90/90	
	90	90			

$n_1 = 90$ & $n_2 = 90$.

$D_0 = 19/90$

$$\chi^2 = 4\left(\frac{19}{90}\right)^2\left[\frac{90(90)}{90+90}\right]$$

$\chi^2_{cal} = 8.022$.

Critical Value, we have $\alpha = 0.05$ & $d.f = 2$.

$\chi^2_{tab} = \chi^2_{2,0.05} = 5.991$.

Decision

Since, $\chi^2_{cal} > \chi^2_{tab}$ so we reject H_0.

Date
Page

W

Kruskal-Wallis H Test (H ⇒ p-value) (At least 3 sample)

or Kruskals Wallis One way ANOVA Test.

It is used to test wheather the three or more populations differ in location (median) or not

Notation:

k = no. of samples (≥ 3)

n_1 = size of first sample

n_2 = Size of second sample

n_k = Size of kth sample.

Procedure:

H_0: $Md_1 = Md_2 = \ldots = Md_k$ i.e. there is no significant difference between k-pop^n medians or k-sample medians.

H_1: $Md_1 \neq Md_2 \neq \ldots \neq Md_k$ i.e. there is significant .... or at least one of the median is different.

Small Sample Size i.e. $k=3$ and $n_i \leq 5$; $i = 1, 2, \ldots, k$

Test Statistic: Under H_0,

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

R_1 = Sum of the combined ranks given to the first sample.

R_k = Sum of the combined ranks given to the kth sample.

$n = n_1 + \ldots + n_k$.

If tied occurs then corrected test-statistic is obtained as follows.

$$H_{corrected} = \frac{H}{cf}$$

where, cf = correction factor

$$= 1 - \frac{\sum (t_i^3 - t_i)}{n^3 - n}$$

t_i = no. of times ith rank is repeated

Critical Value.

For different values of n_i $(i = 1, 2, \ldots, k)$, the probability p_0 associated with the value as extreme as observed H (Real) is obtained from the K-W table i.e. $p_0 = P(H \geq H_{obs})$.

Decision

If p-value $[p_0] \leq \alpha$ then we reject H_0 otherwise do not reject H_0.

Large Sample Size i.e. $k \geq 3$ and $n_i > 5$

In this case, the sampling distribution of 'H' can be approximated by a chi-square distribution with $(k-1)$ degree of freedom.

$$\chi^2_{tab} = \chi^2_{(k-1),\alpha}$$

Decision:

If $H \geq \chi^2_{(k-1),\alpha}$ then we reject H_0 otherwise don't reject H_0.

Q. For the following scores of 3 groups, apply kruskal Wallis H test to test the hypothesis that the three groups are not significantly different.

Group	Scores.				
A	96	128	83	61	101
B	82	124	132	135	104
C	115	149	166	147	—

Soln.

$k = 3$, $n_1 = 5$, $n_2 = 5$ and $n_3 = 4$ $\therefore n = 14$

H_0: Three groups are not significantly different.

H_1: Three groups are significantly different.

Calculation:

	Scores. (combined rank)					R_i	R_i^2/n_i
A	96 (4)	128 (9)	83 (3)	61 (1)	101 (5)	$R_1 = 22$	$R_1^2/n_1 = 96.8$
B	82 (2)	124 (8)	132 (10)	135 (11)	104 (6)	$R_2 = 37$	$R_2^2/n_2 = 273.8$
C	115 (7)	149 (13)	166 (14)	147 (12)		$R_3 = 46$	$R_3^2/n_3 = 529$
							$\sum \frac{R_i^2}{n_i} = 899.6$

Test-Statistic : Under H_0,

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{14(14+1)} \cdot (899.6) - 3(14+1)$$

$\therefore H = 6.405.$

Critical Value : (look near to 6.405)

For, $H = 6.405$, $n_1 = 5$, $n_2 = 5$, $n_3 = 7$ and $k = 3$ we have

p-value (P_0) = 0.044.

Decision :

We have, $\alpha = 0.05$ and p-value = 0.044.

Since p-value $< \alpha$ so we reject H_0. Hence, there is significant difference between three groups.

Q. The following data represent the operating times in hour for 3 types of Scientific pocket calculators before a charge is required. Calculator A 4.9, 6.1, 4.3, 4.6, 5.3 } Harrelled $\chi^2_{(k-1),0.01}$

B 5.5, 5.4, 6.2, 5.8, 5.5, 5.2, 4.8

C 6.4, 5.8, 5.6, 6.5, 6.3, 6.6

Use Kruskal Wallis test, at the 0.01 level of significance test the hypothesis that the operating times for all three calculators are equal.

