B.Sc. MATHEMATICS – I YEAR

DJM1C : STATISTICS

SYLLABUS

UNIT I : Correlation - Karl Pearson's coefficient of correlation, Lines of Regression - Regression coefficient - Rank Correlation.

UNIT II : Probability - Definition - Applications of Addition and multiplication, theorems - Conditional, Probability - Mathematical Expectations - Moment generating function - Special Distributions, (Binomial Distributions, Poisson Distribution, Normal Distribution - Properties).

UNIT III : Association of Attributes - Coefficient of Association - Consistency - Time Series - Definition - Components Of Time Series - Seasonal and cyclic variations.

UNIT IV : Sampling - Definition - Large samples. Small samples- Population with one samples and population with two samples - Students - t - test - Applications - chi - square test and goodness of fit - applications.

UNIT V : Index Numbers - Types of index numbers - Tests - Unit test commodity reversal test, time reversal test, factor reversal test - Chain index numbers - cost of living index - Interpolation - Finite differences operators - Newton's forward, backward interpolation formulae, Lagrange's formula.

Books:

1. Statistics: S. Arumugam & others
2. Statistics: D.C.Sancheti & Kapoor
3. Statistics: Mangaladas & others
4. Statistics: T.Sankaranarayana & others

UNIT I : **CORRELATION AND REGRESSION**

Correlation - Karl Pearson's coefficient of correlation, Lines of Regression - Regression coefficient - Rank Correlation.

CORRELATION AND REGRESSION

CORRELATION:

Definition:

Consider a set of bivariate data (x_i, y_i);i=1,2….n. If there is a change in one variable corresponding to a change in the other variable we say that the variables are correlated.

If the two variable two variables deviate in the same direction the correlation is said to be direct or positive.

Definition:

The covariance between x and y is defined by

Cov (x,y) = $\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n}$

Hence $\gamma_{xy} = \frac{cov\ (x,y)}{\sigma_x \sigma_y}$

Example:

The heights and weights of five students are given below.

Height in c.m x	160	161	162	163	164
Weight in kgs y	50	53	54	56	57

$Here\ \bar{x} = 162; \bar{y} = 54\ ;\ \sigma_x = \sqrt{2}\ and\ \sigma_y = \sqrt{6}$

Now $\sum(x_i - \bar{x})(y_i - \bar{y}) = (-2)(-4) + (-1)(-1) + 0 + (1 \times 2) + (2 \times 3) = 17$

$\therefore \gamma_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n\sigma_x\sigma_y} = \frac{17}{5\sqrt{2}\sqrt{6}} = \frac{17\times\sqrt{12}}{60} = \frac{17\times 3.46}{60}$ = 0.98

Theorem:

$$\gamma_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\left[n\sum x_i^2 - (\sum x_i)^2\right]^{1/2}\left[n\sum y_i^2 - (\sum y_i)^2\right]^{1/2}}$$

Proof:

$$\gamma_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + n\bar{x}\bar{y}$$

$$= \sum x_i y_i - \bar{x}(n\bar{y}) - \bar{y}(n\bar{x}) + n\bar{x}\bar{y}$$

$$= \sum x_i y_i - n\bar{x}\bar{y}$$

$$= \sum x_i y_i - (\tfrac{1}{n}) \sum x_i y_i$$

$$= \tfrac{1}{n}[\mathrm{n}\sum x_i y_i - \sum x_i \sum y_i] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots.. (2)$$

Also,

$$\sigma_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$$

$$= \tfrac{1}{n}\ [\sum x_i^2 - 2\bar{x}\sum x_i + n(\bar{x})^2]$$

$$= \tfrac{1}{n}\ [\sum x_i^2 - 2n(\bar{x})^2 + n\ (\bar{x})^2]$$

$$= \tfrac{1}{n}\left[\sum x_i^2 - \tfrac{1}{n}(\sum x_i)^2\right]$$

$$= \tfrac{1}{n^2}[n\sum x_i^2 - (\sum x_i)]^2$$

$$\therefore \sigma_x = \tfrac{1}{n}\ [n\sum x_i^2 - (\sum x_i)^2]^{1/2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots. (3)$$

Similarly,

$$\sigma_y = \frac{1}{n}\left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]^{1/2} \qquad \ldots\ldots\ldots\ldots\ldots.(4)$$

Substituting (2),(3) and (4) in (1) we get the required result.

Theorem:

-1 $\leq \gamma \leq 1$

Proof:

$$\gamma_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n\sigma_x\sigma_y}$$

$$= \frac{\left(\frac{1}{n}\right)\sum(x_i-\bar{x})(y_i-\bar{y})}{\left[\frac{1}{n}\sum(x_i-\bar{x})^2\right]^{1/2}\left[\frac{1}{n}\sum(y_i-\bar{y})^2\right]^{1/2}}$$

Let $a_i = x_i - \bar{x}$ and $b_i = y_i - \bar{y}$

$$\therefore \gamma_{xy}^2 = \frac{\left(\sum a_i b_i\right)^2}{\left(\sum a_i^2\right)\left(\sum b_i^2\right)}$$

By Schwartz inequality we have,

$$\left(\sum a_i b_i\right)^2 \leq \left(\sum a_i\right)^2\left(\sum b_i\right)^2$$

Hence $\gamma_{xy}^2 \leq 1$

$$\therefore \left|\gamma_{xy}\right| \leq 1$$

$$\therefore -1 \leq \gamma \leq 1$$

Note: 1

If $\gamma = 1$ *the correlation is perfet and positive.*

Note: 2

If $\gamma = -1$ the correlation is perfect and negative.

Note: 3

If $\gamma = 0$ the variables are uncorrelated.

Note: 4

If the variables x and y are uncorrelated then cov (x,y) = 0

Problem: **1**

Ten students obtained the following percentage of marks in the college internal test (x) and in the final university examination (y). Find the correlation coefficient between the marks of the two tests.

x	51	63	63	49	50	60	65	63	46	50
y	49	72	75	50	48	60	70	48	60	56

Solution:

Choosing the origin A = 63 for the variable x and B= 60 for y and taking $u_i = x_i - A$ and $v_i = y_i - B$.

We have the following table:

x_i	u_i	y_i	v_i	u_i^2	v_i^2	u_iv_i
51	-12	49	-11	144	121	132
63	0	72	12	0	144	0
63	0	75	15	0	225	0
49	-14	50	-10	196	100	140
50	-13	48	-12	169	144	156
60	-3	60	0	9	0	0
65	2	70	10	4	100	20
63	0	48	-12	0	144	0
46	-17	60	0	289	0	0
50	-13	56	-4	169	16	52
Total	-70	-	-12	980	994	500

$$\gamma_{xy} = \gamma_{uv}$$

$$= \frac{n\sum u_iv_i - \sum u_i \sum v_i}{[n\sum u_i^2 - (\sum u_i)^2]^{1/2}[n\sum v_i^2 - \sum(v_i)^2]^{1/2}}$$

$$= \frac{10\times500-(-70)\times(-12)}{[10\times980-(-70)^2]^{1/2}[10\times994-(-12)^2]^{1/2}}$$

$$= \frac{4160}{70\times98.97}$$

$$=0.6$$

Problem: 2

If x and y are two variable. Prove that the correlation coefficient between $ax + b$ and $cy + d$ is

$$\gamma_{ax+b,cy+d} = \frac{ac}{|ac|}\gamma_{xy}\ \ if\ a,c\ \neq 0.$$

Proof:

Let $u = ax + b$and $v = cy + d$

$$\therefore \bar{u} = a\bar{x} + b\ and\ \bar{v} = c\bar{y} + d$$

$\sigma_u^2 = \frac{1}{n}\sum(u - \bar{u})2$

$= \frac{a^2}{n}\sum(x_i - \bar{x})2$

$=a^2\sigma_x^2$

Similarly,

$$\sigma_v^2 = c^2\sigma_y^2$$

Now,

$\gamma_{uv} = \frac{\sum(u-\bar{u})(v-\bar{v})}{n\sigma_u\sigma_v}$

$= \frac{\sum a(x-\bar{x})c(y-\bar{y})}{n|ac|\sigma_x\sigma_y}$

$= \frac{ac}{|ac|}\ \gamma_{xy}\,.$

Problem: 3

A programmer while writing a program for correlation coefficient between two variable x and y from 30 pairs of observations obtained the following results $\sum x = 300;\ \sum x^2 = 3718, \sum y = 210;\ \sum y^2 = 2000;\ \sum xy = 2100.$ At the time of checking it was found that he had copied down two pairs (x_i, y_i) as (18, 20) and (12, 10) instead of the correct values (10,15) and (20,15). Obtain the correct value of the correlation coefficient.

Solution:

Corrected $\sum x = 300 - 18 - 12 + 10 + 20$

$= 300$

Corrected $\sum y = 210 - 20 - 10 + 15 + 15$

$= 210$

Corrected $\sum x^2 = 3718 - 18^2 - 12^2 + 10^2 + 20^2$

$= 3750$

Corrected $\sum y^2 = 2000 - 20^2 - \ 10^2 + 15^2 + 15^2$

$= 1950$

Corrected$\sum xy = 2100 - (18 \times 20) - \ (12 \times 100) + (10 \times 15) + (20 \times \ \ 15)$= 2070

After correction the correlation coefficient is,

$$\gamma_{xy} = \frac{n\sum xy - \ \sum x \sum y}{[n\sum x^2 - (\sum x)^2]^{1/2}[n\sum y^2 - (\sum y)^2]^{1/2}}$$

$$\gamma_{xy} = \frac{30 \times 2070 - 300 \times 210}{[30 \times 3750 - 300^2]^{1/2}\left[[30 \times 1950 - 210^2\ ]^{\frac{1}{2}}\right]}$$

$$= \frac{62100 - 63000}{(112500 - 9000)^{1/2}(58500 - 44100)^{1/2}}$$

$$= \frac{-900}{(22500)^{\frac{1}{2}}(14400)^{1/2}}$$

$$= - \frac{900}{150 \times 120}$$

$$= - \frac{1}{20}$$

$$= - 0.05$$

Problem: 4

$If\ x, y\ and\ z$ are uncorrelated variable each having same standard deviation obtain the coefficient of correlation between $x + y\ and\ y + z$.

Solution:

$Given\ \sigma_x = \sigma_y = \sigma_z = \sigma$.

$X\ and\ y$ are uncorrelated ⇒ $\sum(x - \bar{x})(y - \bar{y}) = 0$

$Y\ and\ z$ are uncorrelated ⇒ $\sum(y - \bar{y})(z - \bar{z}) = 0$

$Z\ and\ x$ are uncorrelated ⇒ $\sum(z - \bar{z})(x - \bar{x}) = 0$

Let $u = x + y\ and\ v = y + z$

$\bar{u}$ = $\bar{x}$+$\bar{y}$ and $\bar{v}$=$\bar{y}$+$\bar{z}$

Now ,

$$\sigma_u^2 = \frac{1}{n}\sum(u - \bar{u})^2$$

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

$$=\frac{1}{n}[\sum(x - \bar{x})^2 + \sum(y - \overline{y)}^2 + 2\sum(x - \bar{x})(y - \bar{y})]$$

$$= \sigma_x^2 + \sigma_y^2\ \ [since\ \sum(x - \bar{x})(y - \bar{y}) = 0]$$

$$= 2\sigma^2$$

Similarly,

$$\sigma_v^2 = 2\sigma^2$$

Now, $\sum(u - \bar{u})(v - \bar{v}) = \sum[\{x - \bar{x}) + (y - \bar{y})\}\{(y - \bar{y}) + (z - \bar{z})\}]$

$$= \sum(x - \bar{x})(y - \bar{y}) + \sum(y - \bar{y})^2 + \sum(x - \bar{x})(z - \bar{z}) +$$

$$y{-}y(z{-}z)$$

$$= 0 + n\sigma_y^2 + 0 + 0$$

$$= n\sigma^2$$

$$\gamma_{uv} = \frac{\sum(u-\bar{u})(v-\bar{v})}{n\sigma_u\sigma_v}$$

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

$$= ½.$$

Problem: 5

Show that the variable $u = xcos\alpha + y\,sin\,\alpha\ and\ v = ycos\alpha - xsin\alpha\ are\ uncorrelated$. If $\alpha = \frac{1}{2}\tan^{-1}\left(\frac{2\gamma_{xy\,\sigma_x\sigma_y}}{{\sigma_x}^2-{\sigma_y}^2}\right)$

Solution:

$$u_{i=}\ x_i cos\alpha + y_i\ sin\alpha\ and\ v_i = y_i\ cos\alpha - x_i\ sin\alpha$$

$$\bar{u} =\ \bar{x}cos\alpha + \bar{y}sin\alpha\ and\ \bar{v} = \bar{y}cos\alpha - \bar{x}sin\alpha$$

$$u_i - \bar{u} = (x_i - \bar{x})\ cos\alpha + (y_i - \bar{y})sin\alpha$$

The variable u_i and v_i are uncorrelated if $\sum(u_i - \bar{u})(v_i - \bar{v}) = 0$

$$\sum[(x_i-\bar{x})cos\alpha + (y_i-\bar{y})sin\alpha][(y_i-\bar{y})cos\alpha - (x_i-\bar{x})sin\alpha] = 0$$

$\therefore \sum(x_i - \overline{x})\ (y_i - \bar{y})\cos^2\alpha - \sum(x_i-\bar{x})(y_i-\bar{y})\sin^2\alpha - cos\alpha sin\alpha[\sum(x_i-\bar{x})^2 -$ *yi−y2]=0*

$$\therefore n\gamma_{xy\,\sigma_x\sigma_y}(\cos^2\alpha - \sin^2\alpha) =\ ncos\alpha sin\alpha\left(\sigma_x^2 - \sigma_y^2\right)$$

$$\therefore \gamma_{xy\,\sigma_x\sigma_y}\cos 2\alpha = \frac{1}{2}sin2\alpha\left(\sigma_x^2 - \sigma_y^2\right).$$

$$\therefore tan2\alpha =\ \frac{2\gamma_{xy}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}$$

$$\therefore\ \alpha = \frac{1}{2}\tan^{-1}\left(\frac{2\gamma_{xy}\sigma_x\sigma_y}{\sigma_x^2-\sigma_y^2}\right)$$

Rank correlation:

Suppose that a group of n individuals are arranged in the order of merit or efficiency with respect to some characteristics. Then the rank is a variable which takes only the values 1,2,3…n. assuming that there is no tie.

Hence $\bar{x}$= $\frac{1+2.\ldots\ldots\ldots\ldots+n}{n}$ = $\frac{n+1}{2}$ and the variance is given by $\sigma_x^2 = \frac{1}{12}\,(n^2-1)$.

Theorem:

Rank correlation ρ is given by

$$\rho = 1 - \frac{6\sum(x-y)^2}{n(n^2-1)}.$$

Proof:

Consider a collection of n individuals

let x_i and y_i be the ranks of the i^{th} individual in the two different rankings.

$\therefore\ \bar{x} = \frac{1}{2}(n+1) = \bar{y}$ and $\sigma_x^2 = \frac{1}{12}(n^2-1) = \sigma_y^2$.

Now, $\sum(x-y)^2 = \sum[(x-\bar{x})-(y-\bar{y})]\,^2 (since\ \bar{x} = \bar{y})$

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

$$= n\sigma_x^2 + n\sigma_y^2 - 2n\rho\sigma_x\sigma_y$$

$$= 2\text{n}\sigma_x^2(1-\rho) \quad (since\ \sigma_x^2 = \sigma_y^2)$$

$$= \frac{1}{6}\ n\ (n^2-1)(1-\rho)$$

$$1-\rho = \frac{6\sum(x-y)^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6\sum(x-y)^2}{n(n^2-1)}.$$

Problem: 6

Find the rank correlation coefficient between the height in cm and weight in kg of 6 soldiers in Indian Army.

Height	165	167	166	170	169	172
Weight	61	60	63.5	63	61.5	64

Solution:

Height	**Rank in height x**	**Weight**	**Rank in weight y**	**x-y**	$(\boldsymbol{x-y})^2$
165	6	61	5	1	1
167	4	60	6	-2	4
166	5	63.5	2	3	9
170	2	63	3	-1	1
169	3	61.5	4	-1	1
172	1	64	1	0	0
Total	-	-	-	-	-16

$$\rho = 1 - \frac{6\sum(x-y)^2}{n(n^2-1)} = 1\text{-}\ \frac{6\times 16}{6\times 35}$$

$$= 1\text{-}0.457$$

$$= 0.543.$$

Problem: 7

From the following data of marks obtained by 10 students in physics and chemistry. Calculate the rank correlation coefficient.

Physics (P)	35	56	50	65	44	38	44	50	15	26
Chemistry (Q)	50	35	70	25	35	58	75	60	55	35

Solution:

We rank the marks of physics and chemistry and we have the following table.

P	**Rank in p x**	**Q**	**Rank in Q y**	**x-y**	$(\boldsymbol{x-y})^2$
35	8	50	6	2	4
56	2	35	8	-6	36

50	3.5	70	2	1.5	2.25
65	1	25	10	-9	81
44	5.5	35	8	-2.5	6.25
38	7	58	4	3	9
44	5.5	75	1	4.5	20.25
50	3.5	60	3	0.5	0.25
15	10	55	5	5	25
26	9	35	8	1	1
Total	-	-	-	-	185

We observe that in the values of x the marks 50 and 44 occurs twice. In the values of y the mark 35 occurs thrice.

Hence in the calculation of the rank correlation coefficient $\sum(x-y)^2$ is to be corrected by adding the following correction factors

$$\left[\frac{2(2^2-1)}{12}+\frac{2(2^2-1)}{12}\right]+\frac{3(3^2-1)}{12}=3$$

After correction $\sum(x-y)^2 = 188$.

Now,

ρ= 1- $\frac{6\sum(x-y)^2}{n(n^2-1)}$

= 1 - $\frac{6\times188}{10\times99}$

= 1- $\frac{1128}{990}$

= 1-1.139

= -0.139.

Problem: 8

Three judges assign the ranks to 8 entries in a beauty contest.

Judge Mr.x	1	2	3	4	5	6	7	8
Judge Mr.y	3	2	1	5	4	7	6	8
Judge Mr.y	1	2	3	4	5	7	8	6

Which pair of Judges has the nearest approach to common in beauty?

Solution:

Table for the rank correlation coefficients $\rho_{xy}\,,\rho_{yz}\,,\rho_{zx}$

x	y	Z	x-y	$(x-y)^2$	y-z	$(y-z)^2$	z-x	$(z-x)^2$
1	3	1	-2	4	2	4	0	0
2	2	2	0	0	0	0	0	0
3	1	3	3	9	-2	4	-1	1
4	5	4	-2	4	1	1	1	1
5	4	5	3	9	-1	1	-2	4
6	7	7	-1	1	0	0	1	1
7	6	8	-1	1	-2	4	3	9
8	8	6	0	0	2	4	-2	4
Total			-	28	-	18	-	20

ρ_{xy} = 1- $\frac{6\sum(X-y)^2}{n(n^2-1)}$

= 1- $\frac{6\times 28}{8\times(8^2-1)}$

= 1- $\frac{168}{504}$

=1-0.333= 0.667.

ρ_{yz} = 1- $\frac{6\times18}{8\times63}$

= 1- $\frac{108}{504}$

= 1-o.214

= 0.786.

ρ_{zx}= 1- $\frac{6\times20}{8\times63}$

$= 1- \frac{120}{504}$

$= 1-0.238$

$= 0.762.$

Since ρ_{yz} is greater than ρ_{xy} andρ_{xz} the judges Mr. Y and Mr. Z have nearest approach to common taste in beauty.

Problem: 9

The coefficient of rank correlations of marks obtained by 10 students in mathematics and physics was found to be 0.8. It was later discovered that the differences in rank in two subjects obtained by one of the students was wrongly taken as 5 instead of 8. Find the correct coefficient of rank correlation.

Solution:

$$\rho_{xy} = 1- \frac{6\sum(x-y)^2}{n(n^2-1)}.$$

Given ρ_{xy} = 0.8 and n=10

$$0.8 = 1- \frac{6\sum(x-y)^2}{10(10^2-1)}$$

$$= 1- \frac{6\sum(x-y)^2}{990}$$

$$\frac{6\sum(x-y)^2}{990} = 1-0.8 = 0.2$$

$$6\sum(x-y)^2 = 990 \times 0.2$$

$$= 198$$

$$\therefore \sum(x-y)^2 = 33$$

Corrected $\sum(x-y)^2 = 33 - 5^2 + 8^2 = 72$

Now, after correction $\rho_{xy} = 1- \frac{6\times72}{10(10^2-1)}$

$$= 1- \frac{432}{990}$$

$$= 1-0.486$$

= 0.564

The correct coefficient of rank correlation is 0.546.

Exercises:

1. Ten students got the following percentage of marks in two subjects.

Economics	78	65	36	98	25	75	82	90	62	39
Statistics	84	53	51	91	60	68	62	86	58	47

2. The following table shows how 10 students were ranked according to their achievements in the laboratory and lecture portions of a biology course. Find the coefficient of rank correlation.

Laboratory	8	3	9	2	7	10	4	6	1	5
Lecture	9	5	10	1	8	7	3	4	2	6

REGRESSION:

If there is a functional relationship between the two variable x_i and $y_i$the points in the scatter diagram will cluster around some curve called a line of regression. If the curve is a straight line it is called a line of regression between the two variables.

Definition:

It we fit a straight line by the principle of least squares to the points of the scatter diagram in such a way that the sum of the squares of the distance parallel to the y-axis from the points to the line is minimized we obtain a line of best fit for the data and its is called the regression line of y and x.

Similarly we can define the regression line of $x\ on\ y$.

Theorem:

The equation of the regression line of $y\ on\ x$ is given by $y - \bar{y} = \gamma \frac{\sigma_x}{\sigma_y} (x - \bar{x})$

Proof:

Let $y = ax + b$ be the line of regression of on x.

According to the principle of least square the constants a and b are to be determined in such a way that S= $\sum[\, y_i - (ax_i + b)]\,^2 is\ minimum$

$\frac{\partial s}{\partial a} = 0 \Rightarrow -2\ \sum[(y_i - (ax_i + b)]x_i = 0$

$\Rightarrow \sum x_i y_i = a \sum x_i^2 + b \sum x_i$ ………(1)

$\frac{\partial s}{\partial b} = 0 \Rightarrow -2\ [\sum(y_i - (ax_i + b)] = 0$

$\Rightarrow \sum y_i = a \sum x_i + nb$ … … … … … (2)

Equations(1) and(2) are called normal equations.

From (2) we obtain $\bar{y} = a\bar{x} + b$………(3)

∴The line of regression passes through the point $(\bar{x}, \bar{y})$.

Now, shifting the origin to this point $(\bar{x}, \bar{y})$ by means of the transformation

$X_i = x_i - \bar{x}$ and $Y_i = y_i - \bar{y}$.

We obtain $\sum x_i = 0 = \sum y_i$ and the equation of the line of regression becomes

$$y = ax \quad \text{…………(4)}$$

Corresponding to this $line\ y = ax$ the constant a can be determined from the normal equation .

a$\sum X_i^2 = \sum x_i y_i$

a= $\frac{\sum X_i Y_i}{\sum X_i^2}$

$= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})2}$

$= \frac{\gamma \sigma_x \sigma_y}{\sigma_x^2}$

$= \gamma \frac{\sigma_x}{\sigma_y}$

The required regression line (4) becomes $Y = \left(\gamma \frac{\sigma_y}{\sigma_x}\right) X$

$\therefore y - \bar{y} \ = \ \gamma \frac{\sigma_y}{\sigma_x} \ (x - \bar{x}).$

Definition:

The scope of the regression line of y on x is called the regression coefficient of y on x and it is denoted by b_{yx}.

Hence $b_{yx} = \gamma \frac{\sigma_y}{\sigma_x}$

The regression coefficient of x on y is given by

$b_{xy} \ = \ \gamma \frac{\sigma_x}{\sigma_y}$.

Theorem:

Arithmetic mean of the regression coefficients is greater than or equal to the correlation coefficient.

Proof:

Letb_{xy} and b_{yx} be the regression coefficients.

we have to prove ½ $(b_{xy} \ + \ b_{yx}) \geq \gamma$

Now, ½ $(b_{xy} \ + \ b_{yx})\ ) \geq \gamma$

$$\Leftrightarrow b_{yx} + b_{xy} \ \geq 2\gamma$$

$$\Leftrightarrow \gamma \frac{\sigma_y}{\sigma_x} + \gamma \frac{\sigma_x}{\sigma_y} \geq 2\gamma$$

$$\Leftrightarrow \sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y \geq 0$$

$$(\sigma_x - \sigma_y)^2 \geq 0.$$

This is always true.

Hence the theorem.

Theorem:

Regression coefficient are independent of the change of origin but dependent on change of scale.

Proof:

Let $u_i = \frac{x_i - A}{h}$ and $v_i = \frac{y_i - B}{k}$

Let $x_i = A + hu_i$ and $y_i = B + kv_i$

We know that, $\sigma_x = h\sigma_u, \sigma_y = k\sigma_v \; and \; \gamma_{xy} = \gamma_{uv}$

Now $b_{yx} = \gamma_{xy} \frac{\sigma_y}{\sigma_x}$

$$= \gamma_{uv} \left(\frac{k\sigma_v}{h\sigma_u}\right)$$

$$= \frac{k}{h} \; b_{uv} \ldots \ldots \ldots \ldots \ldots \ldots . (1)$$

similarly $b_{xy} = (\frac{h}{k}) \; b_{uv}$ ……………(2)

From (1) and (2) ⇒ b_{yx} and b_{xy} depend upon the scales h and k, but not on the origins A and B .

Hence the theorem.

Theorem:

The angle between two regression line is given by $\theta = \tan^{-1}\left[\left(\frac{1-\gamma^2}{\gamma}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right]$

Proof:

The equations of lines of regression of y on x and x on y respectively are

$y - \bar{y} = \gamma \frac{\sigma_y}{\sigma_x} (x - \bar{x})$…………………………(1)

$$x - \bar{x} = \gamma \frac{\sigma_x}{\sigma_y} \; (y - \bar{y}) \ldots \ldots \ldots \ldots . (2)$$

(2) can also be written as

$y - \bar{y} = \frac{1}{\gamma} \frac{\sigma_y}{\sigma_x} \; (x - \bar{x})$ … … … ….(3)

Slopes of the two lines (1) and (2) are $\gamma \frac{\sigma_y}{x}$ and$\frac{\sigma_y}{\gamma\sigma_x}$.

Let θ be the acute angle between the two lines of regression.

$$\therefore tan\theta = \frac{\gamma \frac{\sigma_y}{\sigma_x} - \frac{\sigma_y}{\gamma \sigma_x}}{1 + \left(\gamma \frac{\sigma_y}{\sigma_x}\right)\left(\frac{y}{\gamma \sigma_x}\right)}$$

$$= \frac{\gamma^2 - 1}{\gamma}\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)$$

$$= \frac{1-\gamma^2}{\gamma}\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)$$

(since $\gamma^2 \leq 1$ and θ is $acute$).

$$\therefore \theta = \tan^{-1}\left[\left(\frac{1-\gamma^2}{\gamma}\right)\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)\right]$$

Problem: 10

The following data relate to the marks of 10 students in the internal test and the university examination for the maximum of 50 in each.

Internal marks	25	28	30	32	35	36	38	39	42	45
University marks	20	26	29	30	25	18	26	35	35	46

i) Obtain the two regression equations and determine.

ii) The most likely internal mark for the university mark of 25.

iii) the most likely university mark for the internal mark of 30.

Solution:

(i) Let the marks of internal test and university examination be denoted by x and y respectively.

We have $\bar{x} = \frac{1}{10}\sum x_i = 35$ and $\bar{y} = \frac{1}{10}\sum y_i = 29$.

For the calculation of regression we have the following table.

x_i	x_i-35	$(x_i - 35)^2$	y_i	y_i-29	$(y_i - 29)^2$	(x_i-35)$(y_i - 29)$
25	-10	100	20	-9	81	90
28	-7	49	26	-3	9	21
30	-5	25	29	0	0	0
32	-3	9	30	1	1	-3
35	0	0	25	-4	16	0
36	1	1	18	-11	121	-11
38	3	9	26	-3	9	-9
39	4	16	35	6	36	24
42	7	49	35	6	36	42
45	10	100	46	17	289	170
Total	0	358	-	0	598	324

$$\sigma_x^2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{1}{10}\sum(x_i - 35)^2 = 35.8$$

$$\sigma_y^2 = \frac{\sum(y_i - \bar{y})^2}{n} = \frac{1}{10}\sum(y_i - 29)^2 = 59.8$$

$$\sigma_x = 5.98 \ \ and \ \sigma_y = 7.73$$

$$\therefore \gamma = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y}$$

$$= \frac{324}{10 \times 5.98 \times 7.73}$$

$$= \frac{324}{462.254}$$

= 0.7 (approximately)

Now the regression of y on x is $y - \bar{y} = \gamma \frac{\sigma_y}{\sigma_x} (x - \bar{x})$

$$\therefore \gamma \frac{\sigma_y}{\sigma_x} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x^2}$$

$= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

$= \frac{324}{358} = 0.905$

Similarly, $\gamma \frac{\sigma_x}{\sigma_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2} = \frac{324}{598} = 0.542$

The regression line $of\ y\ on\ x$ is $y = 29\ =\ 0.905\ (x - 35)$

(ie), $y = 0.905\ x - 2.675$ ……..(1)

The regression line $of\ x\ on\ y\ is\ x - 35\ =\ 0.542\ (y - 29)$

(ie), $x = 0.542y\ +\ 19.282$…………..(2)

(1) and (2) are the required regression equations.

ii) the most likely internal mark for the university mark of 25 is got from the regression equation $of\ x\ on\ y$ by putting $y\ =\ 25$

(2)⇒ $x = 0.542\ \times 25 + 19.282 = 32.83$

iii) The most likely university mark for the internal mark of 30 is got from the regression equation $of\ y\ on\ x$ by putting $x\ =\ 30$

(1)⇒ $y = 0.905\ \times 30 - 2.675 = 24.475$

Problem: 11

The two variable x and y have the regression lines 3x+2y-26 = 0 and 6x+y-31=0. Find

i)The mean values of x and y
ii)The correlation coefficient between x and y
iii)The variance of y if the variance of x is 25

Solution:

(i) Since the two lines of regression pass through ($\bar{x}$,$\bar{y}$)

we have 3$\bar{x}$+$\bar{y}$=26……(1)

6$\bar{x}$+$\bar{y}$=31……(2)

Solving(1) and(2) we get $\bar{x} = 4\ and\ \bar{y} = 67$

(ii) As in the previous problem we can prove that $y = \frac{-3}{2}x + 13$ *and* $x = \frac{-1}{6}y + \frac{31}{6}$ represent the regression lines of y on x and x on y by respectively.

Hence we get the regression coefficients $as\ b_{yx} = \frac{-3}{2}$

$$and\ b_{xy} = -1/6$$

Now,

$$\gamma^2 = \left(\frac{-3}{2}\right) \times \left(\frac{-1}{6}\right) = ¼$$

$$\gamma = \pm\frac{1}{2}$$

Since both the regression coefficients are negative we take $\gamma = -\frac{1}{2}$

iii) Given $\sigma_x = 5$

We have $b_{yx} = \gamma \frac{\sigma_y}{\sigma_x}$

$$\therefore \frac{-3}{2} = \left(\frac{-1}{2}\right)\left(\frac{\sigma_y}{5}\right)$$

$$\sigma_y = 15\ .$$

Problem: 12

If θ is the acute angle between the two regression lines.
Show that $\theta \leq 1 - \gamma^2$

Solution:

We know that if θ is the acute angle between the two regression on lines we have,

$\tan\theta = \left(\frac{1-\gamma^2}{\gamma}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$ ………..(1)

We claim that $\sigma_x^2 + \sigma_y^2 \geq 2\sigma_x\sigma_y$.

Suppose not, then $\sigma_x^2 + \sigma_y^2 < 2\sigma_x\sigma_y$

(i.e.) $\sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y < 0$

$\left(\sigma_x - \sigma_y\right)^2 < 0$.This is impossible.

$$\text{Hence } \sigma_x^2 + \sigma_y^2 \geq 2\sigma_x\sigma_y$$

$$\therefore \frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2} \leq \frac{1}{2}.$$

$$(1) \Rightarrow \tan\theta \leq \left(\frac{1-\gamma^2}{\gamma}\right)\left(\frac{1}{2}\right)$$

$$\therefore \tan\theta \leq \left(\frac{1-\gamma^2}{2\gamma}\right)$$

$$\text{Hence } \sin\theta \leq \left(\frac{1-\gamma^2}{1+\gamma^2}\right)$$

$$\sin\theta \leq 1-\gamma^2.$$

Exercise:

1.calculate the coefficient of correlation of correlation and obtain the lines of regression for the following data.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

Correlation coefficient for a bivariate frequency distribution:

The correlation coefficient between x and y is given by $\gamma_{xy} = \frac{\cos(x,y)}{\sigma_x\sigma_y}$

$$\therefore \gamma_{xy} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}\, x_i y_j - \frac{1}{N}\left(\sum_{i=1}^{n} g_i x_i\right)\left(\sum_{j=1}^{m} f_j y_j\right)}{\sqrt{\sum_{i=1}^{n} g_i x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{n} g_i x_i\right)^2} \times \sqrt{\sum_{j=1}^{m} f_j y_j^2 - \frac{1}{N}\left(\sum_{j=1}^{m} f_j y_j\right)}^2}$$

Note: Since correlation coefficient is independent of origin and scale if x and y are transformed to u and v by the formula u = $\frac{x-A}{h}$ and v = $\frac{y-B}{k}$ then we $have\ \gamma_{xy} = \gamma_{uv}$.

Problem: 13

Find the correlation coefficient between x and y from the following table:

x	5	10	15	20
y				
4	2	4	5	4
6	5	3	6	2
8	3	8	2	3

Solution:

X		**x1**	**x2**	**x3**	**x4**	**Total**
Y		**5**	**10**	**15**	**20**	
y1	**4**	**2**	**4**	**5**	**4**	**F1=15**
y2	**6**	**5**	**3**	**6**	**2**	**F2=16**
y3	**8**	**3**	**8**	**2**	**3**	**F3=16**
Total		**g1=10**	**g2=15**	**g3=13**	**g4=9**	**N=47**

Correlation coefficient between x and y is given by

$$\therefore\ \gamma_{xy} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}\, x_i y_j - \frac{1}{N}\left(\sum_{i=1}^{n} g_i x_i\right)\left(\sum_{j=1}^{m} f_j y_j\right)}{\sqrt{\sum_{i=1}^{n} g_i x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{n} g_i x_i\right)^2} \times \sqrt{\sum_{j=1}^{m} f_j y_j^2 - \frac{1}{N}\left(\sum_{j=1}^{m} f_j y_j\right)^2}}$$

Where i=1,2,3,4 and j=1,2,3.

$$\sum g_i x_i = 50 + 150 + 195 + 180 = 575$$

$$\sum f_j y_j = 60 + 96 + 128 = 284$$

$$\sum g_i x_i^2 = 250 + 1500 + 2925 + 3600 = 8275$$

$$\sum f_j y_j^2 = 240 + 576 + 1024 = 1840$$

$\sum\sum f_{ij}\, x_i y_j = (40 + 160 + 300 + 320) + (150 + 180 + 540 + 240) + (120 +$ *640+240+480=3410*

$$\gamma_{xy} = \frac{3410 - \frac{1}{47}(575 \times 284)}{\sqrt{8275 - \frac{1}{47}(575)^2} \times \sqrt{1840 - \frac{1}{47}(284)^2}}$$

$$= \frac{3410 \times 47 - (575 \times 284)}{\sqrt{8275 \times 47 - 575^2} \times \sqrt{1840 \times 47 - 284^2}}$$

$$= \frac{160270 - 163300}{\sqrt{388925 - 330625} \times \sqrt{86480 - 80656}}$$

$$= \frac{-3030}{\sqrt{58300} \times \sqrt{5824}} = \frac{-3030}{241.5 \times 76.3}$$

$$= \frac{-3030}{18426.5}$$

= -0.16

Problem: 14

Find the correlation coefficient between the heights and weight of 100 students which are distributed as follows.

Height in c.m	Weight in Kgs					Total
	30-40	40-50	50-60	60-70	70-80	
150-155	1	3	7	5	2	18
155-160	2	4	10	7	4	27
160-165	1	5	12	10	7	35
165-170	-	3	8	6	3	20
Total	4	15	37	28	16	100

Solution:

Let x_i denote the mid value of the classes of weights and y_j denote the mid value of the classes of heights.

$$Let\ u_i = \frac{x_i - 55}{10}\ and\ v_j = \frac{y_j - 157.5}{5}$$

Then the 2 -way frequency table is given below.

	35	45	55	65	75	f_j	v_j	$f_j v_j$	$f_j v_j^2$	$f_{ij} u_i v_j$
152.5	(2) 1	(3) 3	(0) 7	(-5) 5	(-4) 2	18	-1	-18	18	(-4)
157.5	(0) 2	(0) -4	(0) 10	(0) 7	(0) 4	27	0	0	0	(0)
162.5	(-2) 1	(-5) 5	(0) 12	(10) 10	(14) 7	35	1	35	35	(17)
167.5	-	(-6) 3	(0) 8	(12) 6	(12) 3	20	2	40	80	(18)
g_i	4	15	37	28	16	100	-	57	133	(31)
u_i	-2	-1	0	1	2	-				
$g_i u_i$	-8	-15	0	28	32	37				
$g_i u_i^2$	16	15	0	28	64	123				
$f_{ij} u_i v_j$	(0)	(-8)	(0)	(17)	(22)	(31)				

$$\gamma_{xy} = \gamma_{uv} \quad = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij}\, u_i v_j - \frac{1}{N}\left(\sum_{i=1}^{n} g_i u_i\right)\left(\sum_{j=1}^{m} f_j v_j\right)}{\sqrt{\sum_{i=1}^{n} g_i u_i^2 - \frac{1}{N}\left(\sum_{i=1}^{n} g_i u_i\right)^2} \times \sqrt{\sum_{j=1}^{m} f_j v_j^2 - \frac{1}{N}\left(\sum_{j=1}^{m} f_j v_j\right)^2}}$$

$$= \frac{31 - \frac{1}{100}(37 \times 57)}{\sqrt{123 - \frac{1}{100} 37^2} \times \sqrt{133 - \frac{1}{100} 57^2}}$$

$$= \frac{3100 - 37 \times 57}{\sqrt{12300 - 37^2} \times \sqrt{13300 - 57^2}}$$

$$= \frac{991}{104.5 \times 100.25}$$

$$= 0.09.$$

UNIT II: PROBABILITY

Probability - Definition - Applications of Addition and multiplication, theorems - Conditional, Probability - Mathematical Expectations - Moment generating function - Special Distributions, (Binomial Distributions, Poisson Distribution, Normal Distribution - Properties).

PROBABILITY

Introduction:

In this chapter we develop the mathematical theory of probability and introduce the concept of random variables which form the basis for various types of theoretical distributions.

Definition:

An experiment is defined as an action which we conceive and do or intend to do.

Each experiment ends with an outcome. For example, a research student in “statistics” when undertaking a pre election sample survey.

An experiment is called a random experiment if, when repeated under the same conditions, it is such that the outcome cannot be predicted with certainty but all possible outcomes can be determined prior to the performance of the experiment.

Each performance of the random experiment is called trail. The collection of all possible outcomes of a random experiment is called the sample space **S**. The elements of sample space are called sample points.

Example:

When two cons are tossed at a time the outcome is an ordered pair (H,H) or (H, T) or (T,H) or (T,T). Hence for the random experiment of tossing two coins, sample space S={(H,H), (H,T), (T,H), (T,T)}.

Definition:

Any subset A of a sample space S is called an event.

The event S is called a sure event and the event φ is called an impossible event.

Definition:

Let S be a sample space associated with an experiment. Let A be event suppose an experiment is repeated N times and suppose the event A happens f times. Then f/N is called the relative frequency of the event A. clearly $0 \leq f/N \leq 1$.

The following desirable properties to be satisfied by a probability function P.

a) P(A) $\geq$0 for all events
b) P(A) $\leq$ 1for all events
c) P(S) = 1
d) If A and B are disjoint events

P(AUB) = P(A) +P(B)

Example:

When two cons are tossed at a time the outcome is an ordered pair (H,H) or

(H, T) or (T,H) or (T,T). Hence for the random experiment of tossing two coins, sample space S={(H,H), (H,T), (T,H), (T,T)}.

Definition:

Any subset A of a sample space S is called an event.

The event S is called a sure event and the event φ is called an impossible event.

Definition:

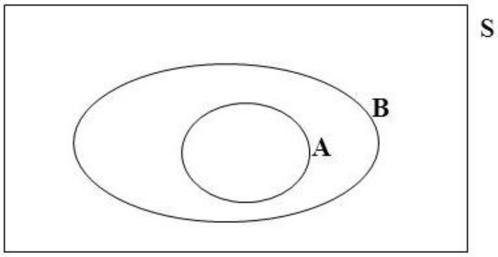
Let S be a sample space associated with an experiment. Let A be event suppose an experiment is repeated N times and suppose the event A happens f times. Then f/N is called the relative frequency of the event A. clearly $0 \leq f/N \leq 1$.

The following desirable properties to be satisfied by a probability function P.

a) P(A) $\leq$ 1for all events
b) P(S) = 1
c) If A and B are disjoint events

P(AUB) = P(A) +P(B)

Proof:



Let $A \subseteq B$.

Then B=A∪($\bar{A}$ ∩B).A and $\bar{A} \cap B$ disjoint events of S.

Hence $P(B) = P(A) + P(\bar{A} \cap B)$

But $P(\bar{A} \cap B) \geq P(A)$

Hence $P(B) \geq (PA)$

Corollary:

For each A$\subseteq$S, 0≤P(A) ≤1.

Proof:

∅ $\subseteq$A⊆S

Hence P(∅)≤P(A) ≤P(S)

Hence 0≤P(A) ≤1.

Theorem:

If A and B are any two events of a sample space S then

P(AUB)=P(A) + P(B) –P(A∩B).

Proof:

$A \cap B \ = \ A \cup (\bar{A} \cap B)$ and A and $\bar{A} \cap B$ are disjoint sets.

∴ P(A∪B) = P(A) + P $(\bar{A} \cap B)$ ……………..(1)

Now, B = (A∩B) ∪($\bar{A}$ ∩B) and A∩B and $\bar{A}$ ∩B are disjoint sets.

∴ P(B) = P(A∩B) + P($\bar{A}$ ∩B)

∴ P($\bar{A}$ ∩B) = P(B) - P(A∩B)………………(2)

(1)⇒P(A∪B) = P(A) + P(B) –P(A∩B) [(2) in (1)]

Example:

Let S = {(H,H), (H,T) (T,H),(T,T)}.

Let us assign the probability of ¼ to each element of the sample space S.

Let A={(H,H),(H,T)}; B={(H,H), (T,H)}

A∪B = { (H,H) (H,T), (T,H)}

A∩B = {(H,H)}

P(A) = $\frac{1}{2}$, P(B) = $\frac{1}{2}$, P(A∪B) = $\frac{3}{4}$

And P(A∩ $B) = \frac{1}{4}$

We have, P(A∪ B) = P(A) + P(B) – P(A∩B)

$= \frac{1}{2} + \frac{1}{2} - \frac{1}{4}$

$= \frac{2+2-1}{4} = \frac{3}{4}$

Hence it is verified.

Example:

Let S= {(i,j) /i,j ∈N, 1≤i≤ $6, 1$ ≤j≤6} be the sample space of the random experiment of throwing two dice. We assign the uniform probably of 1/36 to each of the 36 sample points in the sample space S.

Let A denote the event of getting 1 in the second die.

Then A = {(1,1) (2,1), (3,1), (4,1), (5,1) (6,1) }

P(A) = $\frac{6}{36} = \frac{1}{6}$

Let B = {(1,2), (2,2) (3,2)}

Then P(B) = $\frac{3}{36} = \frac{1}{12}$

A∩B = ∅, P(A∩ $B) = 0$

P(A∪ $B) = P(A) + P(B) - P(A \cap B)$

$$= \frac{6}{36} + \frac{3}{36} = \frac{9}{36} = \frac{1}{4}$$

Definition:

Let S denote a sample space associated with an experiment. Let $A_1, A_2, A_3 \ldots . A_{n, \ldots\ldots}$ be a sequence of Subsets of S.

If $A_i \cap A_j$ = ∅ for all i, j with i ≠ j then the sequence of subsets is said to be mutually disjoint.

If $\bigcup_{n=1}^{\infty} A_n$ then the sequence of events is said to be exhaustive.

Example:

Let S= $\{(i,j) \,/\, i, j \in N, 1 \leq i \leq 6, 1 \leq j \leq 6\}$

Let A be an event of getting the sum i + j and odd number as B be an event of getting the sum as an even number.

Clearly $A \cap B = \emptyset$ and $A \cup B = S$ Hence A and B are mutually exclusive and exhaustive events.

Theorem:

For any two events A and B, $P(\bar{A} \cap B) = P(B) - P(A \cap B)$

Proof:

Let $\bar{A} \cap B$ and $A \cap B$ are disjoint events and

$$(\bar{A} \cap B) \cup (\bar{A} \cap B) = B$$

$P(B) = P[(A \cap B) \cup (\bar{A} \cap B)]$

$P(B) = P(A \cap B) + P(\bar{A} \cap B)$

$\therefore P(\bar{A} \cap B) = P(B) - P(A \cap B)$

Remark:

Similarly we shall get $P(A \cap \bar{B}) = P(A) - P(A \cap B)$

Theorem:

If $B \subset A$, then (i) $P(A \cap \bar{B}) = P(A) - P(B)$

(ii) $P(B) \leq P(A)$

Proof:

i).When B⊂A, B and (A∩ $\bar{B}$) are mutually exclusive events and their union is A.

∴ P(A) = P[B∪ $(A \cap \bar{B})$]

P(A) = P(B) + P(A∩ $\bar{B}$)

∴ P(A∩ $\bar{B}$) = P(A) – P(B).

ii) Using axiom (i)

P(A∩ $\bar{B}$)≥ 0 ⇒ $P(A) - P(B) \geq 0$.

$$P(B) \leq P(A).$$

Corollary:

If (A∩B) ⊂ A and (A∩B) ⊂(B) then$P(A \cap B) \leq P(A)$ and $P(A \cap B) \leq P(B)$

Law of addition of probabilities:

Statement:

If A and B are any two events (subsets of a sample space S) and are not disjoint then, P(A∪B) = P(A) + P(B) – P(A∩B)

Proof:

We have, A∪ $B = A \cup (\bar{A} \cap B)$

Since A and $(\bar{A} \cap B)$ $are\ disjoint$.

P(A$\cup B) = P(A) + P(\bar{A} \cap B)$

=P(A) + [P($\bar{A}$ $\cap$B) + P(A$\cap B) - P\ (A \cap B)$]

= P(A) + P[$(\bar{A} \cap B)\ \cup (A \cap B)] - P\ (A \cap B)$

$[\therefore (\bar{A} \cap B)\ and\ (A \cap B) are\ disjoint]$

$\therefore P(A \cup B)$ = P(A) + P(B) – P(A$\cap B$)

Theorem: (Extension of General law of addition of probabilities]

For n events $A_1, A_2 \ldots \ldots \ldots, A_n$, wehave

P($\bigcup_{i=1}^{\infty} A_i$) = $\sum_{i=1}^{n} p(A_i) - \sum\sum P\big(A_i \cap A_j\big) + \cdots ..$

$+ (-1)^{n-1} P(A_1 \cap A_2 \cap \ldots . \cap A_n)$for all $1 \leq i \leq j \leq n$.

Proof:

For two events $A_1\ and\ A_2$

We have,

$$P\ (A_1 \cup A_2)\ =\ P(A_1)\ +\ (P(A_2) - P(A_1 \cap A_2)$$

It is true for n=2

Suppose that, it is true for n=r

Then, P $(\bigcup_{i=1}^{r} A_i) = \sum_{i=1}^{r} P(A_i) - \sum\sum P\big(A_i \cap A_j\big) + \cdots .$

$+ (-1)^{r-1} P(A_1 \cap A_2 \cap \ldots . \cap A_r)$

for all $1 \leq i \leq j \leq r$.

Now ,

$$P\,(\bigcup_{i=1}^{r+1} A_i) = P\,[(\bigcup_{i=1}^{r} A_i) \cup A_{r+1}]$$

$$=P(\bigcup_{i=1}^{r} A_i) + P(A_{r+1}) - P[(\bigcup_{i=1}^{r} A_i) \cap A_{r+1}]$$

$$= P(\bigcup_{i=1}^{r} A_i) + P(A_{r+1}) - P[\bigcup_{i=1}^{r}(A_i \cap A_{r+1})]$$

(Distributive Law)

$$=\sum_{i=1}^{r} P\,(A_i) - \sum\sum P\left(A_i \cap A_j\right)+\ldots+(-1)^{r-1}P(A_1 \cap A_2 \cap \ldots \cap A_r) + P(A_{r+1}) - P[\bigcup_{i=1}^{r} A_i \cap A_{r+1}]$$

For all $1\leq i \leq j \leq r$

$$=\sum_{i=1}^{r+1} P(A_i)-\sum\sum P\left(A_i \cap A_j\right) + \cdots + (-1)^{r-1}P$$

$$(A_1 \cap A_2 \cap \ldots \cap A_r) - \left[\sum_{i=1}^{r} P\,(A_i \cap A_{r+1}) - \sum\sum P\left(A_i \cap A_j \cap A_{r+1}\right) + (-1)^{r-1}P\,(A_1 \cap A_2 \cap \ldots \cap A_r \cap A_{r+1})\right]$$

For all $1\leq i \leq j \leq r$

$$=\sum_{i=1}^{r+1} P(A_i) - \left[\sum\sum P\left(A_i \cap A_j\right) + \sum_{i=1}^{r} P\,(A_i \cap A_{r+1})\right] + (-1)^r P\,(A_1 \cap A_2 \cap \ldots \cap A_r \cap A_{r+1})$$

For all $1\leq i \leq j \leq r$

$$=\sum_{i=1}^{r+1} P(A_i) - \sum\sum P\left(A_i \cap A_j\right) + \cdots + (-1)^r P\,(A_1 \cap A_2 \cap \ldots \cap A_r \cap A_{r+1})$$

For all $1\leq i \leq j \leq r+1$

Hence it is true for n=r and it is also true for n=r+1.

Hence by the principle of mathematical induction, it is true for all positive integral values of n.