$k = 3$, $n_1 = 5$, $n_2 = 7$ and $n_3 = 6$.

H_0: The operating time of all three calculators are same.

H_1: The operating time of all three calculators are different.

Calculations:

		Combined Rank					
Calculator A	4.9 (4)	6.1 (12)	4.3 (1)	4.6 (2)	5.3 (9)		
B	5.5 (8.5)	5.4 (7)	6.2 (13)	5.8 (11)	5.5 (8.5)	5.2 (5)	4.8 (3)
C	6.4 (15)	5.8 (18)	5.6 (10)	6.5 (16)	6.3 (14)	6.6 (17)	

$t_j = 2$, $R_1 = 45$, $R_2 = 56$ and $R_3 = 90$.

Since 5.5 is repeated two times.

Correction factor,

$$C.F = 1 - \frac{\sum (t_i^3 - t_i)}{n^3 - n} = 1 - \frac{2^3 - 2}{18^3 - 18} = 0.999$$

Test-Statistic: Under H_0

$$H_{corrected} = \frac{H}{C.F} = \frac{\frac{12}{n(n+1)} \sum_{i=1}^{3} \frac{R_i^2}{n_i} - 3(n+1)}{0.999}$$

$$= \frac{12/18(18+1)\left[\frac{25}{5} + \frac{56^2}{7} + \frac{90^2}{6}\right] - 3(18+1)}{0.999} = 10.48$$

Critical Value

We have, $k = 3$ and $\alpha = 0.05$

$$\chi^2_{tab} = \chi^2_{(k-1),\alpha} = \chi^2_{2, 0.05} = 5.991$$

Decision: Since, $H_{corrected} > \chi^2_{tab}$ so we reject H_0. Hence, the operating time of all three calculators are different.

Friedman Test or Friedman two way ANOVA test. ($k \geq 3$)

The friedman test is designed to test whether the population of $k (\geq 3)$ treatments are identical or not i.e. to test whether all the treatments have equal effect or not, or to test whether all treatments are equally effective or not.

Procedure:

H_0: $Md_1 = Md_2 = \ldots = Md_k$ i.e. there is no significant difference between k-population or k-sample medians.

H_1: $Md_1 \neq Md_2 \neq \ldots \neq Md_k$ i.e. there is significant ..... or at least one median is different.

Test-Statistic: Under H_0,

$$F_r = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1) \quad ; n = \text{size of each sample,}$$

[Note that, here ranks are given according to block or column wise]

If there is tie in the observation then the corrected value of the test-statistic is calculated as,

$$\text{Fr}_{\text{corrected}} = \frac{\text{Fr}}{\text{Cf}} \quad ; \quad \text{Cf} = 1 - \frac{\sum (t_i^3 - t_i)}{n(k^3 - k)}$$

where, t_i = no. of times ith rank is repeated.

Case-I :- Small Sample Size. i.e.

(a) $k = 3$ and $2 \le n \le 9$.

(b) $k = 4$ and $2 \le n \le 5$.

Critical Value.

For given sample size n, no. of samples k and the calculated value of the test-statistic i.e. Fr, we obtain the p-value P_0 from the Friedman probability table.

Decision : If P-value (P_0) $> \alpha$ then we do not reject H_0. Otherwise reject H_0.

Case-II : Except above two cases we use large sample size case. Here, the dist^n of the test-statistic follows χ^2-distribution with $(k-1)$ d.f. i.e. Fr $\sim \chi^2_{(k-1)}$.

Decision : If $\text{Fr} > \chi^2_{(k-1),\alpha}$ then we reject H_0. Otherwise don't reject H_0.

Q. The scores of 3 matched groups under the six conditions are given below.

Group	Conditions.					
	I	II	III	IV	V	VI
A	9	5	2	5	6	7
B	6	4	3	4	6	5
C	5	1	3	3	6	5

Apply the Friedman two way ANOVA test to identify if there is significant difference in the variation between matched groups. Use 5% level of significance.

Soln, We have, $k=3$ and $n=6$

H_0: There is no significant difference in the variation between three matched groups.

H_1: There is significant ,,,,,,,,,,,,

Calculation:

Conclusion.