Theorem: [Boole's inequality]

For n events $A_1, A_2, \ldots A_n$ we have

a) $P\,(\bigcap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$

b) $P\,(\bigcup_{i=1}^{n} A_i\,) \leq \sum_{i=1}^{n} P(A_i)$

Proof:

The result is now prove by mathematical introduction

P($A_1 \cup A_2$) = P(A_1) + P(A_2) – P ($A_1 \cap A_2$) ≤1

P($A_1 \cup A_2$) ≥P(A_1) + P((A_2) – 1

Hence it is true for n=2

Suppose that it is true for n=r such that

P($\bigcup_{i=1}^{r} A_i$)≥ $\sum_{i=1}^{r} P(A_i) - (r-1)$

Then P ($\bigcap_{i=1}^{r+1} A_i$) = P ($\bigcap_{i=1}^{r} A_i \cap A_{r+1}$)

≥ P ($\bigcap_{i=1}^{r} A_i$)+P(A_{r+1})-1

$$\geq \sum_{i=1}^{r} P(A_i) - (r-1) + P(A_{r+1}) - 1$$

⇒P($\bigcap_{i=1}^{r+1} Ai$) ≥ $\sum_{i=1}^{r+1} P(A_i) - r$

∴ It is true for n=r+1 also

b) Let $\bar{A}_1\bar{A}_2$……..$\bar{A}_n$ be the events

We get

P($\bar{A}_1 \cap \bar{A}_2$ ….∩ $\bar{A}$n) ≥[P($\bar{A}_1$) + P($\bar{A}_2$) + ……..+P ($\bar{A}_n$)] – (n-1)

= [1-P(A_1) + [1-P(A_2)] + …….+ [1-P(A_n)] – (n-1)

= 1-P(A_1) – P(A_2) - …….-P (A_n)

⇒ P(A_1) + P(A_2) +…..P(A_n)≥1-P ($\bar{A}_1 \cap \bar{A}_2 \cap$…..∩ $\bar{A}_n$)

= 1-P($\overline{A_1 \cup A_2 \cup \ldots \cup A_n)}$

= P($A_1 \cup A_2 \cup \ldots \cup A_n$)

⇒ P($A_1 \cup A_2 \cup \ldots \cup A_n$)≤P($A_1$) + P($A_2$) +…..P($A_n$)

Theorem:

For n events $A_1, A_2, \ldots A_{n.}$

we have P[$\bigcup_{i=1}^{n} A_i$] ≥ $\sum_{i=1}^{n} P(A_i) - \ \sum\sum P\left(A_i \cap A_j\right), for\ all\ \ 1 \le i \le j \le n$

Proof:

We shall prove this theorem by the method of induction.

We have,

P(A_1+A_2+A_3) = P(A_1) +P(A_2) + P(A_3)- [P($A_1 \cap A_2$) + P($A_2 \cap A_3$) + P($A_2 \cap\ A_1$)] + P($A_1 \cap A_2 \cap A_3$)

P($\bigcup_{i=1}^{3}$) ≥ $\sum_{i=1}^{3} P(A_i) - \ \sum\sum_{1 \le i \le j \le 3} P\left(A_i \cap A_j\right)$

Thus the result is true for n=3

Suppose that the result is true for n=r

Then P($\bigcup_{i=1}^{r} A_i$) ≥ $\sum_{i=r} P(A_i)$ - $\sum\sum_{1 \le i \le j \le r} P\left(A_i \cap A_j\right)$

Now,

P($\bigcup_{i=1}^{r+1} A_i$) =P[$\bigcup_{i=1}^{r} A_i \cup A_{r+1}$]

= P($\bigcup_{i=1}^{r}(A_i) +\ P(A_{r+1}) -\ P\ [(\bigcup_{i=1}^{r} A_i) \cap A_{r+1})]$

= P($\bigcup_{i=1}^{r}(A_i) +\ P(A_{r+1}) -\ P\ [(\bigcup_{i=1}^{r}(A_i \cap A_{r+1})]$

≥ $\left[\sum_{i=1}^{r} P\ (A_i) - \sum\sum_{1 \le i \le j < r} P\ \left(A_i \cap A_j\right)\right] + P(A_{r+1}) -\ P[\bigcup_{i=1}^{r}(A_i \cap A_{r+1})]$

From Boole’s inequality we get

P[$\bigcup_{i=1}^{r}(A_i \cap A_{r+1})$] $\le \sum_{i=1}^{r} P(A_i \cap A_{r+1})$

⇒ -P [$\bigcup_{i=1}^{r}(A_i \cap A_{r+1})$] $\ge -\sum_{i=1}^{r} P(A_i \cap A_{r+1})$

$\therefore\ P(\bigcup_{i=1}^{r+1} A_i) \ge \sum_{i=1}^{r+1} P\ (A_i) - \sum\sum_{1 \le i \le j \le r} P\ \left(A_i \cap A_j\right) - \sum_{i=1}^{r} P(A_i \cap A_{r+1})$

$P(\bigcup_{i=1}^{r+1} A_i) \ge \sum_{i=1}^{r+1} P\ (A_i) - \sum\sum_{1 \le i \le j \le r+1} P\ \left(A_i \cap A_j\right)$

Hence, if the theorem is true for n=r, it is also true for n=r+1

Hence by mathematical induction,

the result is true for all positive integral values of n.

Multiplication law of probability and conditional probability:

Theorem:

For two events A and B

P(A∩B) = P(A). P(B│A), P(A)>0

= P(B).P(A│B), P(B)>0

Where P(B│A) represents the conditional probability of occurrence of B when the event A has already happened and P(A│B) is the conditional probability of happening of A, given that B has already happened.

Proof:

Suppose the sample space contains N occurrences of which n_A occurrences belong to the event A and n_B occurrences belong to the event B.

Let n_{AB} be the number of occurrences favorable to the compound event A∩B then, the unconditional probabilities are given by

P(A) = $\frac{n_A}{N}$, P(B) = $\frac{n_B}{N}$ and P$(A \cap B) = \frac{n_{AB}}{N}$

Now, the conditional probability P(A│B) refers to the sample space of n_B occurances, out of which n_{AB} occurrences pertain to the occurrence of A, when B has already happened.

P(A│B) = $\frac{n_{AB}}{n_B}$

Similarly P(B│A) = $\frac{n_{AB}}{n_A}$

Now, P(A ∩ B) = $\frac{\mathrm{n_{AB}}}{N}$

= $\frac{n_{AB}}{n_A} \cdot \frac{\mathrm{n_A}}{N}$

= P(B│A) P(A)

And P(A∩B) = $\frac{\mathrm{n_{AB}}}{N} = \frac{n_{AB}}{n_B} \cdot \frac{\mathrm{n_B}}{N}$

= P(A│B) P(B)

$\therefore$ P(B | A) = $\frac{P(A\cap B)}{P(A)}$ and P(A | B) = $\frac{P(A\cap B)}{P(B)}$

Thus the conditional probabilities P(B | A) and P(A | B) are defined iff P(A) ≠0 and P(B) ≠0 respectively.

Extension of multiplication law of probability :

Theorem:

For n events $A_1, A_2, \ldots A_{n.}$ we have P($A_1 \cap A_2 \cap \ldots \cap A_n$)= P($A_1$) P($A_2 \mid A_1$) P($A_3 \mid A_1 \cap A_2$)…P($A_n \mid A_1 \cap A_2 \cap \ldots\ldots \cap A_{n-1}$) …..

Where,

P($A_i \mid A_j \cap A_k \cap \ldots\ldots \cap A_i$) represents the conditional probability of the event A_i, given that the events A_j, A_k,…A_i have already happened.

Proof:

We have for three events A_1, A_2, and A_3

$$P(A_1 \cap A_2 \cap A_3) \quad = P(A_1 \cap \overbrace{A_2 \cap A_3})$$

$$= P(A_1)\, P(A_2 \cap A_3 \mid A_1)$$

$$= P(A_1)\, P(A_2 \mid A_1) P(A_3 \mid A_1 \cap A_2)$$

It is true for n=2 and n=3

Suppose that it is true for n=k.

So that,

$$P(A_1 \cap A_2 \cap \ldots . \cap A_k) = P(A_1)\, P(A_2 \mid A_1) P(A_3 \mid A_1 \cap A_2) \ldots$$

$$P(A_k \mid A_1 \cap A_2 \cap \ldots \cap A_{k-1})$$

Now,

$$P(\overbrace{A_1 \cap A_2 \cap \ldots \cap A_k} \cap A_{k+1}) = P(A_1 \cap A_2 \cap \ldots . \cap A_k)\, P(A_{k+1} \mid A_1 \cap A_2 \cap \ldots \cap A_k)$$

$$=P(A_1)\, P(A_2 \mid A_1) \ldots P(A_k \mid A_1 \cap A_2 \cap \ldots \cap A_{k-1} \times P(A_{k+1} \mid A_1 \cap A_2 \cap \ldots \cap A_k)$$

It is true for n=k+1 also

Since it is true for n=2 and n=3, by the principle of mathematical induction, it is true for all integral values of n.

Theorem:

For any three events A,B,C;P(A∪ $B|C) = P(A|C) + P(B|C) - \ P(A \cap B|C)$

Proof:

We have

P(A∪B) = P(A) + P(B) – P(A∩B)

⇒ P[(A∩C) ∪(B∩ C)] $= P(A \cap C) + \ P(B \cap C) - \ P(A \cap B \cap C)$

Dividing both sides by P(C) we get

$\frac{P(A\cap C)\cup(B\cap C)}{P(C)} = \frac{P(A\cap C)+ P(B\cap C)- P(A\cap B\cap C)}{P(C)}$,P(C) >0

$= \frac{P(A\cap C)}{P(C)} + \frac{P(B\cap C)}{P(C)} - \frac{P(A\cap B\cap C)}{P(C)}$

⇒ $\frac{P[(A\cup B)\cap C]}{P(C)}$ = P(A|C) + P(B|C) – P(A∩ $B)|C$)

⇒ P(A∪ $B|C) = P(A|C) + P(B|C) - P(A \cap B|C)$

Theorem:

For any three events A, B and C P(A ∩ $\bar{B}|C$) + P(A ∩ $B|C$)=P(A|C)

Proof:

P(A∩ $\bar{B}|C) + P\big(A \cap B|C\big) = \frac{P(A\cap\bar{B}\cap C)}{P(C)} + \frac{P(A\cap B\cap C)}{P(C)}$

$= \frac{P(A\cap\bar{B}\cap C)+ P(A\cap B\cap C)}{P(C)}$

$= \frac{P\ (A\cap C)}{P(C)}$ = P(A|C)

Theorem:

For a fixed B with P(B)>0, P(A|B) is a probability function

Proof:

i. P(A | B) = $\frac{P(A\cap B)}{P(B)} \geq 0$

ii. A(S | B) = $\frac{P(s\cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$

iii. If {An} is any finite or infinite sequence of disjoint events, then

$$P\left[\cup_n A_n \mid B\right] = \frac{P[(\cup_n A_n)\cap B]}{P(B)}$$

$$= \frac{P[\cup_n A_n B]}{P(B)}$$

$$= \frac{\sum_n P\ (A_n B)}{P(B)} = \sum_n \frac{P(A_n\ B)}{P(B)}$$

$$= \sum_n P\left(A_n \mid B\right)$$

Hence the theorem

Theorem:

For any three events A,B and C defined on the sample space S such that B⊂C and P(A>0),P(B | A) ≤P(C | A)

Proof:

$$P(C \mid A) = \frac{P(C\cap A)}{P(A)}$$

$$= \frac{P[(B\cap C\cap A)\cup(\bar{B}\cap C\cap A)]}{P(A)}$$

$$= \frac{P(B\cap C\cap A)}{P(A)} + \frac{P(\bar{B}\cap C\cap A)}{P(A)}$$

$$= P(B\cap C \mid A) + P(\bar{B} \cap C \mid A)$$

Now, B⊂C ⇒ B∩C=B

∴ P(C | A) = P (B | A) + P($\bar{B}$ ∩ C | A)

⇒ P(C | A)≥P(B | A)

Theorem:

If A and B are independent events then A and $\bar{B}$ are also independent events

Proof:

WE have P(A $\cap \bar{B}) = P(A) - (A \cap B)$

= P(A) – P(A) P(B) [A and B are independent]

= P(A) [1-P(B)]

=P(A) P(B)

⇒ A and $\bar{B}$ are independent events

Theorem:

If A and B are independent events then $\bar{A}$ and $\bar{B}$ are also independent events

Proof:

We are given P(A$\cap B) = P(A)P(B)$

Now P($\bar{A} \cap \bar{B}) = P(\overline{A \cup B})$

=1-P(A$\cup B$)

= 1-[P(A) + P(B) – P(A$\cap B$)]

=1-[P(A) + P(B) – P(A) P(B)]

=1-P(A)-P(B) + P(A) P(B)

=[1-P(B)] –P(A) [1-P(B)]

= [1-P(B)] [1-P(A)]

= [1-P(A)] [1-P(B)]

= P($\bar{A}$) P($\bar{B}$)

∴ $\bar{A}$ and $\bar{B}$ are independent events

Theorem:

If A,B,C are mutually independent events when A$\cup$ B *and* C *are also independent*.

Proof:

We are required to prove

$P[(A\cup B)\cap C] = P(A\cup B)P(C)$

L.H.S. = $P[A\cap C)\cup(B\cap C)]$ [*Distributive Law*]

$= P(A\cap C) + P(B\cap C) - P(A\cap B\cap C)$

$= P(A)\,P(C) + P(B)\,P(C) - P(A)\,P(B)\,P(C)$

[A, B, and C mutually independent)

$= P(C)\,[P(A) + P(B) - P(\cap B)]$

$= P(C)\,P(A\cup B) = R.H.S$

Hence $(A\cup B)$ *and C are independent.*

Theorem:

Prove that if A,B and C are random events in a sample space and if A,B,C are pair wise independent and A is independent of $(B\cup C)$, *then A, B and C are mutually independent*

Proof:

We are given,

$$\left.\begin{aligned}&P(A\cap B) = P(A)P(B)\\ &P(B\cap C) = P(B)P(C)\\ &P(A\cap C) = P(A)P(C)\\ &P(A\cap(B\cup C)] = P(A)\,P(B\cup C)\end{aligned}\right\}\text{(1)}$$

Now,

$P[A\cap(B\cup C)] = p[(A\cap B)\cup(A\cap C)$

$= P(A\cap B) + P(A\cap C) - P[(A\cap B)\cap(A\cap C)]$

$= P(A)\,P(B) + P(A)\,P(C) - P(A\cap BC)$………………..(2)

And

$P(A)P(B\cup C) = P(A)[P(B) + P(C) - P(B\cap C)]$

$= P(A)\,P(B) + P(A)\,P(C) - P(A)\,P(B\cap C)$…………(3)

From (2) and (3) on using (1) we get

$$P(A\cap B\cap C) = P(A)P(B\cap C)$$

$$=P(A)\,P(B)\,P(C)$$

Hence A, B, C are mutually independent

Theorem:

For any two events A and B, $P(A\cap B) \leq P(A) \leq P(A\cup B)P(A) + P(B)$

Proof:

We have

$A = (A\cap \bar{B})\ \cup (A\cap B)$

We have $P(A) = P\{\ (A\cap \bar{B}) \cup (A\cap B)$

$= P(A\cap \bar{B}) + P(A\cap B)$

But $P(A\cap \bar{B}) \geq 0$

$\therefore P(A) \geq P(A\cap B)$

Similarly $P(B) \geq P(A\cap B)$

$\Rightarrow P(B) - P(A\cap B) \geq= o$

Now $P(A\cup B) = P(A) + \ [P(B) - \ P(A\cap B)]$

$P(A\cup B)\ \geq P(A)$

$\Rightarrow P(A) \leq P(A\cup B)$

Also $P\ (A\cup B)\ \leq P(A) + \ P(B)$ From (2)

Hence from (1),(2) and (3) we get

$P(A\cap B) \leq P(A) \leq P(A\cup B) \leq P(A) + \ P(B)$

Example:

Two dice, one green and the other red, are thrown. Let A be the event that the sum of the points on the faces shown is odd and B the event of at least one ace (number ‘1’)

a. Describe the

i) complete sample space.

ii) events A,B, $\bar{B}$, A∩B, A∪B, and A∩ $\bar{B}$ and find their probabilities assuming all the 36 saple points have equal probabilities.

a. Find the probabilities of the events

i. $(\bar{A} \cup \bar{B})$

ii. $(\bar{A} \cap \bar{B})$

iii. $(A \cap \bar{B})$

iv. $(A \cap B)$

v. $\overline{A \cap B}$

vi. $\bar{A} \cup B$

vii. $(\overline{A \cup B})$

viii. $\bar{A} \cap (A \cup B)$

ix. $A \cup (\bar{A} \cap B)$

x. (A | B) and (B | A) and (A | $\bar{B}$) and ($\bar{B}$ | $\bar{A}$)

Solution:

a. The sample space S, consists of the 36 elementary events

{(1,1); (1,2); (1,3); (1,4); (1,5); (1,6)
(2,1); (2,2); (2,3); (2,4); (2,5); (2,6)
(3,1); (3,2); (3,3); (3,4); (3,5); (3,6)
(4,1); (4,2); (4,3); (4,4); (4,5); (4,6)
(5,1); (5,2); (5,3); (3,4); (3,5); (3,6)
(4,1); (4,2); (4,3); (4,4); (4,5); (4,6)
(5,1); (5,2); (5,3); (5,4); (5,5); (5,6)
(6,1); (6,2); (6,3); (6,4); (6,5); (6,6)}

for example, the ordered pair (4,5) refers to the elementary event that the green die shows 4 and the red die shows 5.

A= the event that the sum of the numbers shown by the two dice is odd.

= {(1,2); (2,1); (1,4); (2,3); (3,2); (4,1); (1,6); (2,5); (3,4); (4,3); (5,2); (6,1); (3,6); (4,5); (5,4); (6,3); (5,6); (6,5)}

Therefore,

P(A) = $\frac{n(A)}{n(S)}$ = $\frac{18}{36}$

B= the event that at least one face is 1

= {(1,1); (1,2); (1,3); (1,4); (1,5); (1,6); (2,1); (3,1) (4,1); (5,1); (6,1)}

therefore

P(B) = $\frac{n(B)}{n(S)}$

= $\frac{11}{36}$

$\bar{B}$ = the event that each of the face obtained is not an ace

= {(2,2); (2,3); (2,4); (2.5); (2,6); (3,2); (3,3); (3,4); (3,5); (3,6); (4,2); (4,3); (4,4); (4,5); (4,6); (5,2); (5,3); (5,4); (5,5); (5,6); (6,2); (6,3); (6,4); (6,5); (6,6)}

therefore

P($\bar{B}$) = $\frac{n(\bar{B})}{n(S)}$

=$\frac{25}{36}$

A∩B = the event that sum is odd and atleast one face is an ace.

= {(1,2); (2,1); (1,4); (4,1); (1,6); (6,1)}

$\therefore P(A \cap B) = \frac{n(A \cap B)}{n(s)} = \frac{6}{36} = \frac{1}{6}$

A∪B = {(1,2); (2,1); (1,4); (2,3); (3,2); (4,1); (1,6), (2,5); (3,4); (4,3); (5,2); (6,1); (3,6); (4,5); (5,4); (6,3); (5,6); (6,5); (1,1); (1,3); (1,5); (1,5) (3,1), (5,1)}

$\therefore P(A \cup B) = \frac{n(A \cup B)}{n(s)} = \frac{23}{36}$

A∩ $\bar{B}$ = {(2,3); (2,5); (3,2); (3,4); (3,6); (4,1); (4,5); (5,2); (5,4) (5,6), (6,3) (6,5)}

$$P(A \cap \bar{B}) = \frac{n(A \cap \bar{B})}{n(S)}$$

$$= \frac{12}{36}$$

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

b. i. P ($\bar{A} \cup \bar{B}$) $= P(\overline{A \cap B})$

= 1-P ($A \cap B$)

$= 1-\frac{1}{6} = \frac{5}{6}$

ii) P$(\bar{A} \cap \bar{B}) = P(\overline{A \cup B})$

$= 1\text{-P}(\text{A} \cup B)$

$= 1\text{- P(A)} – \text{P(B)} + \text{P}(\text{A} \cap B)$

$=1\text{-} \frac{18}{36} \text{-} \frac{11}{36} + \frac{6}{36}$

$= \frac{13}{36}$

iii. $P(A \cap \bar{B}) = P(A) - P(A \cap B)$

$= \frac{18}{36} \text{-} \frac{6}{36}$

$= \frac{12}{36}$

$= \frac{1}{3}$

iv) P$(\bar{A} \cap B) = P(B) - P(A \cap B)$

$= \frac{11}{36} - \frac{6}{36}$

$= \frac{5}{36}$

V) P$(\text{A} \cap \bar{B}) = 1 - P(A \cap B)$

$= 1\text{-}\frac{1}{6}$

$= \frac{5}{6}$

vi) P$(\bar{A} \cup B) = P(\bar{A}) + P(B) - P(\bar{A} \cap B)$

$= (1\text{-}\frac{18}{36}) + \frac{11}{36} \text{-} \frac{5}{36}$

$= \frac{2}{3}$

vii) P$(\overline{\text{A} \cup \text{B}}) = 1 - P(A \cup B)$

$= 1\text{-}\frac{23}{36} = \frac{13}{36}$

viii) P$(\bar{A} \cap (A \cup B)] = P[(A \cap \bar{A}) \cup (\bar{A} \cap B)$

$= \text{P}\ (\bar{A} \cap B) \quad [\ \text{A}\cap \bar{A} = \emptyset]$

$= \frac{5}{36}$

ix) P(A∪ $(\bar{A} \cap B)] = P(A) + P(\bar{A} \cap B) - P(A \cap \bar{A} \cap B)$

$= \text{P(A)} + \text{P}(\bar{A} \cap B)$

$= \frac{18}{36} + \frac{5}{36}$

$= \frac{23}{36}$

x. P(A│B) $= \frac{P(A\cap B)}{P(B)}$

$= \frac{6/36}{11/36}$

$= \frac{6}{11}$

P(B│A) $= \frac{P(A\cap B)}{P(A)}$

$= \frac{6/36}{18/36}$

$= \frac{6}{18}$

$=\frac{1}{3}$

xi. P$(\bar{A}$│B) $= \frac{P(\bar{A}\cap B)}{P(B)}$

$= \frac{13/36}{25/36}$

$=\frac{13}{25}$

P $(\bar{B}/\bar{A}) = \frac{P(\bar{A}\cap\bar{B})}{P(\bar{A})}$

$$= \frac{13/36}{25/36} = \frac{13}{18}$$

Example:

If two dice are thrown, what is the probability that the sum is a) greater than 8 and b) neither 7 nor 11?

Solution:

a.) If S denotes the sum on the two dice then we want P(S>8)

i. S=9, (ii) S=10, iii) S=11 iv) S=12

Hence by addition theorem of probability

P(S>8) = P(S=9) + P(S=10) + P(S=11) + P(S=12)

n(S) = 36

The number of favorable cases can be enumerated as follows

S=9: (3,6), (6,3), (4,5), (5,4)

i.e. 4 sample points

P(S = 9) = $\frac{4}{36}$

S= 10; (4,6), (6,4), (5,5) i.e. 3 sample points

P(S=10) = $\frac{3}{36}$

S=11: (5,6), (6,5) i.e. 2 sample points

P(S=11) = $\frac{2}{36}$

S=12: (6,6) i.e. 1 sample point

P(S = 12) = $\frac{1}{36}$

P(S>8) = $\frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{10}{36} = \frac{5}{18}$

b. Let A denotes the event of getting the sum of 7 and B, the event of getting the sum of 1 with a pair of dice.

S=7; (1,6), (6,1), (2,5), (5,2), (3,4), (4,3)

ii. i.e. 6 distinct sample points

P(A) = P(S=7) = $\frac{6}{36} = \frac{1}{6}$

S=11; (5,6), (6,5)

P(B) = P(S=11) = $\frac{2}{36} = \frac{1}{18}$

Required probability = P($\bar{A} \cap \bar{B}$)

= 1-P(A∪ B)

= 1-$[P(A) + P(B)]$ ($\therefore$ A and B are $disjoint$ $events$) =1$- \frac{1}{6} - \frac{1}{18}$

$= \frac{7}{9}$

Example:

An urn contains 4 tickets numbered 1,2,3,4 and another contains 6 tickets numbered 2,4,,6,7,8,9. If one of the two urns is chosen at random and a ticket is drawn at random from the chosen urn, find the probability that the ticket drawn bears of the number

i. 2 or 4 (ii) 3 (iii) 1 or 9

Solution:

1. Required event can happen in the following mutually exclusive ways,

I. First urn is chosen and then a ticket is drawn

II. Second urn is chosen and then a ticket is drawn

Since the probability of choosing any urn is ½ the required probability P is given by

P= P(I) + P(II)

$= \frac{1}{2} \times \frac{2}{4} + \frac{1}{2} \times \frac{2}{6} = \frac{5}{12}$

ii) Required probability = $\frac{1}{2} \times \frac{1}{4} + \frac{1}{2} \times 0$

$$= \frac{1}{8}$$

iii) Required probability = $\frac{1}{2} \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{6} = \frac{5}{24}$

Example:

A card is drawn from a well – shuffled pack of playing cards. What is the probability that it is either a spade or on ace.