Group	I	II	III	IV	V	VI	R_i	R_i^2
A	9(3)	5(3)	2(1)	5(3)	6(2)	7(3)	15	225
B	6(2)	4(2)	3(1.5)	4(2)	6(2)	5(1.5)	12	144
C	5(1)	1(1)	3(2.5)	3(1)	6(2)	5(1.5)	9	81

$t_1=2$, $t_2=3$, $t_3=2$.

Calculation:

$$cf = 1 - \frac{\sum(t_i^3 - t_i)}{nk(k^2-1)} = \frac{1-(2^3-2)(3^3-3)(2^3-2)}{6(3^3-3)} = 0.75.$$

Test-statistic: Under H_0,

$$F_{(corrected)} = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1) \Big/ cf$$

$$= \left\{ 12/6(3)(4) \left[225+144+81 \right] \right\} - 3(6)(4) \Big/ 0.75 = 4.$$

Critical Value: We have $k=3$, $n=6$ and $F_{corr} = 4$. We have

P-value $(P_0) = 0.184$.

Decision: we have $\alpha = 0.05$, $P_0 = 0.184$,

Since, P-value $> \alpha$ so we do not reject H_0. there there is no significat difference in the variation between three matched group.

Q. Three different advertising media TV, Radio and News paper are being compared to study their effectiveness in promoting sales of Waiwai noodles. Each advertising media is exposed for specified period of time and sales (000 pkg) from 10 stores located at different areas are recorded.

Advertising Media. Stores.

	A	B	C	D	E	F	G	H	I	J
TV	20	21	15	12	14	17	21	16	20	18
Radio	7	9	11	12	10	10	14	12	8	7
News Paper	8	6	11	12	9	6	8	10	8	6

Are three advertising media equally effective, use Friedmans two way ANOVA test.

Soln, $k = 3$ $n = 10$

H_0: There is no significant difference....

H_1: There is significant difference....

Adv. Media. Stores.

	A	B	C	D	E	F	G	H	I	J	R_i	R_i^2
TV	20(3)	21(3)	15(3)	12(1)	14(3)	17(3)	21(3)	16(3)	20(3)	18(3)	29	841
Radio	7(1)	9(2)	11(1.5)	12(2)	10(2)	10(2)	14(2)	12(2)	8(1.5)	7(2)	18	324
News Paper	8(2)	6(1)	11(1.5)	12(2)	9(1)	6(1)	8(1)	10(1)	8(1.5)	6(1)	13	169

$t_1 = 2$ $t_2 = 3$ $t_3 = 2$

Calculation.

$$CF = 1 - \frac{\sum t_i^2 - t}{n(k^3-k)} = 1 - \left[\frac{(2^3-2)+(3^3-3)+(2^3-2)}{10(3^3-3)}\right] = 0.825$$

Test-statistic: Under H_0

$$\text{Friedman}(t) = \left(\frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1)\right) / CF = 15.76.$$

Critical Value:

We have $k=3$, $n=10$, Friedman $= 15.76$.

$(n-1) = (3-1) = 2 = d.f.$, $\chi^2_{0.01,2} = 5.99$,

$F_r > \chi^2_{0.01,2}$

Decision: Reject H_0

Unit 7

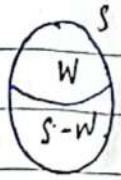
Hypothesis Testing.

Critical Region.

Let us define a sample space 'S' and we divide the whole sample space into two regions or two disjoint subsets say W and S-W such that.

$W \cup S-W = S.$

and $W \cap S-W = \phi$ (null set)



If the sample points fall into the region W, then we can reject the null hypothesis (H_0) and the region is called region of rejection. or it is called critical region and if the sample points fall into the region (S-W), then we can accept the null hypothesis and the region is called region of acceptance.

Type-I error

α = Prob. (Type I error)

= Prob (reject H_0 when H_0 is true)

= Prob ($x \in W / H_0$)

$= \int_{x \in W} f(x/H_0)\, dx$

Type-II error

β = Prob (Type II error)

= Prob. (accept H_0 / H_0 false)

= Prob. ($x \in S-W / H_1$)

$= \int_{x \in S-W} f(x/H_1)\, dx.$

Power of the test.