Solution:

Let A and B denote the events of events drawing a spade card and an ace respectively. Then A consists of 13 sample points and B consists of A sample points.

i.e. P (A) = $\frac{13}{52}$ and P (B) = $\frac{4}{52}$

The compound event A$\cap B$consists of only one sample point

i.e. P (A$\cap B) = \frac{1}{52}$

The probability that the card drawn is either a spade or an ace is given by

$$\text{P}A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52}$$

$$= \frac{4}{13}$$

Example:

A box contains 6 red, 4 white and 5 black balls. A person draws 4 balls from that among the balls drawn there is at least one ball of each color

Solution:

The required event E that in a draw of 4 balls from the box at random there is at least one ball of each color can materialize the following mutually disjoint ways.

i) 1 red, 1 white, 2black balls
ii) 2 red, 1 white, 1 black balls
iii) 1 red, 2 white, 1 black balls

Hence by the addition theorem of probability the required probability is given by, P (E) = P(i) + P(ii) + P(iii)

$= \frac{6c_1 \times 4c_1 \times 5c_2}{15c_4} + \frac{6c_2 \times 4c_1 \times 5c_1}{15c_4} + \frac{6c_1 \times 4c_2 \times 5c_1}{15c_4}$

$= \frac{1}{15c_4}$[6×4×10+15×4×5+6×6×5]

$=\frac{4!}{15\times14\times13\times12}$ [240+300+180]

$= \frac{24 \times 720}{15\times14\times13\times12}$ = 0.5275

Example:

A problem is statistic is given to the three students A,B and C whose chances of solving it are ½, -3/4, and ¼ respectively. What is the probability that the problem will be solved?

Solution:

Let A ,B ,C denote the events that the problem is solved by the students A,B,C respectively.

Then, P (A) = ½, P(B) = ¾ and P(C) = 1/4

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

$$= P(A) + P(B) + P(C) - P(A).P(B) - P(A).P(C) - P(B).P(C) + P(A).P(B).P(C)$$

$$= \frac{1}{2} + \frac{3}{4} + \frac{1}{4} - \frac{1}{2} . \frac{3}{4} - \frac{3}{4} . \frac{1}{4} - \frac{1}{2} . \frac{1}{4} + \frac{1}{2} . \frac{3}{4} . \frac{1}{4}$$

$$= \frac{29}{32}$$

Example:

A bag contains 6 white 9 black balls. Four balls are drawn at a time. Find the probability for the first draw to give 4 white and the second to give 4 black balls in each of the following cases.

i. The balls are replaced before the second draw

ii. The balls are not replaced before the second draw

Solution:

1. The experiment of drawing 4 balls from a bag containing 6 white and 9 black balls result in $15C_4$ ways and hence the sample space consist $15C_4$ sample points.

Let A be the event that the first drawing gives 4 white balls and B be the event that the second drawing gives 4 black balls.

The event A consists of $6C_4$ sample points as there are 6 white balls and 4 are to be chosen from them

P(A) = $\frac{6C_4}{15C_4}$

Now, if the drawn balls are not replaced our sample space is reduced to $11C_4$ points only. The event B that the second draw results in 4 black balls.

P($\frac{B}{A}$) $= \frac{9C_4}{11C_4}$

Hence, P(A$\cap B) = P(A)\times\ P(B\,|\,A)$

$$=\frac{6C_4}{15C_4}\times\frac{9C_4}{11C_4}$$

$$=\frac{3}{715}$$

ii) consider the same experiment with replacement

P(A) = $\frac{6C_4}{15C_4}$

Whether A has occurred or not, the probability of drawing 4 black ball in the second draw is $9C_4\,|\,15C_4$

P(A$\cap B) = P(A)\times\ P(B\,|\,A)$

= P(A).P(B), as B is independent of A

$$=\frac{6C_4}{15C_4}\times\frac{9C_4}{15C_4}=\frac{6}{5926}$$

Exercise:

1. A bag contains 6 balls of different colors and a ball is drawn from its. A speaks truth thrice out of 4 times and B speaks truth 7 times out of times. If both A and B say that a red ball was drawn, find the probability of their joint statement being true (Ans : 7/15)

2. A and B are two very weak students of statics and their chances of solving a problem correctly are 1/8 and 1/12 respectively if the probability of their making a common mistake is 1/1001 and they obtain the same answer, find the chance that their answer is correct (Ans : 13/14)

3. A bag contains 10 balls, two of which are red three blue and 5 black. Three balls are drawn at random from the bag, that is every ball has an equal chances of being included is the three what is the probability that

i) The three balls are of different colors

ii) Two balls are of the same colors and

iii) The balls are all of the same color?

Ans: $\frac{30}{120}$; $ii)$ $\frac{79}{120}$ $iii)$ $\frac{11}{120}$

Mathematical expectation.

Definition:

Let x be a discrete random variable which can assume any of the values x_1, x_2,..x_n with corresponding probabilities P_i= P(x=x_i) i=1,2,… Then the mathematical expectation of x, denoted by E(x) is defined by

E(x) = $\sum_i P_i X_i$ provided the series is absolutely convergent.

Example:

1.Let X be the discrete random variable taking the values 1,2,…, 6 with corresponding probabilities P_i = 1/6

Then E(x) = $\sum_{i=1}^{6} P_i X_i$

$$= \frac{1}{6}(1) + \frac{1}{6}\ (2) +\ \ldots\ldots. + \frac{1}{6}(6)$$

$$= \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}$$

2. Let x be a random variable having the p.d.f

$$P(x) = \begin{cases} \frac{x}{6} & if\ x = 1,2,3 \\ 0 & otherwise \end{cases}$$

Then E(x^3+2x^2) = E (x^3) + 2 E (x^2)

$$= \sum_{i=1}^{3} x^3 P(x_i) +\ 2\sum_{i=1}^{3} x^2\ P(x_i)$$

$$= (\frac{1}{6} + \frac{16}{6} + \frac{81}{6}) + 2\ (\frac{1}{6} + \frac{8}{6} + \frac{27}{6})$$

$$= \frac{98}{6} + 2 \times \frac{36}{6}$$

$$= \frac{170}{6}$$

$$= \frac{85}{3}$$

Mathematical expectation of continuous random variable

Definition:

If x is a continuous random variable with p.d.f f (x) then mathematical expectation of x is defined to be E(X) = $\int_{-\infty}^{\infty} xf(x)dx$

Provided the integral is absolutely convergent.

E ($\psi(x)$) =$\int_{-\infty}^{\infty}\ \psi\ (x)f(x)\ dx\ where\ \psi(x) is\ a\ function\ of\ r.v.X.$

Example:

Let x have a p.d.f f(x) = $\begin{cases} \frac{x+2}{18} if -2 < x < 4 \\ 0 \quad \text{otherwise} \end{cases}$

Solution:

i. E(x) = $\int_{-2}^{4} x(\frac{x+2}{18})\, dx$

$= \frac{1}{18}\,[\frac{x^3}{3} + x^2]^{4}_{-2}$

$= \frac{1}{18}\left[\left(\frac{64}{3} + 16\right) - \left(\frac{-8}{3} + 4\right)\right]$

$= \frac{1}{18}\,[\frac{108}{3}] \qquad = 2$

ii. E[(x+2)^2] = E(x^2+4x+4)

= E(x^2) +4E(x) + 4

$= \int_{-2}^{4} x^2(\frac{x+2}{18})\, dx + 4E\,(x) + 4$

$= \frac{1}{18}\,[\frac{x^4}{4} + \frac{2x^3}{3}]_{-2}^{4}$ + (4x2 + 4)

$= \frac{1}{18}\,[\left(64 + \frac{128}{3}\right) - \,(4 - \frac{16}{3})]$+12

$= \frac{1}{18}[\frac{320}{3} + \frac{4}{3}]$ + 12

= 18 + 12 = 30

Definition:

Let X be a r.v E(x) is called the mean value of x and is denoted by μ.

Hence $\bar{X}$ = μ= E(X)

E(X^r), r≥1 is called the rth moment of X about the origin and is denoted by μ_r^1.Hence $\mu_r' = E(X^{r)}$and

$\bar{x} = \mu_1'$

E(X- μ)^2 is called the variance of X and is denoted by $\sigma^{2.}$ The positive square root σ of the variance is called the standard deviation of X.

E(X- μ)^r is called the rth central moment of X and is denoted by μ_{r} .

Hence $\mu_{\text{r=}}$ E(X- μ)^r

Lemma:

σ^2= E(X^2) – [E(x)]^2 = $\mu_2 - \mu_1'^{\,2}$

Proof:

$$\sigma^2 = \text{E}\ [(\text{x} - \mu)^2]$$

$$= E(x^2 - 2\mu x + \ \mu^2)$$

$$= E(x^2) - \ 2\mu E(x) + \ \mu^2$$

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

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

$$= \mu_2' - \mu_1'^{\,2}$$

Lemma:

$$\mu_r = \mu_r' - r_{c_1}\mu\mu_{r+1} + r_{c_2}\mu^2\mu_{r-2}\dots\dots\dots$$

Proof:

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

=$E\left(X^r - r_{c_1}\mu X^{r-1} + \cdots..\right)$

=$\mu_r' - r_{c_1}\mu\mu_{r-1} + r_{c_2}\mu^2\mu_{r-2}\dots\dots\dots$

In particular, $\mu_1 = E(X-\mu) = E(X) - \mu = \mu - \mu = 0$.

$$\mu_2 = \mu_2' - 2\mu_1'\mu + \mu^2\mu_0$$

$\therefore \mu_2 = \mu_2' - \mu_1'^2$ (Since $\mu = \mu_1'\ and \mu_0' = 1$)

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^2$$

Problem:

A random variable X is defined as the sum of the numbers on the faces when two dice are thrown. Find the expected value of X.

Solution:

The probability distribution of X is given by the following table.

x_i	2	3	4	5	6	7	8	9	10	11	12
P(x_i)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

$E(X) = \sum x_i p(x_i)$

=2(1/36)+3(2/36)+4(3/36)+5(4/36)+6(5/36)+7(6/36)+8(5/36)+9(4/36)+10(3/36)+11(2/36)+12(1/36)=252/36=7.

Problem:

A random variable X has the following probability functions.

x_i	-2	-1	0	1	2	3
P(x_i)	0.1	K	0.2	2k	0.3	k

Find (i) the value of k (ii) mean (iii) variance (iv) $p(x \geq 2)$ (v) $p(x < 2)$

(vi) $p(-1 < x < 3)$.

Solution:

(i) $\sum p_i = 1$

⇒ 0.1+k+0.2+2k+0.3+k=1

⇒4k=0.4

⇒k=0.1.

x_i	-2	-1	0	1	2	3
P(x_i)	0.1	0.1	0.2	0.2	0.3	0.1

Hence the probability function is

(ii) Mean=$E(X) = \sum x_i p(x_i)$.

=(-2)(0.1)+(-1)(0.1)+(0)(0.2)+(1)(0.2)+2(0.3)+3(0.1)

=0.8.

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

$=\sum {x_i}^2 p(x_i)$-0.8^2

$=[4(0.1) + 1(0.1) + 1(0.2) + 4(0.3) + 9(0.1)$

=2.8-0.64=2.16.

(iv) $p(x \geq 2) = p(X = 2) + p(= 3)$=0.3+0.1=0.4.

(v) $p(x < 2) = 1 - \ p(x \geq 2) = 0.6$.

(vi) $p(-1 < x < 3) = p(X = 0) + p(X = 1) + p(X = 2)$

=0.2+0.2+0.3=0.7.

Problem:

Let X have the p.d.f f(x)= $\begin{cases} \frac{x+1}{2}, & \text{if } -1<x<1 \\ 0, & \text{otherwise.} \end{cases}$

Find the mean and standard deviation of x.

Solution:

$$\mu = E(x) = \int_{-\infty}^{\infty} x f(x) dx = \int_{-1}^{1} x \left(\frac{x+1}{2}\right) dx$$

$$=1/2\left[\frac{x^3}{3} + \frac{x^2}{2}\right]_{-1}^{1}$$

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

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

$$= \int_{-1}^{1} \frac{x^2}{2}(x+1)dx - \left(\frac{1}{3}\right)^2$$

$$=1/2\left[\frac{x^4}{4} + \frac{x^3}{3}\right]_{-1}^{1} - \frac{1}{9}$$

$$\sigma^2 = \frac{2}{9}$$

Moment Generating Function:

Definition:

The moment generating function (m.g.f) for any random variable X about the origin is defined by $M_X(t) = E(e^{tx}) =$

$$f(x) = \begin{cases} \int (e^{tx})\, f(x)dx \quad if\ X\ is\ a\ continuos\ r.v\ with\ p.d.f\ f(x) \\ \sum_x e^{tx} P(x), \qquad if\ X\ is\ a\ \ discrete\ r.v\ with\ p.d.f\ P(x) \end{cases}$$

Where the integration or summation is taken over the entire range of X and t is a real parameter.

Definition:

More generally the m.g.f of a random X about a point a denoted by $M_{X=a}(t) is\ defined\ by\ M_{X=a}(t) = E\left(e^{t(x-a)}\right)$.

Example :

Define P(x) = $\begin{cases} \frac{6}{\pi^2 x^2},\ if\ x = 1,2,3,\ldots \\ 0, \quad otherwise \end{cases}$

Now, $\sum P(x) = \sum \frac{6}{\pi^2 x^2}$ $= \frac{6}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2}$

$=\frac{6}{\pi^2} \times \frac{\pi^2}{\mathbf{6}}$ (Since 1 + $\frac{1}{2^2}$ + $\frac{1}{3^3}$ ... += $\frac{\pi^2}{6}$)=1

$\therefore P(x) is\ a\ probability\ distribution\ of\ the\ r.v\ X.$

Now, the m.g.f of X, if it exists, is given by $M_X(t) = E(e^{tx}) = \sum e^{tx}\ P(x)$

$=\sum e^{tx} \left(\frac{6}{\pi^2 x^2}\right)$

$=\frac{6}{\pi^2} \sum_{n=1}^{\infty} \left(\frac{e^{tn}}{n^2}\right)$

The above series, by ratio test ,diverges for t>0.

$$\therefore Moment\ generating\ function\ of\ X\ does\ not\ exist.$$

Properties of moment generating function:

1. The rth derivative of m.g.f of r.v X at t=0 is μ_r'.

Proof:

Let $M_X(t) be\ the\ m.g.f\ of\ the\ r.v\ X$

Then $M_X(t) = 1 + \mu_1' + \frac{\mu_2'}{2!} t^2 + \cdots + \frac{\mu_r'}{r!} t^r + \cdots$

$$\therefore \frac{d^r}{dt^r}(M_X(t)) = \frac{\mu_r'\ r!}{r!} + \frac{t\mu_{r+1}'}{(r+1)!} + \dots$$

At t=0, $\frac{d^r}{dt^r}(M_X(t)) = \mu_r'$

Problem:

A random variable X has the probability function P(x)=$\frac{1}{2^x}$; $x = 1,2,3 \ldots\ Find\ its\ (i) m.g.f$

(ii) Mean and (iii) Variance.

Solution:

X is a discrete random variable

(i) $(M_X(t)$=E$(e^{tx}) = \sum_x P(x)$

$= \sum_{x=1}^{\infty} e^{tx} \left(\frac{1}{2^x}\right)$

$= \sum_{x=1}^{\infty} \left(\frac{e^t}{2}\right)^x$

$= \frac{e^t}{2} + \left(\frac{e^t}{2}\right)^2 + \dots$

$$=\frac{e^t}{2}\left[1+\frac{e^t}{2}+\left(\frac{e^t}{2}\right)^2\right]$$

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

$$M_X(t) = \frac{e^t}{2-e^t}$$

(ii) $\frac{d}{dt}M_X(t) = \frac{(2-e^t)e^t+e^te^t}{(2-e^t)^2} = \frac{2e^t}{(2-e^t)^2}$

$$\therefore \mu_1' = \left[\frac{d}{dt}M_X(t)\right]_{t=0} = 2$$

$$\frac{d^2}{dt^2}(M_X(t) = \frac{(2-e^t)^2 2e^t+2e^t 2(2-e^t)}{(2-e^t)^4}$$

$$=\frac{8e^t-2e^{2t}}{(2-e^t)^3}$$

$$\therefore \mu_2' = \left[\frac{d^2}{dt^2}M_X(t)\right]_{t=0} = 6$$

(iii) Variance $\mu_{2\,=}\,\mu_2' - (\mu_1')^2$

=6-4 = 2

Problem :

Find the m.g.f of the r.v X having the p.d.f

$$\boldsymbol{f(x)} = \begin{cases} \frac{\mathbf{-1}}{\mathbf{3}}, & \mathbf{-1} < x < 2 \\ \mathbf{0}, & \boldsymbol{otherwise} \end{cases}$$

Solution:

X is a continuous random variable.

$$M_x(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

$$=\frac{1}{3}\int_{-1}^{2} \boldsymbol{e^{tx}}\, \boldsymbol{dx} \;=\frac{1}{3}\left[\frac{e^{tx}}{t}\right]_{-1}^{2} \quad \text{when } t\neq 0$$

$$=\frac{e^{2t}-e^{-t}}{3t} \qquad \text{when } t\neq 0$$

when t$= 0$, $M_x(t) = \frac{1}{3}\int_{-1}^{2} \boldsymbol{dx} \;=\frac{1}{3}[\boldsymbol{x}]_{-1}^{2}=\boldsymbol{1}$

$$M_x(t) = \begin{cases} \dfrac{e^{2t} - e^{-t}}{3t}, & \text{when } t \neq 0 \\ 1 & \text{when } t = 0 \end{cases}$$

Some Special Distributions:

Introduction:

In this chapter we discuss some important distribution of random variable which are frequently used is statistics. We make a detailed study of binomial distribution, Poisson distribution which are of discrete type and normal distribution which is of continuous type.

Binomial Distribution:

Definition:

Let n be any positive integer and let $0 < p < 1,\; Let\; q = 1 - p$

Define $p(x) \;= \begin{cases} n_{c_x} p^x q^{n-x} & if\; x = 0,1,2, \dots . n \\ 0 & otherwise \end{cases}$

A discrete random variable with the above p.d.f. is said to have binomial distribution and the p.d.f itself is called a binomial distribution.

Note: 1

The two independent constants n and p in the distributions are known as the parameters of the distribution. If x is a binomial variate with parameters n and p we write as

$$X{\sim}B(n, P).$$

Note:2

In this experiment is repeated N times (say) then the frequency function of the binomial distribution is given $by\ f(x) = NP(x) = Nn_{c_x}p^xq^{n-x}, x = 0,1,2, \dots \dots . . n$.

Theorem: 1

m.g.f of a binomial distribution about the origin is

$(q + pe^t)^n$.

Proof:

$$M_X(t) = E(e^{Xt}) = \sum e^{Xt}p(x) = \sum_{x=0}^{n}\left(e^{tx}n_{c_x}p^xq^{n-x}\right)$$

$$=\sum_{x=0}^{n} n_{c_x}\,(pe^t)^xq^{n-x}$$

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

Moments of binomial distribution:

We know that for any random variable x the m.g.f is $M_X(t) = 1 + \mu_1^{'} + \frac{\mu_2^{'}}{2!}t^2 + \dots + \frac{\mu_r^{'}}{r!}t^r + \cdots \dots$

Theorem: 2 (Addition property of binomial distribution)

If $X_1 \sim B(n_1, p), X_2 \sim B(n_2, p)$ are independent random variable then $X_1 + X_2$is $B(n_1 + n_2, p)$.

Proof:

GivenX_1, andX_2 are independent random variable with parameters $n_1, p\ and\ n_2, p$ respectively.

Let us consider m.g.f of X_1 and X_2about origin.

$\therefore M_{X_1} = (q + pe^t)^{n_1}\ M_{X_2}(t) = (q + pe^t)^{n_2}$

Now, $M_{X_{1+}X_2}(t)$=$M_{X_1}(t) + M_{X_2}(t)$ (since X_1, and X_2 are independent)

$$= (q + pe^t)^{n_1} + (q + pe^t)^{n_2}$$

$$= (q + pe^t)^{n_1+n_2}$$

= m.g.f of the binomial $X_{1+}X_2$with parameters $n_1 + n_2$ and P.

Hence the uniqueness theorem $X_{1+}X_2$ is a binomial variable with parameters

$n_1 + n_2$ and P.

Theorem: 3

Characteristic function of binomial distribution is $\left(q + pe^{it}\right)^n$.

Proof:

Let $X \sim B(n, P)$. Hence $\varphi_x(t) = E\left(e^{itx}\right)$

$= \sum_{x=0}^{n} e^{itx} p(x) = \sum_{x=0}^{n} e^{itx} n_{c_x} p^x q^{n-x}$

$= \sum_{x=0}^{n} n_{c_x} \left(pe^{it}\right)^x q^{n-x}$

$= (q + Pe^{it})^n$.

Mode of Binomial Distribution:

Let $X \sim B(n, P).\ then\ P(x) = n_{c_x} p^x q^{n-x}$ .Let x be the mode of the binomial distribution.

Example:

The unbiased coins are tossed and number of heads noted. The experiment is repeated 64 times and the following distribution is obtained.

No.of heads	0	1	2	3	4	5	Total
Frequencies	3	6	24	26	4	1	64

Solution:

Here n=5 and N=64

Since the coins are unbiased p=1/2 =q. So that p/q =1

Now p(0)=q^n=$(1/2)^5$ =1/32

Hence f(0)=N q^n =64×1/32 = 2

Using the recurrence formula $p(x+1) = \left(\frac{n-x}{x+1}\right)\left(\frac{p}{q}\right)p(x)$

P(1)= 5(1/32).

Hence f(1)=10

x	Probabilities $p(x)$	Expected frequencies $f(x) = Np(x)$	Observed frequencies
0	p(0)=1/32	2	3
1	P(1)=5/32	10	6
2	P(2)=10/32	20	24
3	P(3)=10/32	20	26
4	P(4)=5/32	10	4
5	P(5)=1/32	2	1
Total		64	64

Problem:

In a binomial distribution the mean is 4 and the variance is 8/3.Find the mode of the distribution.

Solution:

Given mean =4 and Variance=8/3

$\therefore np = 4$ and npq=8/3

$\frac{npq}{np}=\frac{8}{3\times4} = \frac{2}{3}$

q = 2/3,p=1/3

$\therefore np = 4$

n=12

Consider (n+1)p=13/3=4.3

Hence the mode is 4.

Problem:

A discrete random variable X has the mean 6 and variance 2. If it is assumed that the distribution is binomial. Find the probability that $5\leq x \leq 7$.

Solution:

Given np=6 and npq=2

Hence q=1/3 and p=2/3

Also n=9

Now, $p(5\leq x \leq 7)=p(x=5)+p(x=6)+p(x=7)$

$$=9C_5\left(\frac{2}{3}\right)^5\left(\frac{1}{3}\right)^4+9C_6\left(\frac{2}{3}\right)^6\left(\frac{1}{3}\right)^3+9C_7\left(\frac{2}{3}\right)^7\left(\frac{1}{3}\right)^2$$

$$=\frac{126}{3^9}x2^5+84\times\frac{2^6}{3^9}+36\times\frac{2^7}{3^9}$$

$$=\frac{2^5}{3^9}[126+168+144]$$

$$=\frac{2^5}{3^9}\text{ x }438$$

$$=\frac{32\times 438}{19683}$$

= 0.712

Problem :

An insurance agent accepts policies of 5 men all of identity age and in good health. The probability that man of this age will be alive 30 years hence is 2/3. Find the probability that is 30 years (i) all five men (ii) at least one man (ii) almost three will be alive.

Solution:

Here n=5, p=2/3, q=1-p=1-2/3=1/3

It is a binomial distribution and x~B (5,2/3) Hence

$P(X=x) = nC_xP^xq^{n-x}$

Probability of all 5 will be alive is

$P(x=5) = 5C_5 (2/3)^5 = 32/243$

ii) Probability of at least one being alive = 1- probability of no one being alive. Probability of no one being alive

$P(x=0) = 5C_0 (1/3)^5 = 1/243$

Probability of at least one being alive

= 1-1/243

= 242/243

iii) Probability of almost 3 being alive = probability of one man being alive or probability of 2 men being alive (or) probability of 3 men being a live.

= 1- [Prob. Of 4 men being a live (or) probability of 5 men being alive]

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

$= 1- [P(x=4) + P(x=5)]$

$=1-[5C_4 (2/3)^4 (1/3) + 5C_5 (2/3)^5]$

$= 1-[5(16/243)+ 32/243]$

$= 1-112/243 = 131/243$

Problem :

Six dice are thrown 729 times. How many times do you except at least 3 dice to show a five or six.

Solution:

Here n = 6, N= 729

P=prob. Of getting 5 or 6 with one dice = 2/6 = 1/3

Q = 1-1/3 = 2/3

The expected frequency of 0,1,2,..6. successor are the successive terms of

$729 (1/3 + 2/3)^6$

Excepted number of times at least 3 dice showing five or six

$= 729 (6C_3 (1/3)^3 + 6C_4(1/3)^4 (2/3)^2 + 6C_5 (1/3)^5 2/3 + 6C_6 (1/3)^6]$

$= 729/3^6 \times 233$

$= 169857/729$

$=233.$

Problem :

If the m.g.f. of a r.v.x is of the form $M_x(t) = (0.4e^t+0.6)^8$ find i) E(x)

ii) the m.g.f of the r.v Y= 3X+2

Solution:

We have the m.g.f of the binomial variable X ~ B(n,p) is $(q+pe^t)^n$

Here p=0.4, q=0.6 and n=8

i) We have E(X) = $\mu_1 = np$

E(X) = 8x0.4

= 3.2

ii) $M_y(t) = M_{3X+2}(t)$

$= e^{2t} M_x(3x)$

$= e^{2t}(0.6+0.4e^{3t})^8$

Poisson distribution:

We have the binomial distribution is determined by 2 parameters p and n. If the number of trails is in definitely large and the probability p of success is in definitely small such that np=λ,where λ is a constant then the limiting case of the binomial distribution when n→ ∞ *and* $p \to 0$ becomes a distribution known as Poisson distribution

Definition:

A discrete random variable X is said to be follow a Poisson distribution if it assumes only nonnegative integer values and its probability density function is given by

$$p(x) = p(X = x) = f(x) = \begin{cases} \frac{e^{-x}\lambda^x}{x!}, & if\ x = 0,1,2 \ldots \\ 0, & otherwise \end{cases}$$

Where λ is the parameter of the distribution if X is a Poisson variate with parameter λ.we write X$\sim p(\lambda)$

Example:

If fit a Poisson distribution to the following data

x	0	1	2	3	4	Total
f	123	9	14	3	1	200

Solution:

To fit a Poisson distribution we have to calculate all the excepted frequencies

Here the mean $\lambda = \frac{\sum f_i x_i}{\sum f_i}$

$= \frac{59+28+9+4}{200}$ =0.5

$\lambda = 0.5$

$Hence\ the\ p.d.f\ of\ poisson\ distribution\ is$

$$p(x) = \frac{e^{-0.5}(0.5)^x}{x!}$$

$\therefore p(0) = e^{-0.5}$ = 0.6065

Hence f(0)=Np(0)

=200× (0.6065)

x	Probabilities using p(x+1)=(λ/(x+1))p(x)	Expected frequencies f(x)=Np(x)	Observed frequencies
0	P(0)=0.6065	121.3	123

1	P(1)=0.3033	60.66	59
2	P(2)=0.0758	15.16	14
3	P(3)=0.0126	2.52	3
4	P(4)=0.0027	0.54	1

=1213

Using recurrence formula

VP(x+1)=$\left(\frac{\lambda}{(x+1)}\right)p(x)$

We have p(1) = $\frac{0.5}{p(0)}$

=0.3033

=60.66

Problem :

The probabilities of a Poisson variable taking the values 3 and 4 are equal calculate the probabilities of variates taking the values 0 and 2

Solution

Take x be a Poisson variate with parameter λ

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

Given P(X=3)=P(X=4)

To find P(X=0) and P(X=1)

P(X=3)=P(X=4) ⇒ $\frac{e^{-\lambda}\lambda^{3}}{3!}$

=$\frac{e^{-\lambda}\lambda^{4}}{4!}$ ⇒ λ=4

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

P(X=2)= $\frac{e^{-4}4^{2}}{2!}$=0.146.

Problem :

Assuming that one in 80 births in a case twins. Calculate the probability of 2 or more birth of twins on a day when 30 births occur using (i) binomial distribution (ii) Poisson approximation.

Solution:

Assuming X to be a binomial variate with p=probability of twin births=1/80=0.125.

(q=0.9875) where n=30 we get

$$p(x) = 30_{c_x}(0.0125)^x(0.9875)^{30-x}$$

Probability of 2 or more births of twins on a day is

$$p(X \geq 2) = 1 - p(X < 2)$$

=1-[p(X=0)+p(X=1)]

=1-[$(0.9875)^{30}$+30(0.0125)$(0.9875)^{29}$

=1-0.6943(1.3625)

=0.054

Assuming X to be a Poisson variate with λ=np=0.375

We get $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$=$\frac{e^{-0.375}\,0.375^x}{x!}$

$$p(X \geq 2) = 1 - [p(X = 0) + p(X = 1)]$$

=1-$e^{-0.375}$+$e^{-0.375}$(0.375)

=1-0.6873(1.375)

=0.0550.

NORMAL DISTRIBUTION:

Normal distribution is one of the most widely used distribution in application of statistical methods.

We have $\int_{-\infty}^{\infty} e^{-y^2/2}\, dy = \sqrt{2\pi}$

Hence $\int_{-\infty}^{\infty} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\, dy$=1

Put $y = \frac{x-\mu}{\sigma}$ then we have $\int_{-\infty}^{\infty} \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}\, dx = 1$

Definition:

A continuous random variable X is said to follow a normal distribution if its probability density function is given by f(x)= $\frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$ where $-\infty < x < \infty$

μ *and* σ are constants and σ>0 and are called the parameters of the distribution and we write $X \sim N(\mu, \sigma^2)$.

Fitting of normal distribution:

To fit a normal distribution to given date we first calculate the mean μ *and* $s.d$ σ.Thus the normal curve fitted to the given data is given by

f(x)= $\frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$ where $-\infty < x < \infty$

Properties of normal distribution:

1. The normal probability curve is symmetrical about the ordinate at $x = \mu$.The ordinate decreases rapidly as x increases. The curve extends to infinity on either side of the mean. The X-axis is an asymptote to the curve.

2. The mean, median and mode coincide the maximum ordinate at $x = \mu$ is given by $\frac{1}{\sigma\sqrt{2\pi}}$.

3. $\mu \pm \sigma$ are the points of inflexion of the normal curve and hence the points of inflexion are also equidistant from the median.

4.The area under normal curve is unity. The ordinate at $x = \mu$ divides the area under the normal curve into two equal parts symmetry also ensures that the first and third quartiles of normal distribution are equidistant from the median of course on either side.

5. $p(\mu - \sigma < x < \mu + \sigma) = 0.6826$

$p(\mu - 2\sigma < x < \mu + 2\sigma) = 0.9544$

$p(\mu - 3\sigma < x < \mu + 3\sigma) = 0.9973$

6.Q.D:M.D:S.D=10:12:15.

Problem :

If X is normal distributed with zero mean and unit variance.

Find the expectation of X^2.

Solution:

Given $X \sim N(0,1)$

Hence the normal is $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Now $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx$

$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx$

$= \frac{1}{\sqrt{2\pi}} [-xe^{-x^2/2}]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2} dx$

$= \frac{1}{\sqrt{2\pi}} \left[0 + \sqrt{2\pi}\right]$

=1.

Problem :

If X is normally distributed with mean 8 and s.d 4. Find

(i)P(5$\leq X \leq 10)$ $(ii) P(10 \leq X \leq 15)$

(iii) P($X \geq 5$) (iv) $P(\left|X - 5\right| \leq 15$.

Solution:

Given

$X \sim N(8,4)$

Hence the standard normal variate $Z = \frac{X-8}{4}$

When X=5;z=-0.75

When X=10;Z=0.50

When X=15;Z=1.75

When X=20 ;Z=3

(i) P(5$\leq X \leq 10) = P(-0.75 \leq Z \leq 0.50)$

=P(-0.75$\leq Z \leq 0) + P(0 \leq Z \leq 0.50)$

=P(0$\leq Z \leq 0.75) + P(0 \leq Z \leq 0.50)$

=0.2734+0.1915

=0.4649.

$(ii)P(10 \leq X \leq 15)$=P(0.5$\leq Z \leq 1.75)$

=P(0$\leq Z \leq 1.75) - P(0 \leq Z \leq 0.5)$

= 0.4599 – 1.1915

= 0.2684

iii) P (x≥5) = p(Z ≥1.75)

= 0.5 – p (0 $\leq z \leq 1.75)$

= 0.5 – 0.4599

= 0.0401

iv) p(x $\leq$ 5) = p(z ≤-0.75)

= 0.5 – p(-0.75 $\leq z \leq 0)$

= 0.5 – p(0 $\leq z \leq 0.75)$

$= 0.5 – 0.2734$

$= 0.2266$

Problem :

The marks of 1000 students in a university are found to be normally distributed with mean to and s.d. 5. Estimate the number of students whose marks will be i) between 60 and 75 (ii) more than 75 (iii) less than 68.

Solution :

Let x denote the marks of students .

Hence X∼N (70, 25)

The standard normal variate is

$$Z = \frac{X-\mu}{\sigma} = \frac{X-70}{5}$$

i) To find $p(60 < x < 75)$

When $x = 60$; $z= -2$ and when $x=75$; $z=1$

$P (60 < x < 75) = p (-2 < z < 1)$

$= p (-2 <z < 0) + p (0 < z < 1)$

$= p (0 < z < 2) + p (0 < z < 1)$

$= 0.4772 + 0.3413$

$= 0.8185$

∴ The number of students whose marks is between 60 and 75 is 1000 x 0.8185 = 819.

(ii) To find $p (x> 75)$

When $x = 75$; $z =1$

$\therefore p (x > 75) = p(z > 1)$

$= 0.5 – p (0 <z < 1)$

$= 0.5 – 0.3413$

= 0.1587

∴ The number of students whose marks is more than 75 is 159.

(iii) To find p (x < 68)

When x = 68, $z = \frac{68-70}{5} = -0.4$

∴p (x < 68) = p (z < -0.4)

= p(z > 0.4)

= 0.5 – p(0 < z < 0.4)

= 0.5 – 0.1554

= 0.3546.

Problem :

Assume the mean height of soldiers to be 68.22 inches with variance of 10.8 inches. How many soldiers in a regiment of 2000 soldiers would you expect to be over six feet tall. Assume heights to be normally distributed.

Solution :

Let the variable x denote the height in inches of the solders.

Mean $\mu = 68.22;\ \ \sigma^2 = 10.8, \sigma = 3.286$

Hence X ~N (68.22, 3.286)

∴p(x > 6 feet) = p (x > 72)

When x = 72; $z = \frac{x-\mu}{\sigma}$

$= \frac{72-68.22}{3.286}$

$= \frac{3.78}{3.286}$

= 1.15

∴p (x > 72) = p(z > 1.15)

= 0.5 – p (0 $\leq z \ \leq 1.15$)

= 0.5 – 0.3749

= 0.1251

∴ The number of soldiers in the regimet of 2000 over 6 feet tall is 2000 x 0.1251= 250

Problem :

A set of examination marks is approximately distributed with mean 75 and S.D. of 5. If the top 5% of students get grade A and the bottom 25% get grade B what mark is the lowest A and what mark is the highest B?

Solution :

It x denote the marks in the examination.

Given x is normally distributed with mean $\mu = 75 \ and \ \sigma = 5$

ie) X ~N (75, 25)

Let x_1 be the lowest marks for A and x_2 be the highest marks for B. Given

p(x >x_1) = 0.05 and p(x <x_2) = 0.25

The standard normal variate

$$z = \frac{x_{1-\mu}}{\sigma}$$

$$= \frac{x_1 - 75}{5} = z_1$$

$$Z = \frac{x_2 - \mu}{\sigma} = \frac{x_2 - 75}{5} = -z_2 \qquad \text{…………. (1)}$$

P(0 < z <z_1) = 0.45

z_1 = 0.45

P(-$z_2 < z < 0$) = 0.25

P(0 < z <z_2) = 0.25

z_2 = 0.675

(1)⇒ x_1 = 75 + 5z_1

x_1 = 83.225

$x_1 \approx 83$

$x_2 = 75 - 5z_2$

$x_2 = 71.625$

$x_2 \approx 72$

Hence the lowest mark for grade A is 83 and the highest mark for B is 72.

Problem :

In a normal distribution 31% of the items are under 45 and 8% are over 64. Find the mean and standard deviation.

Solution :

Let x denote the normal variate with mean $\mu \; and \; S.D. \sigma$

Given p(x < 45) = 0.31 and p (x > 64) = 0.08

When x = 45, $z = \frac{45-\mu}{\sigma} = -z_1$

When x = 64, $z = \frac{45-\mu}{\sigma} = z_2$

P(0 < z <z_2) = 0.42 and p($-z_1 < z < 0$) = 0.19

From the area table we get

$$z_1 = 0.496 \; and \; z_2 = 1.405$$

(1)⇒ $45 - \mu = 0.496 \, \sigma$

$64 - \mu = 1.405 \, \sigma$

$$\sigma = 9.99 \approx 10$$

$$\mu = 49.96 \approx 50$$

Problem :

Find the probability of getting between 3 heads to 6 heads in 10 tosses of a fair coin using (i) binomial distribution (ii) the normal approximation to the binomial distribution.

Solution :

(i). Take x as the binomial variate,

P = $\frac{1}{2}$, $q = \frac{1}{2}, n = 10$

We have X $\sim B\ (10, \frac{1}{2})$

Probability of getting at least 3 heads

= p (x ≥ 3)

= p(x=3) + p(x=4) + p(x=5) + p(x=6)

$$= 10c_3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7 + 10c_4\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^6 + 10c_5\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^5 + 10\ c_6\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^4$$

$= \frac{1}{2^{10}}[120 + 210 + 252 + 210]$

$= \frac{792}{2^{10}}$

= 0.7734

(ii). Taking the data as continuous it follows that 3 to 6 heads can be considered as 2.5 to 6.5 heads.

Mean $\mu = np$

= $10\left(\frac{1}{2}\right)$

= 5

$\sigma = \sqrt{npq} = 1.58$

$\therefore\ X \sim N\ (5, 1.58)$

The standard normal variate is $z = \frac{X-\mu}{\sigma}$

For X = 2.5 ;$z = \frac{2.5-5}{1.58}$

= -1.58

X = 6.5 ;$z = \frac{6.5-5}{1.58}$

Z = 0.95

P($2.5 \leq x\ \leq 6.5) =\ p\ (-1.58 < z < 0.95)$

= P (-1.58 < z < 0) + p(0 < z < 0.95)

$= p(0 < z < 1.58) + p (0 < z , 0.95)$

$= 0.4429 + 0.3289 = 0.7718.$

UNIT III : ASSOCIATION OF ATTRIBUTES

Association of Attributes - Coefficient of Association - Consistency - Time Series – Definition - Components Of Time Series - Seasonal and cyclic variations.

THEORY OF ATTRIBUTES

Attributes:

The qualitative characteristics of a population are called attributes and they cannot be measured by numeric quantities. Hence the statistical treatment required for attributes is different from that of quantitative characteristic.

Suppose the population is divided into two classes according to the presence or absence of a single attribute. The positive class denotes the presence of the attributes and the negative class denotes the absence of the attribute. Capital Roman letter such as A,B,C,D… are used to denote positive Greek letters such as $\alpha, \beta, \gamma, \delta$ …… are used to denote negative classes.

For example If A represents the attribute richness then α represents the attribute non-richness (poor).

A class represented by n attributes is called a class of nth order.

For example,

A,B,C, $\alpha, \beta, \gamma, \delta$ are all of first order, AB, A$\beta, \alpha B, \alpha\beta$ are of second order, and ABC, A$\beta\gamma, A\beta C, \alpha\beta\gamma$ are of the third order.

The number of individuals possessing the attributes in a class of nth order is called a class frequency of order 'n' and class frequencies are denoted by bracketing the attributes.

Thus (A) stands for the frequency of A the number of individuals possessing the attribute A and (Aβ) stands for the number of individuals possessing of the attributes A and not B.

Note:

1. Class frequencies of the type (A), (AB), (ABC) are known as positive class frequencies.
2. Class frequencies of the type $(\alpha),(\beta),(\alpha\beta),(\alpha\beta\gamma)\ldots.are$ known as negative class frequencies.
3. Class frequencies of the type $(\alpha B),(A\beta),(A\beta\gamma),(\alpha\beta C)$ .... are known as contrary frequencies.
4. The classes of highest order are called the ultimate classes and their frequencies are called the ultimate class frequencies.

Examples:

1. AB = (ABC) + (ABγ)

Consider, (ABγ) $= AB\gamma.N$

=AB(1-C).N

=AB.N – ABC.N

=(AB) – (BC)

$\therefore (AB) = (ABC) + (AB\gamma)$

2. If there are two attributes A and B we have,

N = (A) + (α) = (B) + (β)

Hence N = (A) + (α)

N = (AB) + (Aβ) + (αB) + ($\alpha\beta$)

And N = (B) +(β) = (AB) + $(\alpha B) + (A\beta) + (\alpha\beta)$

If there are three attributes A,B,C we have N = (A) + (α)

We have N = (A) + (α)

⇒ N = (AB) + (Aβ) +($\alpha\beta$) $+ (\alpha\beta)$

Thus,

N = (ABC) + (ABγ) $+ A\beta C) + (A\beta\gamma) + (\alpha BC) + (\alpha B\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)$

3. Consider two attributes A and B

Now, $(\alpha\beta) = \alpha\beta.N$

= (1-A) (1-B).N
= (1-A-B+ AB). N
= N-A.N- B.N + AB.N
= N-(A) – (B) + (AB)

4. (AB) = AB.N

= (1-γ) $(1-\beta).N$

= (1-$\alpha-\beta+\ \alpha\beta).N$

= $N-\alpha.N-\ \beta.N+\ \alpha\beta.N$

= N- $(\alpha)-(\beta)+\ (\alpha\beta)$

5. $(\alpha\beta\gamma)=\ \alpha\beta\gamma.N=(1-A)(1-B)(1-\ \ C).N$ = N-A.N-B.N – C.N + AB.N + AC.N + BC.N – ABC.N
= N-(A) – (B) –(C) + (AB) +(AC) +(BC) – (ABC)

6. N= (A) +(B) + (C) – (AB) – (BC) – (AC) + (ABC) + $(\alpha\beta\gamma)$

Problem :

Given (A) = 30, (B) = 25, $(\alpha)=30\ (\alpha\beta)=20.$

Find i) (N) (ii) (β) (iii) (AB) (iv) (Aβ) $(\ V)$ (αB)

Solution:

i) N = (A) + $(\alpha)=30+30=60$
ii) $(\beta)=N-(B)=\ 60-25=35$
iii) $(AB)=\ AB.N$

= (1-α) $(1-\beta).N$

= N-$(\alpha)-(\beta)+\ (\alpha\beta)$

= 60-30-35+20

= 15

iv.(A$\beta)=A\beta.N=A(1-B).N$
= (A) – (AB)
= 30-15
= 15

v. (αB) = $\alpha B.N = (1-A)B.N$

= (B)-(AB)

= 25-15

=10

Problem :

Given the following ultimate class frequencies of two attributes A and B. Find the frequencies of positive and negative class frequencies and the total number of observations.

(AB) = 975, $(\alpha B) = 100, (A\beta) = 25, (\alpha\beta) = 950.$

Solution:

Positive class frequencies are (A) and (B)

(A) = (AB) + $(A\beta) = 975 + 25 = 1000$

(B) = (AB) +$(\alpha B) = 975 + 100 = 1075$

Negative class frequencies are (α) and (β)

$(\alpha) = (\ \alpha B) + (\alpha\beta) = \ 100 + 950 = 1050$

$(\beta) = (A\beta) + \ (\alpha\beta) = \ 235 + 950 = 975$

N = (A) + (α) = (B) + (β)

Taking,

N = (A) + $(\alpha) = 1000 + 1050 = 2050$

Problem :

Given the following positive class frequencies find the remaining class frequencies N = 20 (A) = 9; (B) = 12; (C) = 8; (AB) = 6; (BC= 4); (CA) = 4; (CA) = 4; (ABC) = 3

Solution:

There are three attributes A,B,C.

∴ The total number of class frequencies is 3^3=27.

We are given only 8 class frequencies and we have to find the remaining 19 class frequencies. They are

Order 1:

$(\alpha) = N - (A) = \ 20 - 9 = 11.$

(β) = N-(B) = 20-12=8

$(\gamma) = N - (C) = \ 20 - 8 = 12$

Order 2:

$(A\beta) = A(1 - B).N$

= (A) – (B)

= 9-6= 3

$(\alpha B) = (1 - A)B.N$

= (B) – (AB)

= 12-6 = 6

(A$\gamma) = A(1 - C).N$

= (A) – (AC)

= 9-4

= 5

$(\alpha C) = (1 - A)C.N$

=(C) - (AC)

= 8-4=4

(B$\gamma) = B(1 - C)N$

= (B) – (BC)

= 12-4 = 8

$(\beta C) = (1 - B)\ C.N$

= (C) - (BC)

= 8-4=4

$(\alpha\beta) = (1 - A)(1 - B).N = N - (A) - \ (B) + \ (AB)$

= 20-9-12+ 6=5

$(\beta\gamma) = (1 - B)(1 - C).N$

= N – (B) – (C) + (BC)

= 20-12-8+4

= 4

$(\alpha\gamma)\ = (1 - A)(1 - C).N$

= N- (A) – (C) + (AC)

= 20-9-8+4

= 7

Order 3:

$(\mathrm{A}\beta\gamma) = AB(1 - C).N$

= (AB) – (ABC)

= 6-3=3

$(\mathrm{A}\beta C) = A(1 - B)C.N$

= (AC) – (ABC)

= 4-3 = 1

$(\mathrm{A}\beta\gamma) = A\ (1 - B)(1 - C).N$

= (A) – (AC) – (AB) + (ABC)

= 9-4-6+3 = 2

$(\alpha BC) = (1-A)BC.N$

= (BC) – (ABC)

= 4-3=1

$(\alpha B\gamma) = (1-A)(1-C).B.N$

= (B) - (BC) – (AB) + (ABC)

= 12-4-6+3

= 5

$(\alpha\beta C) = (1-A)(1-B)C.N$

=(C) - (AC) – (BC) + (ABC)

= 8-4-4+3=3

$(\alpha\beta\gamma) = (1-A)(1-C).N$

= N-(A)-(B) – (C) + (AB) + (BC) + (CA) – (ABC)

= 20-9-12-8+6+4+4-3 = 2

Problem :

In a class text in which 135 candidates were examined for proficiency in English and Maths. It was discovered that 75 students failed in English, 90 failed in Maths and 50 failed in both. Find how many candidates i) have passed in Maths ii) have passed in English, failed in Maths iii) have passed in both.

Solution:

Let A denote pass in English and B denote pass in Maths .

∴ (α) denotes fail in English and (β) denotes fail in Maths.

Given (α)= 75; $(\beta) = 90; (\alpha\beta) = 50; N = 135$

We have to find (i) (B) (ii) $(A\beta)$ $(iii)(AB)$

i) (B) = N-(β)

= 135-90

= 45

ii) Consider, $(\beta) = (A\beta) + \ (\alpha\beta)$

$$\Rightarrow (A\beta) = (\beta) - \ (\alpha\beta)$$
$$= 90 - 50$$
$$= 40$$

iii) (AB) = (1-α)$(\ 1 - \beta).N$

= N- $(\alpha) - (\beta) + \ (\alpha\beta)$

= 135-75-90 + 50

= 20

Problem :

Given N = 1200; (ABC) = 600; $(\alpha\beta\gamma) = 50; (\gamma) = \ 270;$

$(A\beta) = \ 36; (\beta\gamma) = \ 204; (A) - (\gamma) = \ 192; (B) - \ (\beta) = \ 620.$

Find the remaining ultimate class frequencies .

Solution:

Since there are 3 attributes there are 2^3=8.Ultimate class frequencies we are given two.

Hence we have find the remaining six

They are (i) (ABγ) (ii)(AβC)

$(iii)\ (\alpha BC)\ (iv)(A\beta\gamma)\ \ (v)(\alpha B\gamma) and\ (vi)(\alpha\beta C)$

To find the frequencies of positive classes: (A), (B), (C); (AB), (BC), (AC).

First order: (A) – $(\alpha) = 192$

$(A) + (\alpha) = 1200(= N)$

Adding,

2(A)=1200+192

2(A)=1392

(A)=696

(B) – $(\beta) = 620$

(B) – $(\beta) = 620$ (=N)

Hence (B) = 910

Now, (C) = N- (γ)

= 1200 – 270

= 930.

Second order:

(AB) = (A) – (Aβ) $= 696 - 36$

= 660

(BC) = (B) – (Bγ) = 910 – 204

= 706

We have, N= (A) + (B) + (C) – (AB) – (BC) – (AC) + (ABC) + $(\alpha\beta\gamma)$

(AC) = (A) + (B) + (C) – (AB) – (BC) + (ABC) + $(\alpha\beta\gamma)$

= 696 + 910 + 930 – 660 – 706 + 600 + 50= 620

Third order:

i. (ABγ) $= AB\ (1 - C).N$

= (AB) – (ABC)

= 660 – 600

= 60

ii. (AβC) $= AC(1 - B).N$

= (AC) – (ABC)

= 620 – 600

= 20

iii.$(\alpha BC) = (1 - A)BC.N$

= (BC) – (ABC)

= 706 – 600

= 106

iv.(A$\beta\gamma$) $= A(1 - B)(1 - C).N$

= (A) – (AB) – (AC) + (ABC)

= 696 – 660 – 620 + 600

= 16

v. $(\alpha B\gamma) = (1 - A)\ (1 - C)B.N$

= (B) – (AB) – (BC) + (ABC)

= 910 – 660 – 706+ 600

= 144.

vi.$(\alpha\beta C) = (1 - A)(1 - B)C.N$

= (C) - (AC) – (BC) + (ABC)

= 930 – 620 – 706 + 600 = 204

Problem :

Given that (A) = $(\alpha) = (B) = (\beta) = N/2$

Show that i) (AB) ii) $(\alpha\beta)$ $(ii)(A\beta) = (\alpha B)$

Solution:

i. (AB) = AB.N

= (1-α) $(1-\beta).N$

= N- $(\alpha) - (\beta) + (\alpha\beta)$

= N – N/2 – N/2 + $(\alpha\beta)$

(AB) = $(\alpha\beta)$

ii. (Aβ) $= A\beta.N$

= (1-α) $(1-B).N$

= N – $(\alpha) - (B) + (AB)$

= N – N/2 – N/2 + (αB)

(Aβ) = (αB)

Problem :

Of 500 men in a locality exposed to cholera 172 in all were attacked, 178 were inoculated and of these 128 were attacked. Find the number of persons.

i) not inoculated not attacked

ii) inoculated not attacked

iii) not inoculated attacked

Solution:

Denote the attribute A as attacked and the attribute B as inoculated.

Hence α denote “NOT ATTACKED”; β DENOTES “NOT INOCULATED”.