$1-\beta$ = 1 - Prob (Type II error)

= 1 - Prob (accept H_0 / H_0 is false)

= Prob (reject H_0 / H_0 is false)

= Prob (accept H_1 / H_1 is true)

= Prob (correct decision)

Q. Tom a coin 5 times, if 2 or 3 heads turn up then the coin is accepted as an unbiased coin. Find α, β and the power of the test when $H_0: P = 1/2$ against $H_1: P = 3/4$.

Soln,

Since, 1 coin is tossed five times so, $S = \{0,1,2,3,4,5\}$, $n=5$.

Let X = no. of heads then $X \sim B(n,p)$ so

$P(X=x) = {}^{n}C_x\, p^x q^{n-x}$; $x = 0,1,2,\ldots n$, $p+q=1$

Here,

$S-W = \{2,3\}$, $W = \{0,1,4,5\}$.

Now,

$P(X/H_0) = {}^{5}C_2 (1/2)^2 (1/2)^{5-x} = {}^{5}C_x (1/2)^5$.

and,

$P(x/H_1) = {}^{5}C_x (3/4)^x (1-3/4)^{5-x} = {}^{5}C_x (3/4)^x (1/4)^{5-x}$.

$\alpha = Prob(x \in 0,1,4,5 / H_0)$

Here, $\alpha = 1 - Prob(x \in 2,3 / H_0)$

$= 1 - [P(X=2/H_0) + P(X=3/H_0)]$

$= 1 - [{}^{5}C_2 (1/2)^5 + {}^{5}C_3 (1/2)^5]$

$= 1 - 0.625$

$\therefore \alpha = 0.375$.

$\beta = Prob.(x \in S-W / H_1)$

$= Prob.(x \in W / H_1)$

$= Prob(X=2/H_1) + P(X=3/H_1)$

$= {}^{5}C_2 (3/4)^2 (1/4)^{5-2} + {}^{5}C_3 (3/4)^3 (1/4)^{5-3}$

$=$

Power of test, $1-\beta =$

Q Let P be the Prob. of getting a head in a single toss of a coin. The coin is tossed 5 times and A is desired to test $H_0: P = 1/2$ against $H_1: P = 2/3$. The null hypothesis H_0 is rejected if more than 3 heads are obtained. Find the prob of type I, type II error Also, find power of test

Given, $n=5$

X = no. of heads.

$S = \{0,1,2,3,4,5\}$

$H_0 : P = 1/2$ against $H_1 : P = 2/3$.

Here, $X \sim B(n,p)$

So, $P(X=x) = {}^{n}C_x\, p^x q^{n-x}$; $x = 0,1,\ldots n$.

$W = \{4,5\}$.

$S-W = \bar{W} = \{0,1,2,3\}$.

Now, $\alpha = Prob.(x \in W / H_0)$

$= P(X=4/H_0) + P(X=5/H_0)$

$\int e^{ax}dx = \frac{e^{ax}}{a}$

$\therefore \alpha = {}^5C_4 (1/2)^4 (1/2)^{5-4} + {}^5C_5 (1/2)^5 (1/2)^{5-5}$

and,

$\beta = \text{Prob.}(x \in S-W / H_1)$

$= 1 - P(x \in W / H_1)$

$= 1 - [P(X=4/H_1) + P(X=5/H_1)]$

$= 1 - [{}^5C_4 (2/3)^4 (1/3)^{5-4} + {}^5C_5 (2/3)^5 (1/3)^{5-5}]$

$= 0.54.$

Power of the test $= 1-\beta = 1-0.54 = 0.46.$

Q. If $x \geq 1$ is the critical region for testing $H_0 : \theta = 2$ against alternative hypothesis $H_1 : \theta = 1$ on the basis of single observation from the pop^n,

$f(x, \theta) = \theta e^{-\theta x}, \quad 0 \leq x < \infty.$

obtain the values of type I and type II errors. Also find the power of the test.

Soln, Given $W = \{x : x \geq 1\}$.

and $S-W = \{x : x < 1\}$.

Further, $H_0 : \theta = 2$ Vs $H_1 : \theta = 1$

and $f(x, \theta) = \theta e^{-\theta x}, \quad x > 0$

$\alpha = f(x : x \geq 1 / H_0) = \int_1^\infty f(x/H_0)dx = \int_1^\infty [\theta e^{-\theta x}]_{\theta=2} dx.$

$= \int_1^\infty 2e^{-2x} dx = 2\left[\frac{e^{-2x}}{-2}\right]_1^\infty = [e^{-2x}]_\infty^1 = e^{-2} - e^{-\infty}$

$= e^{-2} - 0$

$\therefore \alpha = 0.13.$

$\beta = \text{Prob}(x \in S-W / H_1)$

$= f(x : x < 1 / \theta = 1)$

$= \int_0^1 f(x/\theta=1)dx$

$= \int_0^1 [\theta e^{-\theta x}]_{\theta=1} dx.$

$= \int_0^1 e^{-x} dx$

$= \left[\frac{e^{-x}}{-1}\right]_0^1 = [e^{-x}]_1^0 = e^0 - e^{-1}$

$\beta = 1 - e^{-1} = 0.632$

$\therefore$ Power of test $= 1-\beta = 1-0.632$

Q. Given the frequency function, $f(x,\theta) = \begin{cases} 1/\theta & ; \ 0 \le x \le \theta \\ 0 & ; \ \text{otherwise} \end{cases}$

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, type II error, if you choose the interval (i) $0.5 \le x$ (ii) $1 \le x \le 1.5$ as the critical region? Also obtain the power of test.

Given,

$f(x,\theta) = 1/\theta \ ; \ 0 \le x \le \theta$

$H_0 : \theta = 1$ Vs. $H_1 : \theta = 2$

(i). $w = \{x : x \ge 0.5\}$

and $S - w = \{x : x < 0.5\}$.

Now, $\alpha = f(x \in w / H_0)$

$$= \int_{0.5}^{\theta} \left[\frac{1}{\theta}\right]_{\theta=1} dx = \int_{0.5}^{1} dx = [x]_{0.5}^{1} = 0.5.$$

$\beta = f(x \in S-w / H_1)$

$$= \int_{0}^{0.5} \left[\frac{1}{\theta}\right]_{\theta=2} dx = \frac{1}{2}[x]_{0}^{0.5} = \frac{1}{2}(0.5) = \frac{1}{4} = 0.25.$$

$\therefore \beta = 0.25$

$\therefore$ Power of the test $(1-\beta) = 0.75$.

(ii). $w = \{x : 1 \le x \le 1.5\}$ and $S - w = \{x : x < 1 \text{ or } x > 1.5\}$.

$\alpha = f(x \in w / H_0)$

$$= \int_{1}^{1.5} \left[\frac{1}{\theta}\right]_{\theta=1} dx$$

Here, the prb. does not exist,

Since, $0 \le x \le \theta$ and $\theta = 1$ $\therefore \alpha = 0$

$\beta = f(x \in S-w / H_1)$

$$= 1 - \int_{x \in w} f(x/H_1)\,dx = 1 - \int_{1}^{1.5} \left[\frac{1}{\theta}\right]_{\theta=2} dx = 1 - \frac{1}{2}[x]_{1}^{1.5} = 1 - \tfrac{1}{2}[1.5-1]$$

$$= 0.75.$$

Power of test $= 1 - \beta = 0.25$

$X \sim N(\mu, \sigma^2)$

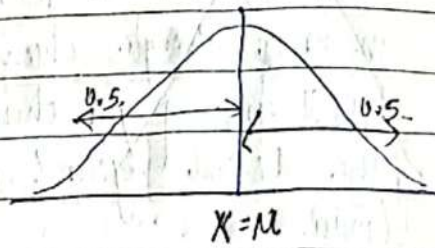
X follows Normal Distribution with mean μ and variance σ^2

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

↓

Standard Normal Distribution.