Given, N= 500; (A) = 172; (B) = 178; (AB) = 128

To find (i) $(\alpha\beta)$ $(ii)(\alpha B)$ $(iii)(A\beta)$

i. $(\alpha\beta) = \alpha\beta.N$

= (1-A) (1-B).N

= N-(A) –(B) + (AB)

= 500 -172-178 + 128

= 278

i. $(\alpha B) = \alpha B.N = (1-A)B.N$

= (B) – (AB)

= 178 – 128 = 50

iii). (A$\beta) = A\beta.N = A(1-B).N$

= (A) – (AB)

= 172 – 128 = 44

Problem:

There were 200 students is a college whose results in the first semester, second semester and the third semester are as follows: 80 passed in the first semester; 75 passes in the second semester. 96 passed in the third semester 25 passed in all the three semester 46 failed in all the three semester 29 passed in the first two and failed in the third semester 42 failed in the first two semester but passed in the third semester. Find how many students passed in atleast two semesters

Solution:

Denoting "pass in first semester" as "A' Pass in second semester 'B' and pass in the third semester as 'C' we get.

N = 200; (A) = 80, (B) = 75 ; (C) = 96

(ABC) = 25; $(\alpha\beta\gamma) = 46; (AB\gamma) = 29; (\alpha\beta C) = 42$

We have to find (ABγ) $+ (\ \alpha BC) + (A\beta C) +\ (ABC)$

Consider, (C) = (AC) + (αC)

= (ABC) + (A$\beta C) + (\alpha BC) +\ (\alpha\beta C)$

$\therefore (ABC) +\ (\alpha BC) +\ (A\beta C) = (C) -\ (\alpha\beta C)$

= 96 – 42 = 54

$\therefore$ (ABC) + $(\alpha BC) + (A\beta C) +\ (AB\gamma) =\ 54 + 29 = 83$

Thus the number of students who passed in atleast two semester is 83.

Problem :

Given (ABC) = 149; (ABγ) $= 738; (A\beta C) =\ 225\ ;\ (A\beta\gamma) = 1196; (\alpha BC) = 204;\ (\alpha B\gamma) = 1762; (\alpha\beta C) =\ 171; (\alpha\beta\gamma) = 21842. find\ (A), (B), (C), (AB), (AC), (BC) and\ N.$

Solution:

N = (ABC) + (ABγ) $+ (A\beta C) +\ (A\beta\gamma) +\ (\alpha BC) +\ (\alpha B\gamma) +\ (\alpha\beta C) +\ (\alpha\beta\gamma)$

= 149 + 738 + 225 + 1196 + 204 + 1762 + 171 + 21842.

= 26287

(A) $= (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) = 149 + 738 + 225 + 1196$

$= 2308$

(B) $= (ABC) + (AB\gamma) + (\alpha BC) + (\alpha B\gamma) = 149 + 738 + 204 + 1762$

$= 2853$

(C) $= 749$

$(AB) = (ABC) + (AB\gamma) = 149 + 738 = 887$

$(AC) = (ABC) + (A\beta C) = 149 + 225 = 374$

$(BC) = (ABC) + (\alpha BC) = 353$

Problem :

In a very hotly fought battle 70% of the solders at least lost an eye 75% at least lost an ear 80% at least an arm and 85% at least lost a leg. How many at least must have lost all the four?

Solution:

Denoting “loosing an eye” A, “loosing a ear by B” “loosing an arm by C” and “loosing a leg by D”

We have

N= 100, $(A) \geq 70, (B) \geq 75, (C) \geq 80, (D) \geq 85.$

To find the least value of ABCD

$(ABCD) \geq (A) + (B) + (C) + (D) - 3N$

$\geq 70 + 75 + 80 + 85 - 300$

$= 10$

$(ABCD) \geq 10$

At least 10% of the soldiers lost all the four.

Problem :

A company producers tube lights and conducts a test on 5000 lights for production defects of frames (F); chokes (C); starters (S) and tubes (T). The following are the records of defects.

$(F) = 130, (C)=120, (S) = 115, (T) = 86$

$(FC) = 100, (CS) = 130, (ST) = 75, (FT) = 60$

$(CT) = 54, (FS) = 37, (FCS) = 90 , (CST) = 85$

$(FST) = 112, (FCT) = 108, (FCST) = 5.$

Find the percentage of the tube lights which pass all the four tests.

Solution:

Number of tube lights passing the four tests

$$= (1-F)\ (1-C)\ (1-S)\ (1-T)\ .N$$

$$= [1-(F+C+S+T) + (FC+CS+ST+FT+CT+FS) - (FCS+CST+STF+FCT) + FCST].N$$

$$= N-[(F)+(C)+(S)+(T)]+ [(FC)+(CS)+(ST) + (FT)+(CT)+(FS)] - [(FCS)+(FCT)+((FST)+(CST)] + (FCST)$$

$$= 5000- (130+20+115+86) + (100+130 +75+60+54+37)-(90+108+112+85) +5$$

$$= 5000-451+456-395+5$$

$$= 5461-846=4615$$

Out of 5000 tube lights 4615 pass the four tests for defects.

Percentage of tube lights which pass the four tests

$= \frac{4615}{5000} \times 100$ = 92.3%

Exercises:

1.Given the frequencies (A) = 1150, $(\alpha) = 1120,\ (AB) = 1075\ (\alpha\beta) = 985. Find\ remaining$ class frequencies and total number of observations.

2. Given the following ultimate class frequencies find the frequencies of the positive and negative classes and the total number of observations.

(AB) = 733, (A$\beta) = 840, (\alpha B) = 699; (\alpha\beta) = 783.$

3. A survey reveals that out of 1000 people in locality 800 like coffee, 700 like tea, 660 like both coffee and tea. Find how many people like neither coffee nor tea.

4. An examination result shows the following data. 56% at least failed in part I Tamil, 76% at least failed in part II English 82% at least failed in major – chemistry and 88% at least failed ancillary maths. How many at least failed in all the four?

5. In a university examination 95% of the candidates passed partI, 70% passed in part II, 65% passed part III. Find how many at least should have passed the whole examination.

Consistency of data:

Definition:

A set of class frequencies is said to the consistent if none of them is negative otherwise the given set of class frequencies is said to be inconsistent.

We have the following set of criteria for testing the consistency in the case of single attributes and three attributes.

Attributes	Condition consistency	Equivalent positive class condition	Number of conditions
A	$(A)\geq 0$ $(\alpha)\geq 0$	$(A)\geq 0$ $(A)\leq N$(Since $(\alpha)=(1-A)N\geq 0$)	2
A,B	$(AB)\geq 0$ $(A\beta)\geq 0$ $(\alpha B)\geq 0$ $(\alpha\beta)\geq 0$	$(AB)\geq 0$ $(AB)\leq A$ $(AB)\leq B$ $(AB)\geq(A)+(B)-N$	2^2
A,B,C	$(ABC)\geq 0$ $(AB\gamma)\geq 0$ $(A\beta C)\geq 0$ $(\alpha BC)\geq 0$ $(A\beta\gamma)\geq 0$ $(\alpha B\gamma)\geq 0$ $(\alpha\beta C)\geq 0$ $(\alpha\beta\gamma)\geq 0$	i) $(ABC)\geq 0$ ii) $(ABC)\leq(AB)$ iii) $(ABC)\leq(AC)$ iv) $(\alpha BC)\leq(BC)$ v)$(ABC)\geq(AB)+(AC)-(A)$ vi)$(ABC)\geq(AB)+(BC)-(B)$ vii)$(ABC)\geq(AC)+(BC)-(C)$ viii)$(ABC)\leq(AB)+(BC)+(AC)-(A)-(C)+(N)$	2^3

Note:

In the case of 3 attributes conditions

(i) and (Viii)

$$\Rightarrow (AB)+(BC)+(AC)\geq (A)+(B)+(C)-N \ldots\ldots (ix)$$

Similarly,

(ii) and (vii)

⇒ (AC) +(BC) – (AB)≤ (C) … … … … … … … … … … … … … .. (x)

(iii) and (vi)

⇒(AB)+(BC)-(AC) ≤(B) ………………………… ….(xi)

iv) and (v)

⇒ (AB) +(AC) – (BC)   (A) ……………………(xii)

conditions (ix) to (xii) can be used to check the consistency of data when the class of first and second order alone are known.

Problem :

Find whether the following data are consistent. N= 600; (A) = 300;(B) = 400; (AB)=50.

Solution:

We calculate the ultimate class frequency $(\alpha\beta), (\alpha B) and\ (A\beta)$

$(\alpha\beta) = \ \alpha\beta . N = (1 - A)(1 - B). N$

= N-(A) –(B) +(AB)

= 600 – 400 +50

= -50

Since $(\alpha\beta) < 0,$ the data are inconsistent.

Problem :

Show that there is some error in the following data: 50% of people are wealthy and healthy 35% are wealthy but not healthy 20% are healthy but not wealthy.

Solution:

Taking “wealth” as A and “health as “B” we get the following data

N=100, (AB) =50; (Aβ) = 35, (αB)=20

To check the consistency of data we find $(\alpha\beta)$

$(\alpha\beta) = \ \alpha\beta.N = (1 - A)(1 - B).N$

= N-(A) – (B) + (AB)

But (A) = (AB) +(Aβ)

= 50+35=85

(B) = (AB) + (αB)

= 50+20

= 70

$(\alpha\beta) = 100 - 85 - 70 + 50$

= -5

$(\alpha\beta)$<0

Hence there is error in the data.

Problem :

Of 2000 people consulted 1854 speak Tamil; 1507 speak Hindi; 572 Speak English; 676 speak Tamil and Hindi; 286 speak Hindi and English; 114 speak Tamil; Hindi and English. Show that the information as it stands is incorrect.

Solution:

Let A,B,C denote the attribution of speaking Tamil, Hindi, English respectively.

Given, N= 2000, (A) = 1854, (B) = 1507 (C) = 572;

(AB)= 676; (AC)= 286, (BC) = 270, (ABC)= 114

Consider $(\alpha\beta\gamma) = \alpha\beta\gamma.N$

=(1-A) (1-B) (1-C).N

=N – (A) – (B) –(C) + (AB) +(BC) +(AC) – (ABC)

=2000 – 1854 – 1507 – 572 +676 + 270 + 286 – 114

=- 815

$$\therefore (\alpha\beta\gamma) < 0.$$

Hence the data are inconsistent.

∴The information is incorrect.

Problem :

Find the limits of (BC) for the following available data.

N = 125, (A) = 48, (B) = 62, (C) = 45

$(\text{A}\beta) = 7 \; and \; (A\gamma) = 18$

Solution:

To find (AB) and (AC)

(AB) = (A) – (Aβ)

= 48-7 = 41

(AC) = (A) – (Aγ)

= 48-18 = 30

Now, by condition of consistency (ix)

(AB) + (BC) +(AC) $\geq (A) + (B) + (C) - N$

41+(BC) + 30 $\geq 48 + 62 + 45 - 125$

$(BC) \geq -41 \dots\dots\dots\dots\dots..(i)$

Also using (xii)

(AB) +(AC)-(BC)$\leq (A)$

⇒(BC) $\geq (AB) + (AC) - (A)$

= 41+30-48= 23

(BC) $\geq 23 \dots\dots\dots\dots\dots\dots\dots\dots\dots.(ii)$

Using (xi), (AB) + (BC) – (AC) $\leq (B)$

⇒(BC) $\leq (B) + (AC) - (AB)$

= 62+ 30-41

=51

∴(BC) $\leq 51 \dots\dots\dots\dots\dots\dots\dots\dots\dots.(iii)$

Using (x), (AC) + (BC) – (AB) $\leq (C)$

$\Rightarrow$(BC) $\leq (C) + (AB) - (AC)$

= 45+41-30

= 56

$\therefore$(BC)= 56 ……………………………..(iv)

From (i), (ii), (iii) and (iv) we get

$23 \leq (BC) \leq 56$

Problem :

Find the greatest and least value of (ABC) if (A)=50, (B)=60, (C)= 80, (AB) = 35, (AC)= 45 and (BC)=42

Solution:

The problem involves 3 attributes and we are given positive class frequencies of first order and second order only.

Using positive class conditions (ii), (iii), (iv) of consistency for 3 attributers

(ABC) $\leq (AB) \Rightarrow (ABC) \leq 35$

(ABC) $\leq (BC) \Rightarrow (ABC) \leq 42$

(ABC) $\leq (AC) \Rightarrow (ABC) \leq 45$

$$\Rightarrow (ABC) \leq 45 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (i)$$

Using (v) (vi) and (vii)

(ABC) $\geq (AB) + (AC) - (A)$

$$\Rightarrow (ABC) \geq 35 + 45 - 50 = 30$$

(ABC) $\geq (AB) + (BC) - \ (B)$

$\Rightarrow$ (ABC) $\geq 35 + 42 - 60 = 17$

(ABC) $\geq (AC) + (BC) - \ (C)$

$\Rightarrow$ (ABC) $\geq 45 + 42 - 80 = 7$

Thus (ABC) ≥ 30

(ABC) ≥ 17

(ABC) ≥ 7

$\Rightarrow$ (ABC) $\geq 30 \ldots \ldots \ldots \ldots \ldots \ldots \ldots (2)$

From (1) and (2) we get 30$\leq (ABC) \leq 35$

∴The least value of (ABC) is 30 and the greatest value of (ABC) is 35.

Problem :

If $\frac{(A)}{N} = x; \quad \frac{(B)}{N} = 2x, \frac{(C)}{N}$= 3x and

$\frac{(AB)}{N} = \quad \frac{(AC)}{N} = \frac{(BC)}{N} = y$. prove that neither $x \; nor \; y$ can exceed ¼ .

Solution:

Clearly $x \; and \; y$ are positive integers. The condition of consistency

$$(AB) \leq (A)$$

$$\Rightarrow \frac{(AB)}{N} \leq \frac{(A)}{N}$$

$$y \leq x$$

Similarly,

(BC)$\leq (B) \ \Rightarrow \ y \leq 2x$

$\Rightarrow$ y$\leq x \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$

Now, (AB)$\geq (A) + (B) - N$

$\Rightarrow \frac{(AB)}{N} \geq \frac{(A)}{N} + \frac{(B)}{N} - 1$

Thus, (AB) $\geq (A) + (B) - N$

y$\geq 3x - 1$

Similarly

(BC) $\geq (B) + (C) - N$

$\Rightarrow$y$\geq 5x - 1$

$\Rightarrow$y$\geq 5x - 1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots. (2)$

(AC) $\geq (A) + (C) - N$

By (1) and (2) 5x -1 $\leq y \leq x$.

Taking 5x-1$\leq x\, weget\, x \leq \frac{1}{4}$

Taking y$\leq x\, weget\, y \;\leq \frac{1}{4}$

Neither $x\, nor\, y$ can exceed ¼.

Exercises:

1. Examine the consistency of data when

i) (A)=800; (B)= 700, (AB)=660; (N)= 1000

ii) (A)=600; (B)= 500, (AB)= 50; N= 1000

iii) N=2100; (A)=1000, (B)=1300; (AB)=1100

iv) N=100; (A)=45; (B)=55, (C) = 50; (AB)=15 , (BC)= 25, (AC)= 20, (ABC)=12

v)N=1800; (A)=850; (B)=780; (C)=326; (AB)=250; (BC)=122; (AC)=144;(ABC)= 50

2. A market investigator returns the following data of 2000 people consulted 1754 liked chocolates 1872 liked toffee and 572 liked biscuits, 678 liked chocolate and coffee, 236 liked chocolates and biscuits, 270 liked chocolates and biscuits, 270 liked toffee and biscuits, 114 liked all the three .Show that the information it started must be incorrect.

3. If (A) = 50; (B)= 60; (C)=50; (A$\beta) = 5$;

$(A\gamma) = 20\ and\ N = 100.$ Find the least and greatest value of (BC).

Independence and Association of Data:

Two attributes A and B are said to be independent if there is same proportion of A's amongst B as amongst β's.

Thus A and B are independent iff

$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$……………………………….(i)

or

$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$……………………………….(ii)

From (i) we get

$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} = \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{N}$

∴(AB) = $\frac{(A)(B)}{N}$………………….(1)

And (A$\beta) = \frac{(A)(\beta)}{N}$ …………………(2)

Again from (1) we get

$1-\frac{(AB)}{(B)} = 1-\frac{(A\beta)}{(\beta)}$

$\frac{(B)-(AB)}{(B)} = \frac{(\beta)-(A\beta)}{(\beta)}$

$\frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}$

$\therefore \frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}$

$= \frac{(\alpha\beta)+(\alpha B)}{(\beta)+(B)}$

$= \frac{(\alpha)}{N}$

$(\alpha\beta) = \frac{(\alpha)(\beta)}{N}$……………………….(3)

And $(\alpha B) = \frac{(\alpha)(B)}{N}$…………………………….(4)

(1),(2),(3),(4) are all equivalent conditions for independent of the attribute A and B.

Association and Coefficient of Association:

If (AB) $\neq \frac{(A)(B)}{N}$ we say that A and B are associated. There are two possibilities.

If (AB) > $\frac{(A)(B)}{N}$ we say that A and B are positively associated and If (AB) < $\frac{(A)(B)}{N}$ we say that A and B are negatively associated.

Let us denote $\delta = (AB) - \frac{(A)(B)}{N}$

ie. $\delta = \frac{1}{N}$ [(AB) $(\alpha\beta) - (A\beta)(\alpha\beta)$]

Note:

i. A and B are independent if $\delta = 0$.

ii. A and B are positively associated if $\delta > 0$ and negatively associated if $\delta < 0$.

Coefficient of association:

There are several measures indicating the intensitivity of association between two attribution

A and B.

The most commonly used measures are the Yule's coeficiency of association Q and coefficient of colligation Y which are defined as follows.

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$Q = \frac{N\delta}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$Y = \frac{\left[1 - \sqrt{\left\{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}\right\}}\right]}{\left[1 + \sqrt{\left\{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}\right\}}\right]}$$

Problem :

Check whether the attributes A and B are independent given that (i) = 30 (B)= 60, (AB)= 12, N= 150

(ii)(AB) = 256, $(\alpha B) = 768, (A\beta) = 48\ , (\alpha\beta) = 144.$

Solution:

Given class frequencies are of first order condition for independence is

(AB) = $\frac{(A)(B)}{N}$

Consider,

$= \frac{(A)(B)}{N} == \frac{30\times60}{150} = 12 = (AB)$

$\therefore$ (AB)= $\frac{(A)(B)}{N}$

Hence A and B are independent.

ii) (A) = (AB) + (Aβ) $= 256 + 48 = 304$

(B) = (AB) + $(\alpha B\ ) = 256 + 768 = 1024$

$(\alpha) = (\alpha B) + \ (\alpha\beta) = \ 768 + 144 = 912$

$(\beta) = (A\beta) + \ (\alpha\beta) = \ 48 + 144 = 192$

N = (A) + $(\alpha) = 304 + 912 = 1216$

Now = $\frac{(A)(B)}{N} = \frac{304 \times 1024}{1216} = 256 = (AB)$

$\therefore$ (AB) = $\frac{(A)(B)}{N}$

Hence A and B are independent.

Problem :

In a class test in which 135 candidates were examined for proficiency in physics and chemistry, it was discovered that 75 students failed in physics, 90 failed in chemistry and 50 failed in both. Find the magnitude of association and state if there is any association between failing in physics and chemistry.

Solution:

Denoting “fail in Physics” as A and “fail in Chemistry” as B we get

(A) = 75, (B) = 90, (AB) = 50, N= 135

The magnitude of association is measured by

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)\beta(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$(\alpha) = N - (A) = 135 - 75 = 60$$

$$(\beta) = N - (B) = 135 - 90 = 45$$

$$(\alpha B) = (B) - (AB) = 90 - 50 = 40$$

$$(\mathrm{A}\beta) = (A) - (AB) = 75 - 50 = 25$$

$$(\alpha\beta) = (\alpha) - (\alpha B) = 60 - 40 = 20$$

$$Q = \frac{50\times20-25\times40}{50\times20+20\times40}$$

$$Q = 0$$

∴ A and B are independent hence failure in physics and chemistry are completely independent of each other.

Problem :

Show whether A and B are independent or positively associated or negatively associated in the following cases.

i) N = 930, (A) = 300, (B) = 400, (AB) = 230

ii) (AB) = 327, $(\mathrm{A}\beta) = 545, (\alpha B) = 741, (\alpha\beta) = 235$

iii) (A) = 470, (AB) = 300, $(\alpha) = 530, (\alpha B) = 150$

iv. (AB) = 66, $(\mathrm{A}\beta) = 88, (\alpha B) = 102; (\alpha\beta) = 136$

Solution:

i) $\frac{(A)(B)}{N} = \frac{300 \times 400}{930} = 129.03$

Now, $\delta = (AB) - \frac{(A)(B)}{N}$

= 230 – 129.03

= 100.97

Here $\delta > 0$

Hence A and B are positively associated.

ii) $Q \;=\; \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

$= \frac{327\times235-545\times741}{327\times235+545\times741}$

$= \frac{76845-4038845}{76845+403845}$

$= \frac{-32700}{480690}$

= -0.6803

$Q < 0$.

Hence A and B are negatively associated.

iii) N= (A) + (α)

= 470 + 530

= 1000

(A)= (AB) + (αB)

= 300 + 150

= 450

Now, $\frac{(A)(B)}{N} = \frac{470\times450}{1000}$ = 2115

$\therefore \delta = (AB) - \frac{(A)(B)}{N}$

= 300- 2155

= - 1825

$\therefore \delta < 0.$

$Hence\ A\ and\ B\ are\ negatively\ associated.$

iv. $Q\ =\ \frac{(AB)(\alpha\beta\,) - (A\beta\,)(\alpha B\,)}{(AB)(\alpha\beta\,) - (A\beta\,)(\alpha B\,)}$

$= \frac{66\times136-88\times102}{66\times136+88\times102}$ = 0.∴ A and B are independent.

Problem :

Calculate the co-efficient of associate between intelligence of father and son from the following data.

Intelligent father with intelligent sons 200.Intelligent fathers with dull sons 50. Dull fathers with intelligence sons 110. Dull fathers with dull sons 600. Comment on the result.

Solution:

Denoting the “intelligence of fathers” as A and intelligence of sons” by B

we have

(AB) = 200, (Aβ) $= 50$, $(\alpha B) =\ 110, (\alpha\beta) = 600$

$$Q = \frac{(AB)(\alpha\beta) -\ (A\beta)(\alpha B)}{(AB)(\alpha\beta) +\ (A\beta)(\alpha B)}$$

$$= \frac{200\times 600-50\times 110}{200\ \times 600+50\times 110}$$

= 0.91235

Since Q is positive it means that intelligent fathers are likely to have intelligent sons.

Problem :

Investigate from the following data between inoculations against small pox prevention from attack.

	Attacked	Not attacked	Total
Inoculated	25	220	245
Not inoculated	90	160	250
Total	115	380	495

Solution:

Denoting A as “inoculated” and B as “attacked” we have (AB)= 25, (A$\beta) =$ $220, (\alpha B) = 90\ and$

$(\alpha\beta) = 160.$

$$Q = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$= \frac{25\ X\ 160-220\ X\ 90}{25\ X\ 160+220\ X\ 90}$$

$$= \frac{400-19800}{400+19800}$$

$$= \frac{-15800}{23800}$$

= -0.6638.

Attributes A and B have negative association.

i.e. “Inoculation” and “attack from small pox” are negatively associated.

Thus inoculation against small pox can be taken as the preventive measure.

Problem :

From the following data compare the association between marks in physics and chemistry in MKU and MSU

University	MSU	MKU
Total number of candidate	200	1600
Pass in physics	80	320
Pass in chemistry	40	90
Pass in physics and chemistry	20	30

Solution:

Denoting “pass in physics” as A and “pass in chemistry” as B.

We have,

MKU	MSU
N=1600	N=200
(A) = 320	(A)=80
(A)= 90	(A)= 40
(AB) = 30	(AB) = 20

From the above data we get the rest of the class frequencies for MKU and MSU.

MKU	MSU
$(\mathrm{A}\beta) = (A) - \ (AB)$ = 320 – 30 = 290 $(\alpha B) = (B) - \ (AB)$ = 90-30 = 60 $(\alpha\beta) = N - (A) - (B) + \ (AB)$ = 1600- 320 – 90 +30 = 1220	$(\mathrm{A}\beta) = (A) - \ (AB)$ = 80 – 20 = 60 $(\alpha B) = (B) - \ (AB)$ = 40-20 = 20 $(\alpha\beta) = N - (A) - \ (B) + (AB)$ = 200 – 80 – 40 + 20 = 100

We now find the coefficient of association between A and B for MKU and MSU

	Passed	Failed	Total
Married	90	65	155
Unmarried	260	110	370
Total	350	175	525

3. From the figures given in the following table compare the association between literacy and un employment in rural and urban areas- and given reasons for the difference if any

	Urban	Rural
Total adult males	25 lakhs	200 lakhs
Literate males	10 lakhs	40 lakhs
Unemployed males	5 lakhs	12 lakhs
Literate and unemployed males	3 lakhs	4 lakhs

Time series:

Definition:

Time series is a series of values of a variable over a period of time arranged chronologically

Components of a time series:

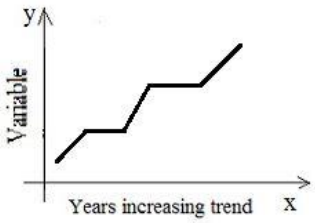
The various forces affecting the values of a phenomenon in a time series may be broadly classified into the following three categories generally known as the components of a time series.

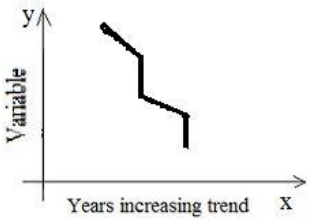
1. Longtime trend (or) secular trend
2. Short term fluctuations (or) periodic movements
3. Irregular fluctuations

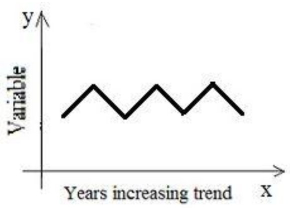
1. Long time trend:

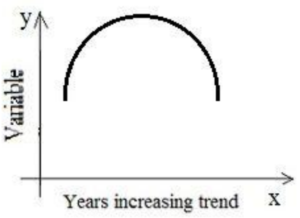
The general tendency of a time series is to increase or decrease or stagnate over a period of several years. Such a long run tendency of a time series to increase or decrease over a period of time is known as secular trend or simply trend. Though the term “long” is a relative term it depends upon the nature of the series under consideration.

The long term trend does not mean that the series should continuously move in one direction only. It is possible that different tendencies of increases and decrease persist together. A graphical representation indicating a long term increase or decrease or stability is given is the following figures.









2. Short term fluctuations:

In most of the time series a number of forces repeat themselves periodically over a period of time preventing the values of the series to move in a particular direction. The variations caused by such forces are called short term fluctuations. This short term fluctuations may broadly be classified into (a) seasonal variation (B) cyclical variation

a) Seasonal variation:

Generally seasonal variations are considered as short term fluctuations that occur within a year. These fluctuations may be regular as well as irregular with in a period of one year

b) Cyclical variation

If the period of oscillation for the periodic movements is a time series is greater than one year then it is called cyclical variation. Generally oscillatory movement is nay business activity is due to the out time of the business cycles normally having four phases namely prosperity recession, depression and recovery. The period between two successive peaks or though is known as the period of the cycle. In cyclical variation generally the period of a cycle is three to eleven years.

3. Irregular fluctuations

The fluctuation which are purely random and due to unforeseen and unpredictable forces are called Irregular fluctuations

Measurement of trends

A graphical representation of a time series exhibits the general upwards and downward tendencies

The following are the four study of measurement of the trend in a time series

i) Graphic method
ii) Method of curve fitting by the principles of least squares.
iii) Method of semi averages
iv)Method of moving averages.

i) Graphic Method

This is the simplest method of determining the trend. In this method all values of the time series are plotted on a graph paper and a smooth curve is drawn by free hand to pass through as many points as possible. The smoothing of the curve eliminates the other components such as seasonal, cyclic and random variations.

ii) The method of curve fitting:

This is the best method of fitting a trend and it is commonly used in practice.

iii) Method of semi averages

In this method the whole time series data is classified into two equal parts with respect to time. Having divided the given series into two equal parts we calculate the arithmetic mean for each part. These means are called semi-averages. Then these average are plotted against the mid values of the respective period covered by each part. The line joining these points give the straight line trend for the time series.

iv) Method of moving averages

This method for measuring the trend consists of obtaining a series of moving average of successive m terms of the time series. This averaging process smoothens the fluctuations and the UPS and down in the given data. It has been observed and proved mathematically that if a trend is liner the period of the moving average is taken to be the period of oscillation.

Measurement for seasonal variation

There is a simple method for measuring the seasonal variation which involves simple averages.

Simple average method

Step 1:

All the data are arranged by years and months.

Step 2:

Compute the simple average $\bar{x}_{\mathrm{i}}$ for ith months

Step 3:

Obtain the overall average x of these average $\bar{x}_i$ and

$$\bar{x} = \frac{\overline{x1}+\bar{x}2+\cdots\ldots\ldots+\bar{x}12}{12}$$

Step4:

Seasonal indices for different months are calculated by expressing monthly average as the percentage of the overall average x

Thus seasonal index for ith month $= \frac{\bar{x}i}{\bar{x}} \times 100$ Take X = x – 1987 and Y = y-42

Then the line of best fit become

Y=ax + b

The normal equations are $\sum xy = a\sum x^2 + b\sum x$

$$\sum y = a\sum x + nb, where\ n = 11$$

From the table,

-19 = 110 a

⇒a = $\frac{-19}{110}$ = - 0.17

17 = 11b

⇒b$\frac{17}{11}$ = 1.55

∴ The line of best fit is Y= - 0.17 x + 1.55

ie.$Y - 42 = -0.17\ (x - 1987) + 1.55$

$$y = -0.17x + 1987 \times 0.17 + 1.55 + 42$$

$$y = -0.17x + 381.34\ is\ the\ straight\ line\ trend$$

Problem:

Use the method least squares and fit a straight line trend to the following data given from 82 to 92. Hence estimate the trend values for 1993.

Year	82	83	84	85	86	87	88	89	90	91	92
Production in quintals	45	46	44	47	42	41	39	42	45	40	48

Solution:

Let the line of best fit be

Y=ax + b Take X = x – 1987 and Y = y-42

Then the line of best fit become

Y=ax + b

The normal equations are $\sum xy = a\sum x^2 + b\sum x$

$$\sum y = a\sum x + nb, where\ n = 11$$

From the table, -19 = 110 a ⇒a = $\frac{-19}{110}$ = - 0.17

X	X= x-1987	Y	Y= y-42	XY	X^2
1982	-5	45	3	-15	25
1983	-4	46	4	-16	16
1984	-3	44	2	-6	9
1985	-2	47	5	-10	4
1986	-1	42	0	0	1

1987	0	41	-1	0	0
1988	1	39	-3	-3	1
1989	2	42	0	0	4
1990	3	45	3	9	9
1991	4	40	-2	-8	18
1992	5	48	6	30	25
1993	0	-	17	-19	110

$$17 = 11b \quad \Rightarrow b\frac{17}{11} = 1.55$$

∴ The line of best fit is Y= - 0.17 x + 1.55

ie. $Y - 42 = -0.17\,(x - 1987) + 1.55$

$$y = -0.17x + 1987 \times 0.17 + 1.55 + 42$$

$$y = -0.17x + 381.34 \; is \; the \; straight \; line \; trend$$

From the line trend

When x =1982, y=44.4

X = 1983, y=44.23, x=1984, y=44.06

X= 1985, y=43.89, x= 1986, y=43.72

X=1987, y= 43.55, x=1988, y=43.38 X=1989, y=43.21, x=1990, y= 43.04

X = 1991, y = 42.87, x= 1992, y=42.7

Thus the trend values are 44.4, 44.23, 44.06, 43.89, 43.72, 43.58, 43.38, 43.21, 43.04, 43.04, 42.87, 42.7

Problem:

Calculate the seasonal variation indices from the following data

Month	Monthly sales in lakhs				Total	$\bar{x}_i$	Seasonal indices $\frac{\bar{x}_i}{\bar{x}}$x 100
	I 1991	II 1992	III 1993	IV 1994			
January	10	11	11.5	13.5	46	11.5	$\frac{11.5}{12}$x100=95.8
February	8.5	8.5	9	10	36	9	$\frac{9}{12}$x100 = 75
March	10.5	12	11	12.5	46	11.5	$\frac{11.5}{12}$x100 = 95.8
April	12	14	16	18	60	15	$\frac{15}{12}$x100 = 125
May	10	9	12	15	46	11.5	$\frac{11.5}{12}$x100 = 95.8
June	10.5	10.5	11	14	46	11.5	$\frac{11.5}{12}$x100 = 95.8
July	12	14	13	17	56	14	$\frac{14}{12}$x100 = 116.7
August	9	8	11	16	44	11	$\frac{11}{12}$x100 = 91.7
September	11	11	12.5	13.5	48	12	$\frac{12}{12}$x100 = 100
October	10	9.5	11.5	13	44	11	$\frac{11}{12}$x100 = 91.7
November	11	12.5	10.5	14	48	12	$\frac{12}{12}$x100 = 100
December	12	13	15	16	56	14	$\frac{14}{12}$x100 = 116.7
Total						$\frac{144}{12}$	

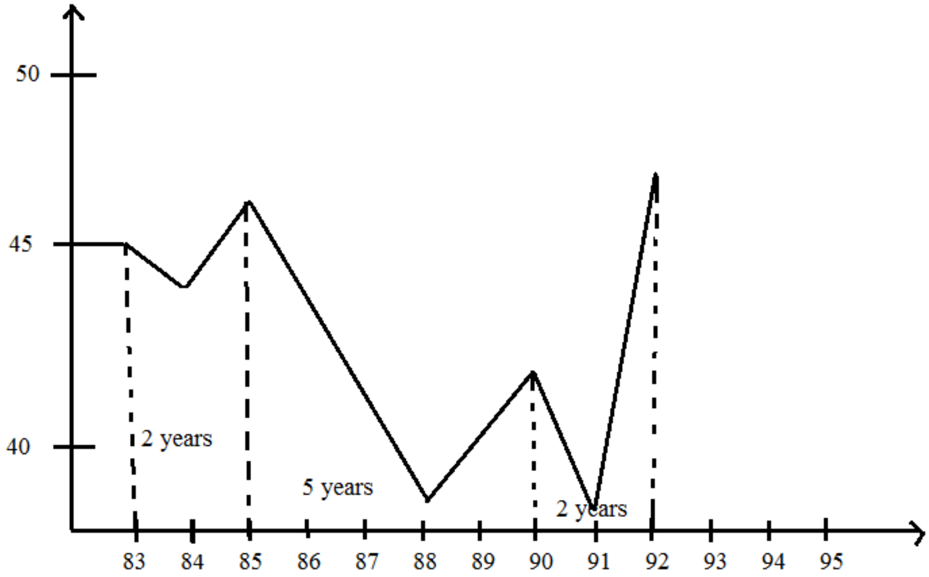
I year	II production in quintals	III 4 yearly moving total	IV 4 yearly moving average	V 2 period moving total	VI trend values (V)/ 2
1982	45	-	-	-	-
1983	46	-	-	-	-
1984	44	182	45.50	90.25	-
1985	47	179	44.75	88.25	45.13
1986	42	174	43.50	85.75	44.13
1987	41	169	42.25	83.25	42.88
1988	39	164	41.00	82.75	41.63
1989	42	167	41.75	83.25	41.38
1990	45	166	41.50	85.85	41.63
1991	40	175	43.75	-	42.93
1992	48	-	-		-

Problem:

Compute the trend values by the method of A yearly moving average for the data given in problem 1.

Problem:

Determine the suitable period of moving average for the data given in problem 1



We observe that the data has peaks at the following years 1983, 1985, 1985, 1990 and 1992.

Thus the data shows 3 cycles with varying periods 2,5,2 respectively.

Hence the suitable period of moving average is taken to be the A.M.

periods.

Hence $\frac{2+5+2}{5}$ = 3 is the period of moving average.

Problem:

Compute the seasonal indices for the following data by simple average method

Princes in different season	Season	1990	1991	1992	1993	1994
	Summer	68	70	68	65	60
	Monson	60	58	63	56	55
	Autumn	61	56	68	56	55
	winter	63	60	67	55	58

Solution:

Year	Summer	Monsoon	Autumn	Winter	Total
1990	68	60	61	63	
1991	70	58	56	60	
1992	68	63	68	67	
1993	65	56	56	55	
1994	60	55	55	58	
total	331	292	296	303	
average	66.2	58.4	59.2	60.6	244.4
Seasonal index	$\frac{66.2}{61.1}$x100 = 108.3	$\frac{58.4}{61.1}$x100 = 95.6	$\frac{59.2}{61.1}$x100 = 69.9	$\frac{60.6}{61.1}$x100 = 99.2	$\bar{x}$ = 61.1

Exercises:

1. Room the data given below calculate the seasonal indicates assuming that trend is absent

Year	I quarter	II quarter	III quarter	IV quarter
1990	40	35	38	40
1991	42	37	39	38
1992	41	35	38	40
1993	45	36	36	41
1994	44	38	38	42

2. Compute the seasonal index for the following data assuming that there is no need to adjust the data for the trend

Quarter	1989	1990	1991	1992	1993	1994
I	3.5	3.5	3.5	4.0	4.1	4.2
II	3.9	4.1	3.9	4.6	4.4	4.6
III	3.4	3.7	3.7	3.8	4.2	4.3
IV	3.6	4.8	4.0	4.5	4.5	4.7

UNIT IV: SAMPLING

Sampling - Definition - Large samples. Small samples - Population with one samples and population with two samples - Students – t – test - Applications - chi - square test and goodness of fit - applications.

TESTS OF SIGNIFICANCE (Large sample)

INTRODUCTION:

Any statistical investigation usually deals with the study of some characteristics of a collection of objects

SAMPLING:

Definition:

A finite subset of population is called a sample and the number of objects in a sample is called the sample size.

Some of the important types of sampling are (i) purposive sampling (ii) Random sampling (iii) Simple sampling (iv) stratified sampling.

(i)Purposive sampling:

If the sample elements are selected with a definite purpose in mind then the sample selected is called purposive sample.

(ii) Random sampling:

A random sample is one in which each element of the population has an equal chance of inclusion in the sample.

(iii) Simple sampling:

Simple sampling is a special type of random sampling in which each element of the population has an equal and independent chance of being included in the sample.

(iv) Stratified sampling:

The sample which is the aggregate of the sampled individuals of each stratum is called stratified sample and the technique of selecting such sample is called stratified sampling.

TESTS OF SIGNIFICANCE FOR LARGE SAMPLES

I. Tests for proportion or percentage

(A) Single proportion (B) Difference of proportions.

II. Tests for means.

(A) (i) Test for single mean if standard deviation of the population σ is known. (ie) $H_{0:}\mu = \mu_0$, σ is known.

(B) (ii) Tests for single mean if σ is not known $H_0: \mu = \mu_0$,σ is unknown.

(C) (i) Test for equality of means of 2 normal populations with Known standard deviations (ie) $H_0: \mu_1 = \mu_2; \sigma_1, \sigma_2$ is known.

(ii) Test for equality of means of 2 normal populations with same standard deviation though unknown $H_0: \mu_1 = \mu_{2,}\ \ \sigma_1 = \sigma = \sigma_2$.

III. Test for standard deviations.

(A) : Test for single standard deviation $H_0: \sigma = \sigma_0$

(B) : Test for equality for 2 standard deviation (ie)$H_0: \sigma_1 = \sigma_2$

1. Test of significance for proportions and percentages.

I(A) **Single proportions.**

If X is the number success in independent trials with constant probability of success **P** for each trial we have E(X)=nP

and V(X)=variance(X)=nPQ where Q=1-P. It has been proved that for large n,

the binomial distribution tends to a normal distribution. Hence for large n,

$$X \ \ \sim N(np, nPQ)$$

$$\therefore Z = \frac{X - E(X)}{S.E\ of\ (X)} = \frac{X - nP}{\sqrt{nPQ}} \sim N(0,1)$$

I (B)Difference of proportions.

Suppose we want to compare 2 distinct populations with regard to possession of an attributes. Let a sample of size $n_1$be chosen from the first population and sample of size $n_2$be chosen from the second population.

Let X_1be number of persons possessing the attribute A in the first sample and X_2 be the number of persons possessing the same attribute in the second sample

$$p_1 = \frac{X_1}{n_1}; p_2 = \frac{X_2}{n_2}$$

As before $E(p_1) = P_1$and $E(p_2) = P_2$ where $P_1\ and\ \ \ P_2$ are the proportions in the populations.$V(p_1) = \frac{P_1Q_1}{n_1}\ and\ V(p_2) = \frac{P_2Q_2}{n_2}$.

$$\therefore Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\ \sim\ N(0,1)$$

An unbiased estimate of population proportion P based on both the samples is given by $P = \frac{(n_1p_1+n_2p_2)}{n_1+n_2}$.Suppose the population proportion*s* $P_1 and P_2$ are given to be different (ie)$P_1 \neq P_2$.

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1)$$

Problem:

A coin is tossed 144 times and a person gets 80 heads. Can we say that the coin is unbiased one?

Solution:

Set the null hypothesis H_0: the coin is unbiased. Given n=144.

Probability of getting a head in a toss P=1/2.HenceQ=1/2.Let X=number of successes=number of getting heads=80.

$$Z = \frac{80 - 144\left(\frac{1}{2}\right)}{\sqrt{144\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = \frac{80 - 72}{\sqrt{36}} = \frac{8}{6} = 1.33 < 1.96.$$

Since │Z│<1.96. We accept the hypothesis at 5% level of significance.

Hence the coin is unbiased.

Problem:

A die is thrown 10000 times and a throw of 1 or 2 was obtained 4200 times. On the assumption of random throwing do the data indicate an unbiased die?

Solution:

P=Probability of getting 1 or 2=1/3 .Hence Q=2/3

Given n=10000,X=4200.The null hypothesis H_0:the die is unbiased.

$$\therefore Z=\frac{X-nP}{\sqrt{nPQ}}=\frac{4200-10000\left(\frac{1}{3}\right)}{\sqrt{10000(\frac{2}{9})}}=\frac{4200-3333.3}{47.14}=\frac{866.7}{47.14}=18.4$$

Since │z│>3, $H_0$is rejected and hence the die is biased one.

Problem:

A manufacturer claimed that at least 95% of the equipment which he supplied to a factory conformed to specification. An examination of a sample of 200 pieces of

equipment revealed that 18 were faulty. Test his claim at a significant level of (i)5% (ii)1%.

Solution:

Out of a sample of 200 equipments 18 were faulty.

X=200-80=182

$$p=\frac{182}{200}=0.91.$$

Set the null hypothesis H_0:P=0.95,Q=0.05.$H_1: P < 0.95$.

$$Z=\frac{p-P}{\sqrt{\frac{PQ}{n}}}=\frac{.91-.95}{\sqrt{\frac{.95\times .05}{200}}}=-\frac{.04}{.0154}=-2.6$$

(i) Since the alternative hypothesis is left tailed and the significant value of Z at 5% level of significant for left tail is -1.645.

Z=-2.6<-1.645.

Hence we accept the null hypothesis at 5% level of significance.

(ii) The critical value of Z at 1% value of significance for left tailed test is -2.33 and Z=-2.6<-2.33.Hence H_0 is accepted at 1% level.

Problem:

A sample of 1000 products from a factory are examined and found to be 2.5% defective. Another sample of 1500 similar products from another factory are found to have only 2% defective. Can we conclude that the products of the first factory are inferior to those of the second?

Solution:

Given $n_1 = 1000, n_2 = 1500.$ Proportion of defectives in the first factory

$p_1 = {}^{25}/_{1000} = .025 \quad p_2 = {}^{30}/_{1500} = .020$

Proportion of defective in the second factory $p_2 = {}^{30}/_{1500} = .020$

$$\therefore P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{25 + 30}{1000 + 1500} = .022$$

Hence Q=1-.022=.978.

Null hypothesis $H_0: P_1 = P_2$

Test hypothesis $Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0,1)$

$$Z = \frac{.025 - .020}{\sqrt{.022(.978)\left(\frac{1}{1000} + \frac{1}{1500}\right)}} = \frac{.005}{\sqrt{.022(.978)/_{600}}} = .83 < 1.9$$

The difference of proportion is not significant on 5% level. Hence this hypothesis is accepted and the two factories are producing similar products. Hence one is not inferior to the other.

Problem:

A machine puts out 16 imperfect articles in a sample of 500 articles. After the machine overhauled it. Puts out 3 defective articles in sample of 100.Has the machine improved?

Solution:

Given $n_1 = 500, n_2 = 100. p_1$ =Proportion defectatives in the first sample=16/500=.032

$$p_2 = \frac{3}{100} = .03$$

Set the null hypothesis $H_0 : P_1 = P_2$

Alternative hypothesis $H_1 : P_1 > P_2$

$$\therefore P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{16 + 3}{500 + 100} = \frac{19}{600} = .032\,; Q = 1 - .032 = .968$$

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{.032 - .030}{\sqrt{.032 \times .968\left(\frac{1}{500} + \frac{1}{100}\right)}} = \frac{.002}{\sqrt{.032 \times \left({}^{6}\!/_{500}\right)}} = \frac{.002}{.019}$$

$$= .105.$$

Sine Z<1.645 it is not significant at 5% level of significance. Hence we can accept the null hypothesis and conclude that the machine has not been improved.

II (B) Test of significance for difference of sample means.

Consider two different normal populations with μ_1 and μ_2 and s.d σ_1 and σ_2 respectively. Let a sample of size n_1 be drawn from the first population and an independent sample of size n_2 be drawn from the second population. Let $\overline{x_1}$ be the mean of the first sample from the first population and $\overline{x_2}$ be the mean of second sample from the second population. If the sample sizes are large we know $\overline{x_1}$ is a normal variate with mean μ_1 and variance $\frac{\sigma_1^2}{n_1}$ and $\overline{x_2}$ is an independent normal variate and normal variate with mean μ_2 and variance $\frac{\sigma_2^2}{n_2}$.

The test statistic becomes $Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ which can be tested at any level of significance.

Problem:

The number of accidents per day were studied for 144 days in Madras city and for 100 days in Delhi city. The mean numbers of accidents and the s.ds were respectively 4.5 and 1.2 for Madras city and 5.4 and 1.5 for Delhi city. Is Madras city more prone to accidents than Delhi city?

Solution:

Given $n_1 = 144\,; \bar{x}_1 = 4.5; \overline{x_2} = 5.4.$

$n_2 = 100; \sigma_1 = 1.2; \sigma_2 = 1.5.$

Set the null hypothesis $H_0 : \mu_1 = \mu_2$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}} = \frac{4.5 - 5.4}{\left({1.2^2}/{144}\right) + \left({1.5^2}/{100}\right)} = -4.99.$$

$\therefore \left|Z\right| = 4.99 > 3$ we reject the hypothesis that the two cities have the same accident rates. However since Delhi city has higher rate of accident than Madras city. Therefore Delhi more prone to accidents.

Problem:

The mean yields of rice from two places in a district were 210 kgs and 220 kgs per acre from 100 acres and 150 acres respectively. Can it be regarded that the sample were drawn from the same district which has the s.d of 11kgs per acre?

Solution:

$$n_1 = 100; \bar{x}_1 = 210;\ \sigma = 11$$

$$n_2 = 150; \bar{x}_2 = 220$$

Set the null hypothesis $H_0: \mu_1 = \mu_2$

$$\therefore Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\left({1}/{n_1}\right) + \left({1}/{n_2}\right)}}$$

$$Z = \frac{210 - 220}{11\sqrt{\left({1}/{100}\right) + \left({1}/{150}\right)}} = \frac{-10}{11\sqrt{{250}/{15000}}} = -7.04$$

$|Z| = 7.04 > 3$.The value is highly significant and hence we reject the null hypothesis. Hence the samples are certainly not from the same district with the s.d 11.

Test of significance for equality of standard deviations of a normal population.

If we want to test whether the two independent samples with known standard deviations $s_1\ and\ \ s_2$ have come from the same population with standard deviation σ. Under the hypothesis $H_0: \sigma_1 = \sigma_2$ the test statistics is $Z = \frac{s_1 - s_2}{\sigma\sqrt{\left({1}/{2n_1}\right) + \left({1}/{2n_2}\right)}}$.

Problem:

The s.d of weight of all students in a first grade college was found to be 4 kgs. Two samples are drawn. The s.ds of the weight of 100 undergraduate students is 3.5kgs and 50 post graduate students are 3 kgs. Test the significance of the difference of standard deviations of the samples at 5% level.

Solution:

Given $n_1 = 100; s_1 = 3.5; \sigma = 4; n_2 = 50; s_2 = 3.$

Set the null hypothesis $H_0: \sigma_1 = \sigma_{2.}$ Then $H_1: \sigma_1 \neq \sigma_2$

$$\therefore Z = \frac{s_1 - s_2}{\sigma\sqrt{\left(1/_{2n_1}\right) + \left(1/_{2n_2}\right)}} = \frac{3.5 - 3}{4\sqrt{\left(1/_{200}\right) + \left(1/_{100}\right)}} = 1.02$$

$|Z|$ =1.02<1.96. It is not significant at 5% level of significance.

Problem:

The mean production of wheat of a sample of 100 plots is 200kgs per acre with s.d of 10 kgs. Another sample of 150 plots gives the mean production of wheat as 220kgs. With s.d of 12kgs. Assuming the s.d of the 11kgs for the universe find at 1% level of significance ,whether two results are consistent.

Solution:

Given σ=11 and

	size	mean	S.D
Sample 1	$n_1 = 100$	$\bar{x}_1 = 200$	$s_1 = 10$
Sample 2	$n_2 = 150$	$\bar{x}_2 = 220$	$s_2 = 12$

Set the null the hypothesis $H_0: \mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$ .For $H_0: \mu_1 = \mu_2$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\left(1/n_1\right) + \left(1/n_2\right)}} = \frac{200 - 220}{11\sqrt{\left(1/100\right) + \left(1/150\right)}} = \frac{-20}{11\sqrt{10/600}} = \frac{-155}{11} = -14.1$$

$\therefore \left|Z\right|$ =14.1>3.Hence the two means differ significantly at 5% level even a 1% level.

For $H_0: \sigma_1 = \sigma_2$.

$$Z = \frac{s_1 - s_2}{\sigma\sqrt{\left(1/2n_1\right)+\left(1/2n_2\right)}} = \frac{10-12}{11\sqrt{\left(1/200\right)+\left(1/300\right)}} = -1.99$$

$\therefore \left|Z\right| = 1.99 > 1.96$ and $\left|Z\right| = 1.99 < 2.58$.

Hence the difference of s.d is significant at 5% level and not significant and 1% level.

∴At 1% level the difference between s.d is not significant but between means it is significant. Hence we can conclude that at 1% level the two results are not consistent.

TEST OF SIGNIFICANCE (SMALL SAMPLES)

TEST OF SIGNIFICANCE BASED ON t-DISTRIBUTION (t-test)

Consider a normal population with mean μ and s.d σ . Let $x_1, x_2, \ldots x_{n.}$ be a random sample of size n with mean $\bar{x}$ and standard deviation s. We know that $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$is the standard normal variate N(0,1).