$E[Z] = 0$, $Var[Z] = 1$



$\mu = 8$, $\sigma = 2$

(i) $P(X > 10) = P\left(Z > \frac{10-8}{2}\right)$

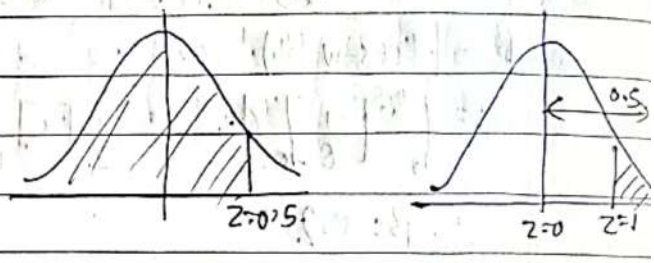
$= P(Z > 1)$

$= 0.5 - P(0 < Z < 1)$

(ii) $P(X < 9) = P\left(Z < \frac{9-8}{2}\right)$

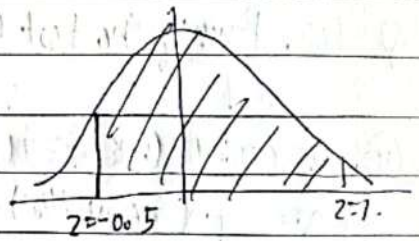
$= P(Z < 0.5)$

$= 0.5 + P(0 < Z < 0.5)$



(iii) $P(7 < X < 10) = P\left(\frac{7-8}{2} < Z < \frac{10-8}{2}\right)$

$= P(-0.5 < Z < 1)$



Q. Suppose it is desired to test the hypothesis $H_0: \mu = 35$ against the alternative $H_1: \mu \neq 35$ on the basis of random sample of size 16 from a Normal population $N(\mu, 1)$. The decision rule is reject H_0, if the sample mean $\bar{X} < 34.65$ or $\bar{X} > 35.35$

(a) Find the prob of type I error

(b) Find the prob of type II error (i) when $\mu = 36$ (ii) when $\mu = 36.1$

(c) Also find the power of the test.

Soln

$H_0: \mu = 35$ vs $H_1: \mu \neq 35$.

$n = 16$ & $\bar{X} \sim N(\mu, 1)$.

$W = \{\bar{x} : \bar{x} < 34.65 \text{ or } \bar{x} > 35.35\}$.

$S - W = \{\bar{x} : 34.65 < \bar{x} < 35.35\}$

$\therefore \bar{x} \sim N[\mu, 1/n]$

Note:

$X \sim N(\mu, \sigma^2)$

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

n = sample size

$\bar{x} \sim N(\mu, \sigma^2/n)$

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

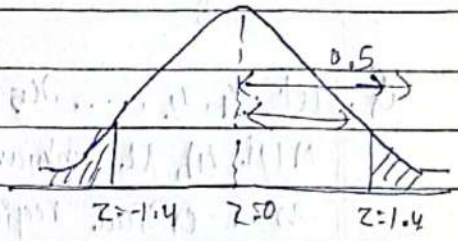
(a) Probability of type I error.

α = probability $(x \in W / H_0)$

$= \text{Prob}(\bar{x} \in \bar{x} < 34.65 \text{ or } \bar{x} > 35.35)$ (under H_0)

$= P(\bar{x} < 34.65 / H_0) + P(\bar{x} > 35.35 / H_0)$

$= P\left(z < \frac{34.65 - 35}{1/\sqrt{16}}\right) + P\left(z > \frac{35.35 - 35}{1/\sqrt{16}}\right)$



$= P(z < -1.4) + P(z > 1.4)$

$= 2P(z > 1.4)$ (By Symmetric Law)

$= 2[0.5 - P(0 < z < 1.4)]$

$= 2[0.5 - 0.4192]$

$= 0.1616$

(b) $\mu = 36$.

Size or probability of type II error $\Rightarrow \beta = P(x \in S - W / H_0)$

$\beta = \text{Prob}(\bar{x} \in 34.65 < \bar{x} < 35.35 / H_1)$

$= P(34.65 < \bar{x} < 35.35 / \mu = 31)$

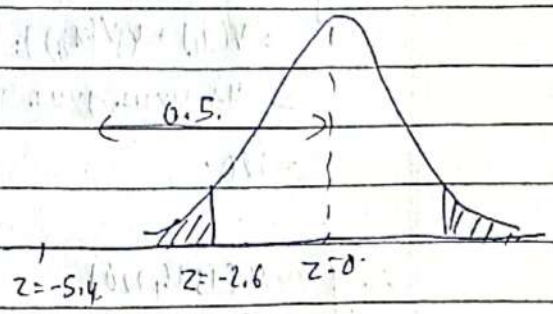
$= P\left(\frac{34.65 - 36}{1/\sqrt{16}} < z < \frac{35.35 - 36}{1/\sqrt{16}}\right)$

$= P(-5.4 < z < -2.6)$.

$= 0.5 - P(-2.6 < z < 0)$

$= 0.5 - 0.4953$

$= 0.0047$



Now,

Power of test $= 1 - \beta = 1 - 0.0047 = 0.9953$

$\bar{X} \sim N(\mu, \sigma^2) \Rightarrow z = \frac{\bar{X}-\mu}{\sigma}$

$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

(n). $\mu_1 = 36.1$

$\beta = P(\bar{X} \leq 34.65 < \bar{X} < 35.35 / H_1)$

$= P(34.65 < \bar{X} < 35.35 / \mu = 36.1)$

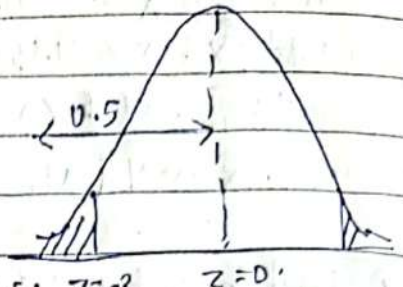
$= P\left(\frac{34.65 - 36.1}{1/\sqrt{16}} < z < \frac{35.35 - 36.1}{1/\sqrt{16}}\right)$

$= P(-5.8 < z < -3)$

$= 0.5 - P(-3 < z < 0)$

$= 0.5 - 0.4987$

$= 0.0013$



Now,

Power of test $= 1 - \beta = 1 - 0.0013 = 0.9987$

Q. Let $x_1, x_2, \ldots, x_5$ be a random sample of size 5 drawn from a $N(\mu, 4)$, μ unknown if you chose $x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 \geq 0$. as a critical region for testing $H_0: \mu = -1$ against $H_1: \mu = 1$, find the size of the critical region and power of the test.

Soln,

$n = 5$, $X \sim N(\mu, 4)$

let $y = x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5$

$E(y) = E(x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5)$

$= E(x_1) + 2E(x_2) + 3E(x_3) + 4E(x_4) + 5E(x_5)$

$= \mu + 2\mu + 3\mu + 4\mu + 5\mu$.

$= 15\mu$.

$V(y) = V[x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5]$

$= V(x_1) + 4V(x_2) + 9V(x_3) + 16V(x_4) + 25V(x_5)$ $\because V(x) = \sigma^2 V(x)$ &

$= 4 + 4\times4 + 9\times4 + 16\times4 + 25\times4$. x_i's are independent

$= 220$. $(\because X \sim N(\mu, 4))$

$\therefore y \sim N(15\mu, 220)$

$W = \{y : y > 0\}$. Critical Region

$S - W = \{y : y < 0\}$ Acceptance Region.

Now,

$\alpha = P(y \geq w / H_0)$ (Type I error)

$= P(y > 0 / \mu = -1)$

$= P\left(z > \frac{0 - (-1)}{\sqrt{220}}\right)$ $\quad z = \frac{\bar{x} - \mu}{\sigma}$

$= P(z > 1.01)$

$= (0.5 - 0.3438)$

$= 0.1562$

$\beta = \text{prob.}(y < s - w / H_1)$

$= P(y < 0 / \mu = -1)$

$= P\left(z < \frac{0 - (-1)}{\sqrt{220}}\right)$

$= P(z < 1.01)$

$= 0.1562$

Unit 7

Testing of Hypothesis.

Cont..

Types of Hypothesis Testing.

(i) Simple Hypothesis.
(ii) Composite Hypothesis

(i) Simple Hypothesis.

The hypothesis which completely specifies the distribution of random variable is known as simple hypothesis. for example. $X \sim N(\mu, \sigma^2)$ and the hypothesis to be tested as, $H_0 : \mu = \mu_0, \sigma^2 = \sigma_0^2$ then under this hypothesis the form of the distribution is completely specified.

(ii) Composite Hypothesis.

A hypothesis which does not completely specifies the distribution then it is said to be composite hypothesis. For example $X \sim N(\mu, \sigma^2)$ and the hypothesis to be tested as $H_0 : \mu = \mu_0$. then under this hypothesis σ^2 is not specified.

Likelihood function.

Let $x_1, x_2, \ldots x_n$ be a random sample of size 'n' drawn from a population with pdf $f(x, \theta)$, then the joint pdf or pmf of $x_1, x_2, \ldots x_n$ is said to be likelihood function of $X = x_1, x_2, \ldots x_n$ and it is denoted by L or $L(\theta)$ or $L(x, \theta)$. It is given by,

$$L = \prod_{i=1}^{n} f(x_i, \theta)$$

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

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

pdf = Prb. density fun.
pmf = Prb. mass fun.
iid = identically independent of

State the likelihood function of $x_i \sim P(\lambda)$; $i = 1, 2, \ldots n$

Soln, Here, $x_i \sim P(\lambda)$

$$\Rightarrow P(x_i) = \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$$

$$\therefore P(x_1) = \frac{e^{-\lambda} \lambda^{x_1}}{x_1!}, \quad P(x_2) = \frac{e^{-\lambda} \lambda^{x_2}}{x_2!}, \ldots, \quad P(x_n) = \frac{e^{-\lambda} \lambda^{x_n}}{x_n!}$$

Now, the likelihood function of $X = x_1, x_2, \cdots, x_n$ is given by,

$$L = \prod_{i=1}^{n} (x_i, \lambda)$$

$$= P(x_1, \lambda), P(x_2, \lambda) \cdots P(x_n, \lambda)$$

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

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

State the likelihood function of $x_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots n$.

Soln, Here,

$x_i \sim N(\mu, \sigma^2)$ So

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \forall\ i = 1, 2, \ldots n.$$

If $X \sim N(\mu, \sigma^2)$ then

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

$; -\infty < x < \infty$

$-\infty < \mu < \infty$

$\sigma > 0$

Now, the likelihood function of $x = x_1, x_2 \ldots x_n$ is given by,

$$L = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2)$$

$$= f(x_1; x_1, \sigma^2)\, f(x_2; \mu, \sigma^2) \ldots f(x_n; \mu, \sigma^2)$$

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

$$\therefore \boxed{L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2}}$$

Best Critical Region and Best test.

A critical region of law size α is said to be a best critical region if it has minimum β or maximum power $(1-\beta)$ among all critical regions.

A test which minimizes β or maximizes the power $(1-\beta)$ for a desired law level of significance α, is known as best test.

Most powerful critical region (MPCR) and most powerful test (MP test)

Let x be an arbitrary sample point in a sample space S. Then a critical region w is said to be a most powerful critical region (MPCR) or (BCR) of size α for testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative hypothesis $H_1 : \theta = \theta_1$, if

$$\alpha = P(x \in w / H_0) \quad \text{— (I)}$$

and, $P(x \in w / H_1) \geq P(x \in w_1 / H_1)$

for every other critical region w_1 satisfying eqn (I) and a test corresponding to this MPCR of size α is called most powerful (MP) test of size α.

Neyman Pearson lemma (NP-Lemma)

NP lemma provides the general method of finding a most powerful test of a simple null hypothesis against the simple alternative hypothesis.

Statement: Let $L(x/H_0)$ and $L(x/H_1)$ be the likelihood function under H_0 and H_1 respectively, the BCR to test the simple null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative hypothesis $H_1 : \theta = \theta_1$ is given by,

$$\left(\frac{L(x/H_1)}{L(x/H_0)}\right) \geq k \quad \forall \; x \in w$$
$$< k \quad \forall \; x \in S - w$$

Such that $\int_{x \in w} L(x/H_0)\,dx \leq \alpha$ fixed.

and k = non-negative constant which is derived from fixed α.

S = Sample space, w = critical region.

Then the test is most powerful test (NP-test) and critical region is BCR.

Proof:

Since, α is the prob. of type first error.

$$\alpha = \text{Prob (Type I error)}$$
$$= \text{Prob.}(x \in w / H_0)$$
$$= \int_{x \in w} L(x/H_0)\,dx \quad \text{— (I)}$$

and, $\beta = \text{Prob (Type II error)}$

$$= \text{prob.}(x \in S - w / H_1)$$
$$= \int_{x \in S-w} L(x/H_1)\,dx \quad \text{— (II)}$$

Let w^* be any other critical region with prob. type I error α_1 (which is less than or equal to α) and prob. of type II error β_1.

$\therefore \alpha_1 = \text{Prob}(x \in w^* / H_0)$

$\Rightarrow \alpha_1 = \int_{x \in w^*} L(x/H_0)\,dx$

Similarly from β_1,

$$\beta_1 = \int_{x \in S-w^*} L(x/H_1)\,dx$$

Since, $\alpha \geq \alpha_1$

$$\int_{x \in w} L(x/H_0)\,dx \geq \int_{x \in w^*} L(x/H_0)\,dx \quad \text{— (iii)}$$

Further,

$1-\beta = \int_{x \in w} L(x/H_1)\,dx$ and $1-\beta_1 = \int_{x \in w^*} L(x/H_1)\,dx$

Now,

$$\beta_1 - \beta = \int_{x \in w} L(x/H_1)\,dx - \int_{x \in w^*} L(x/H_1)\,dx$$