Hence the test statistics is in small sample becomes

$$Z = \frac{\bar{x}-\mu}{\left(s\sqrt{n/n-1}\right)/\sqrt{n}} = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}}.$$ Now let us define $t = \frac{\bar{x}-\mu}{s/\sqrt{n-1}}$.

This follows students's t-distribution with n-1 degrees of freedom

1.Test for the difference between the mean of a sample and that of a population

Under the null hypothesis $H_0: \mu = \bar{x}$.

The test statistic

$t = \frac{\bar{x}-\mu}{s/\sqrt{n-1}} \sim t_{n-1.}$ Which can be tested at any level of significance with n-1 degrees of freedom.

II. Test for the difference between the means of two samples

II.A. If $\bar{x}_1$and $\bar{x}_2$ are the means of two independent samples of sizes $n_1$and n_2 from a normal population with mean μ and standard deviation σ. It found that $\frac{\bar{x}_1-\bar{x}_2}{\sigma\sqrt{(1/n_1)+(1/n_2)}} \sim N(0,1)$.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Which follows a t-distribution with d.f $v = (n_1 + n_2 - 2)$.

II.B. suppose the sample sizes are equal (ie) $n_1 = n_2 = n$.Then we have n pairs of values. Further we assume that the n pair are independent .Then the test statistic t in (1) becomes $t = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{n(s_1^2+s_2^2)}{2n-2}\left(\frac{2}{n}\right)}}$.

$\therefore t = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{(s_1^2+s_2^2)/(n-1)}}$ is a students t variate with

$v = n + n - 2 = 2n - 2.$

II. (C) suppose the sample size are equal and if then n pairs of values in this case are not independent.

The test statistic $t = \frac{\bar{x}-\mu}{s/\sqrt{n-1}}$ to test whether the means of differences is significantly different from zero. In this case the d.f is n-1.

Confidence limits (Fiducial limits). Ifσ is not known and n is small then

1. 95% confidence limits for μ is $\left(\bar{x} - \frac{st_{.05}}{\sqrt{n-1}}, \bar{x} + \frac{st_{.05}}{\sqrt{n-1}}\right)$
2. 99% confidence limits for μ is$\left(\bar{x} - \frac{st_{.01}}{\sqrt{n-1}}, \bar{x} + \frac{st_{.01}}{\sqrt{n-1}}\right)$

Problem:

A random sample of 10 boys has the following I.Q (intelligent quotients). 70, 120, 110, 101, 88, 95, 98, 107, 100. Do these data support the assumption of a population mean of a population mean I.Q of 100?

Solution:

Given n=10; μ=100 . Set H_0 : μ=100

Under H_0 test statistics $\frac{\bar{x}-\mu}{s/\sqrt{n-1}} \sim t_{(n-1)}$ where$\bar{x}$ and s can be calculated from the sample data as $\bar{x} = {}^{972}/_{10} = 97.2$ and

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{1833.60}{10} = 183.36.$$

Hence s =13.54.

$$\therefore t = \frac{97.2 - 100}{13.54/_9} = \frac{-2.8 \times 3}{13.54} = -.6204$$

$$\therefore |t| = .62\text{(nearly).}$$

The table value for 9 d.f at 5% level of significance is $t_{.05}$=2.26

$\therefore |t| = .62 < t_{.05.}$ Hence the difference is not significant at 5% level. Hence H_0 may be accepted at 5% level hence the data support the assumption of population mean 100.

Problem:

It was found that a machine has produced pipes having a thickness .05 mm. to determine whether the machine is in proper working order a sample of 10 pipe is chosen for which the mean thickness is .53mm and s.d is 0.3mm .test the hypothesis that the machine is in proper working order using a level of significance of (1) .05 (2) .01

Solution :

Given μ= .50,$\bar{x}$=.53;$s = .03; n = 10.$

Set the null hypothesis H_0 :μ=50

Under the null hypothesis the test statistic is $t = \frac{\bar{x}-\mu}{{}^{s}/_{\sqrt{n-1}}} = \frac{.53-.50}{.03} \times \sqrt{9}$

$$= \frac{.03 \times 3}{.03} = 3.$$

(i)The table value for $v = 9$d.f at 5% level of significance is $t_{.05}$=2.26
(ie)│t│=3>$t_{.05.}$

∴The difference is significant at 5% level of significance.

∴The null hypothesis is rejected at 5%level of significance .

(ii) The table value for $v = 9$ d.f at 1% level of significance is$t_{.01} = 3.25.$

Hence │t│=3< $t_{.01.}$

∴The difference is not significant at 1% level of significant .
Hence the null hypothesis is accepted at 1% level of significance.

Problem:

A group of 10 rats fed on a diet A and another group of 8 rats fed on a different diet B recorded the following increase in weight in gms.

Diet A	5	6	8	1	12	4	3	9	6	10
Diet B	2	3	6	8	1	10	2	8	-	-

Test whether diet A is superior to diet B .

Solution :

Given $n_1 = 10; n_2 = 8.$

Mean of the first sample $\bar{x}_1 = \frac{5+6+\cdots+10}{10} = \frac{64}{10} = 6.4.$

Mean of the second sample$\overline{x}_2 = \frac{2+3+\cdots+8}{8} = \frac{40}{8} = 5.0.$

Standard deviation $s_1$and s_2 of the first and second sample can be found as

$s_1^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2$=10.24 and $s_2^2 = 10.25$

Set the null hypothesis $H_0 : \mu_1 = \mu_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{6.4 - 5}{\sqrt{\frac{10 \times 10.24 + 8 \times 10.25}{10 + 8 - 2}\left(\frac{1}{10} + \frac{1}{8}\right)}}$$

$$= \frac{1.4}{\sqrt{11.525(.1 + .125)}} = .875.$$

Table value for t at 5% level of significance for $(n_1 + n_2 - 2) = 16\ d.f\ is$ $t_{.05}$=2.12.

Since t=.875 <$t_{.05}$ the difference is not significant at 5% level of significance .

Hence the null hypothesis may be accepted.

Problem:

The table gives the biological values of protein from 6 cows milk and 6 buffalo's milk . Examine whether the differences are significant .

Cow's milk	Buffalo's milk
1.8	2.0
2.0	1.8
1.9	1.8
1.6	2.0
1.8	2.1
1.5	1.9

Solution:

Mean value of protein of cow's milk =1.6

Mean value of protein of buffalo's milk =1.9

Variance of protein of cow's milk =.03

Variance of protein in buffalo's milk=0.1

We notice that the two sets of observations are independent .

Given $n_1 = n_2 = 6; \bar{x}_1 = 1.8; \bar{x}_2 = 1.9; s_1^2 = .03;$

$$s_2^2 = .01.$$

Set null hypothesis $H_0: \bar{x}_1 = \bar{x}_2$. Under this null hypothesis the test statistic is $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n-1}}}$

and the d.f $v = 2n - 2 = 10$.

$$= \frac{-.1}{\sqrt{(.03 + .01)/_5}} = \frac{-.1}{\sqrt{.04/_5}} = -1.11.$$

The table value for $v = 10$d.f at 5% level of significance

is 2.23.

│t│=1.11<2.23.Hence the difference is not significant .

Hence the hypothesis is accepted .

Problem:

Ten soldiers participated in a shooting competition in the first week. After intensive training they participated in the competition in the second week. Their scores before and after coaching were given as follows.

Soldiers	1	2	3	4	5	6	7	8	9	10
Score before(x)	67	24	57	55	63	54	56	68	33	43
Score after(y)	70	38	58	58	56	67	68	75	42	38

Do the data indicate that the soldier have been identified by the training ?

Solution:

Here we are connected with the same set of the soldiers in the 2 competitions and their scores which are related to each other because of the intensive training .we compute the difference in their scores $z = y - x$and calculate the mean $\bar{z}$ and the s.d z as follow

x	y	$z = y - x$	$z - \bar{z}$	$(z - \bar{z})^2$
67	70	3	-2	4
24	38	14	9	81
57	58	1	-4	16
55	58	3	-2	4
63	56	-7	-12	144
54	67	13	8	64
56	68	12	7	49
68	75	7	2	4
33	42	9	4	16
43	38	-5	-10	100
-	-	50	-	482

$\bar{z}=\frac{50}{10} = 5; s^2 = \frac{\sum(z-\bar{z})^2}{10} = \frac{482}{10} = 48.2$

Set the null hypothesis $H_0 : \bar{z}$=0.

Under the null hypothesis the test statistic is $t = \frac{\bar{z}-0}{s/\sqrt{n-1}} = \frac{5}{\sqrt{48.2}} \times \sqrt{9} = \frac{15}{6.94} = 2.16$

The table value for $v = 9$d.f at 55 level of significance is $t_{.05} = 2.26$.

$$\therefore \left| t \right| = 2.16 < t_{.05.}$$

The difference is not significant on 5% level of significance .

Hence the null hypothesis is accepted .We can conclude that there is no significant improvement in the training .

TEST BASED ON χ^2- DISTRIBUTION

INTRODUCTION:

The χ^2 distribution has number of application in statistics. It has three important applications based on χ^2 distribution.

I. χ^2- test for population variance.

II. χ^2-test to test the goodness of fit.

III. χ^2-test to test the independence of attributes.

χ^2-TEST.

$I.\chi^2$-test for population variance

Let $x_1, x_2, \dots x_n$ be a random sample from a normal population with variance σ^2.Set the null hypothesis $H_0: \sigma^2 = \sigma_0^2$. Then the test statistic is $\chi^2 = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_0}\right)^2 = \frac{ns^2}{\sigma_0^2}$ where s^2 is the variance of the sample. Then χ^2=$\frac{ns^2}{\sigma_0^2}$ defined above follows aχ^2 distribution with $n-1$ degrees of freedom.

Problem:

A random sample of size 25from a population gives the sample standard deviation 8.5.Test the hypothesis that the population s.d is 10.

Solution:

Given σ=10,n=25,s=8.5 $H_0: \sigma = 10$

$\chi^2 = \frac{ns^2}{\sigma_0^2}$=$\frac{25\times 8.5^2}{100} = 18.06.$

The table value ofχ^2 for 24 d.f =36.415 at 5% level of significance.

It is not a significant. Hence the null hypothesis is accepted.

Problem:

Test the hypothesis that σ=8 given that s=10 for a random sample of size 51.

Solution:

Given n_1=51,σ=8,s=10.

Let $H_0: \sigma = 8$.

$\chi^2 = \frac{ns^2}{\sigma_0^2} = \frac{51\times10^2}{8^2} = 79.7.$

Since$Z = \sqrt{2\chi^2}$-$\sqrt{2n-1}$=$\sqrt{2\times79.7}$-$\sqrt{2\times51-1}$

=2.58

$$\left|z\right| = 2.58 > 1.96.$$

Hence the difference is significant at 5% level of significance and hence the hypothesis is rejected at 5% level of significance.

$II.\chi^{2}$-TEST TO TEST THE GOODNESS OF FIT

The χ^2 -distribution can be used to test the goodness of fit. This test can also be applied to test for compatibility of observed frequencies and theoretical frequencies. Let $o_1, o_2, \dots o_n$ be the observed frequencies and $e_1, e_2, \dots e_n$ be the corresponding expected frequencies such that $\sum_{i=1}^{n} o_i = N = \sum_{i=1}^{n} e_i$ where N is the number of members in the population.

Define χ^2=$\sum_{i=1}^{n}\frac{(o_i-e_i)^2}{e_i}$ .It is a χ^2 variable with n-1 degrees of freedom.

Problem:

The theory predicts that the proportion of an object available in four groups A,B,C,D should be 9:3:3:1. In an experiment among 1600 items of this object the members in the four groups were 882,313,287 and188.use χ^2-test to verify whether the experimental result support the theory.

Solution:

The observed frequencies o_i are 882,313,287,118.

$\sum o_i$=882+313+287+118=1600

The expected frequencies are in the ratio 9:3:3:1.

∴ The expected frequencies e_i are 900,300,300,100.

$\sum e_i$=1600=$\sum o_i$.

$$\therefore \chi^2 = \sum \frac{(o_i-e_i)^2}{e_i}.$$

$$= \frac{(882-900)^2}{900} + \frac{(313-300)^2}{300} + \frac{(287-300)^2}{300} + \frac{(118-100)^2}{100} \quad = 4.7266$$

Degrees of Table value of χ^2 for 3 d.f at 5% level of significance is 7.851.

Calculated χ^2=4.7266<7.852=table value of χ^2.It is not significant. Hence the null hypothesis may be accepted at 5% level of significance and hence we may conclude that experiment results support the theory.

Problem:

Fit a poisson distribution for the following data and test the goodness of fit.

x	0	1	2	3	4	5	6	Total
f	273	70	30	7	7	2	1	390

Solution:

$$\bar{x}=\frac{\sum fx}{\sum f}=\frac{70+60+21+28+10+6}{273+70+30+7+7+2+1}=\frac{195}{390}$$

$$\lambda = {}^{1}/_{2}.$$

The theoretical frequencies of the poisson distribution are given by

$$f(x)=\frac{Ne^{-\lambda}\lambda^{x}}{x!}=\frac{390}{\sqrt{e}x!}\left(\frac{1}{2}\right)^{x}\ ; x=0,1,2,\ldots 6.$$

freedom =4-1 =3 .

The expected frequencies are by$f(0)=\frac{390}{\sqrt{e}}\left(\frac{1}{2}\right)^{0}=236.4;$

$f(1)=\frac{390}{\sqrt{e}1!}\left(\frac{1}{2}\right)^{1}=118.2$ ....,

$$f(6)=\frac{390}{\sqrt{e}6!}\left(\frac{1}{2}\right)^{6}=0.005.$$

Thus the observed and expected frequencies can be shown below

o_i	273	70	30	7	7	2	1	390
e_i	236.4	118.2	29.5	4.9	.6	.1	0	389.7

Since the sum of the expected frequencies is 389.7.It can be adjusted in the last two frequencies by adding .3.

o_i	273	70	30	17	390
e_i	236.4	118.2	29.5	5.9	390

Set up the null hypothesis H_0:Poisson distribution can be fitted well.

The test statistics is $\chi^2 = \frac{\sum(o_i - e_i)^2}{e_i}$.

$$= \frac{(273-236.4)^2}{236.4} + \frac{(70-118.2)^2}{118.2} + \frac{(30-29.5)^2}{29.5} + \frac{(17-5.9)^2}{5.9} = 46.3.$$

Degrees of freedom=7-1-1-3=2.

The table value of 2 d.f. at 5% level is 5.99.

Since χ^2=46.3>5.99=The table value of $\chi^2_{.05}$ it is much significant at 5% level of significance.

Hence the hypothesis is rejected at 5% level and hence the poisson distribution is not a good fit to the data.

UNIT V: INDEX NUMBERS

Index Numbers - Types of index numbers - Tests - Unit test commodity reversal test, time reversal test, factor reversal test - Chain index numbers - cost of living index – Interpolation - Finite differences operators - Newton's forward, backward interpolation formulae, Lagrange's formula.

INDEX NUMBERS

Index Numbers :

An index number is widely used statistical device for comparing the level of a certain phenomenon with the level of the same phenomenon at some standard period.

In the computation of an index number, if the base year used for comparison is kept constant throughout, then it is called fixed base method. If on the other hand, for every year the previous year is used as a base for comparison, then the method is called chain base method.

Index numbers can be broadly classified into two types.

I. Unweighted or simple index number

II. Weighted index number.

Two standard methods of computation are

A) Aggregate method

B) Average of price relatives method.

I. A). Aggregate method:

In this method total of current prices for various commodities is divided by the total of the base year. In symbols if p_0 denotes the price of the base year and p_1 the price of the current year.

$p_{01} = \frac{\sum P_1}{\sum P_0}\, x\ 100\ where \sum P_1$ is total of the current year

$\sum p_0 \quad is\ the\ total\ of\ the\ base\ year.$

Example :

From the following data construct the simple aggregative index number for 1992.

Commodities	Price in 1991 Rs	Price in 1992 Rs
Rice	7	8
Wheat	3.5	3.75
Oil	40	45
Gas	78	85
Flour	4.5	5.25

Solution :

Construction of price index taking 1991 as base year.

Commodities	Price in 1991 Rs	Price in 1992 Rs
Rice	7	8
Wheat	3.5	3.75
Oil	40	45
Gas	78	85
Flour	4.5	5.25
Total	133.00	147.00

∴aggregate index number $P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$

$$= \frac{147}{133} \times 100$$

$$= 110.5$$

I. B) Average of Price Relatives method (Simple index numbers)

Price relatives denoting the price of a commodity of a base year as P_0 and the price of the current year as P_1 the ratio of the prices $\frac{P_1}{P_0}$ is called the price relatives.

Index number for the current year is = $P_{01} = \frac{P_1}{P_0} \times 100$

i) The Arithmetic mean index number $P_{01} = \frac{\Sigma\left(\frac{P_1}{P_0}\right)\times 100}{n}$

ii) The Geometric mean index number

$$P_{01} = \left[\pi\left(\frac{P_1}{P_0}\right)\right]^{1/n} \times 100, where\ \pi\ denotes\ the\ product$$

$$Hence \log P_{01} = \frac{\sum \log\left(\frac{P_1}{P_0}\right) \times 100}{n}$$

Problem:

For the above example, we find the index number of the price relatives taking 1991 as the base year using i) Arithmetic mean ii) Geometric mean.

Solution:

Commodities	**Price in 1991** $\boldsymbol{P_0}$	**Price in 1992** $\boldsymbol{P_1}$	$\boldsymbol{\left(\frac{P_1}{P_0}\right) \times 100}$	$\boldsymbol{\log\left(\frac{P_1}{P_0}\right) \times 100}$
Rice	7	8	114.3	2.0580
Wheat	3.5	3.75	107.41	2.0298
Oil	40	45	112.5	2.0512

Gas	78	85	109.0	2.0374
Flour	4.5	5.25	116.7	2.0671
Total			**559.6**	**10.2435**

i) Using arithmetic mean the index number

$$P_{01} = \frac{559.6}{5} = 111.92$$

ii) Using geometric mean the index number

$$\log P_{01} = \frac{10.2435}{5} = 2.048\%$$

$$\therefore\ P_{01} = antilog\ (2.0487) = 111.87$$

Problem:

From the following data of the whole sale price of rice for the 5 years construct the index numbers taking (i) 1987 as the base (ii) 1990 as the base.

Years	1987	1988	1989	1990	1991	1992
Price of rice per kg	5.00	6.00	6.50	7.00	7.50	8.00

Solution :

i) Construction of index numbers taking 1987 as base.

Years	**Price of rice per kg**	**Index numbers (base 1987)**

1987	5.00	100
1988	6.00	$\frac{6}{5} \times 100 = 120$
1989	6.50	$\frac{6.5}{5} \times 100 = 130$
1990	7.00	$\frac{7}{5} \times 100 = 140$
1991	7.50	$\frac{7.5}{5} \times 100 = 150$
1992	8.00	$\frac{8}{5} \times 100 = 160$

From the index number table we observe that from 1987 to 1988 these is a increase of 20% in the price of rice per kg; for 1987 to 1989 there is a increase of 30% in the price of rice per kg etc.

ii) Construction of index numbers taking 1990 as base

Years	Price of rice per kg	Index number (Base 1990)
1987	5	$\frac{5}{7} \times 100 = 71.4$
1988	6	$\frac{6}{7} \times 100 = 85.7$
1989	6.50	$\frac{6.5}{7} \times 100 = 92.9$
1990	7	100
1991	7.5	$\frac{7.5}{7} \times 100 = 107.7$
1992	8	$\frac{8}{7} \times 100 = 114.3$

Problem:

Construct the whole sale price index number for 1991 and 1992 from the data given below using 1990 as the base year.

Commodity	Whole sale prices in Rupees per quintal		
	1990	1991	1992
Rice	700	750	825
Wheat	540	575	600
Ragi	300	325	310
Cholam	250	280	295
Flour	320	330	335
Ravai	325	350	360

Solution: Taking 1990 as base year

Commodity	**1990**	**1991**	**1992**	**Relatives for 91**	**Relatives for 92**
Rice	700	750	825	$\frac{750}{700}\times 100$ $= 107.1$	$\frac{825}{700}\times 100$ $= 117.9$
Wheat	540	575	600	$\frac{575}{540}\times 100$ $= 106.5$	$\frac{600}{540}\times 100$ $= 111.1$
Ragi	300	325	310	$\frac{325}{540}\times 100$ $= 108.3$	$\frac{310}{300}\times 100$ $= 103.3$
Cholam	250	280	295	$\frac{280}{250}\times 100$ $= 112$	$\frac{295}{250}\times 100$ $= 118$
Flour	320	330	335	$\frac{330}{320}\times 100$ $= 103.1$	$\frac{325}{320}\times 100$ $= 101.6$
Ravai	325	350	360	$\frac{350}{325}\times 100$ $= 107.7$	$\frac{360}{325}\times 100$ $= 110.8$
Total				**644.7**	**662.7**
Index Number (using AM)				**107.5**	**110.5**

Problem:

From the following average prices of the three groups of commodities given in rupees per unit find (i) fixed base index number (ii) chain base index numbers with 1988 as the base year and

Commodity	1988	1989	1990	1991	1992
A	2	3	4	5	6
B	8	10	12	15	18
C	4	5	8	10	12

Solution:

i) Fixed base index number

Commodity	1988	1989	1990	1991	1992
A	100	$\frac{3}{2}\times\ 100$ $= 150$		$\frac{5}{2}\times\ 100$ $= 250$	$\frac{6}{2}\times\ 100 = 300$
B	100	$\frac{10}{8}\times 100$ $= 125$	$\frac{12}{8}$ $\times 100$ $= 150$	$\frac{15}{8}\times\ 100$ $= 188$	$\frac{18}{8}\times\ 100 = 225$

C	100	$\frac{5}{4}\times 100$ $= 125$	$\frac{8}{4}$ $\times 100$ $= 200$	$\frac{10}{4}\times 100$ $= 200$	$\frac{12}{4}\times 100 = 300$
Total	**300**	**400**	**550**	**688**	**825**
Index number (AM)	**100**	**133.3**	**183.3**	**229.3**	**275**

ii) Chain base index numbers:

Commodity	**1988**	**1989**	**1990**	**1991**	**1992**
A	$\frac{2}{2}$ $\times 100$ $= 100$	$\frac{3}{2}$ $\times 100$ $= 150$	$\frac{4}{3}\times 100$ $= 133.3$	$\frac{5}{4}$ $\times 100$ $= 125$	$\frac{6}{5}$ $\times 100$ $= 120$
B	$\frac{8}{8}$ $\times 100$ $= 100$	$\frac{10}{8}$ $\times 100$ $= 125$	$\frac{12}{10}$ $\times 100$ $= 120$	$\frac{15}{12}$ $\times 100$ $= 125$	$\frac{18}{15}$ $\times 100$ $= 120$
C	$\frac{4}{4}$ $\times 100$ $= 100$	$\frac{5}{4}$ $\times 100$ $= 125$	$\frac{8}{5}\times 100$ $= 160$	$\frac{10}{8}$ $\times 100$ $= 125$	$\frac{12}{10}$ $\times 100$ $= 120$
Total					

	300	400	413.3	375	360
Index number (AM)	**100**	**133.3**	**137.8**	**125**	**120**

II. Weighted Index numbers :

Standard methods of computing weighted index number are

II-A weighted aggregative method

II – B weighted average of price relatives method.

II – A weighted aggregative method:

Though there are many formulae to calculate index number in this method we give below some standard formulae which are very often used.

a) Laspeyre’s index number:

According to Laspeyre’s method the prices of the commodities in the base year as well as the current year are known and they are weighted by the quantities used in the base year.

$$L_{I01} = \frac{\sum P_1 q_0}{\sum p_0 q_0} \times 100$$

b) Paasche’s index number:

According to paasche’s method current year quantities are taken as weights and hence paasche’s index number is defined.

$$L_{I01} = \frac{\sum P_1 q_1}{\sum p_0 q_1} \times 100$$

C) Marshall – Edgeworth’s Index number:

According to this method the weight is the sum of the quantities of the base period and current period.

$$M_{I01} = \frac{\sum P_1q_0 + \sum P_1q_1}{\sum p_0q_0 + \sum p_0q_1} \times 100$$

d) Bowley’s Index number:

The arithemetic mean of Laspeyre’s and paasche’s index number is defined to be Bowley’s index number.

$$BI_{01} = \frac{1}{2}\left[\frac{\sum P_1q_0}{\sum p_0q_0} + \frac{\sum P_1q_1}{\sum p_0q_1}\right] \times 100$$

e) Fisher’s index number:

$$I_{01} = \sqrt{\frac{\sum P_1q_0}{\sum p_0q_0} \times \frac{\sum P_1q_1}{\sum p_0q_1}} \text{x } 100$$

$$I_{01} = \sqrt{LI_{01} \times PI_{01}}$$

f) Kelley’s Index number:

According to Kelley, weight may be taken as the quantities of the period which is not necessarily the

base year or current year. The average quantity of two or more years may be taken as the weight.

$K_{I01} = \frac{\sum P_1q}{\sum P_0q}$x 100. Where q is the average quantity of two or more years.

Example :

Calculate i) Laspeyre’s (ii) Paasche’s iii) Fisher’s index number for the following data given below. Hence or otherwise find Edgeworth and Bowley’s index number.

Commodities	Base year 1990		Current year 1992	
	Price	Quantity	Price	Quantity
A	2	10	3	12
B	5	16	6.5	11
C	3.5	18	4	16
D	7	21	9	25
E	3	11	3.5	20

Solution:

Commodities	**1990**		**1992**		P_0q_0	P_0q_1	P_1q_0	P_1q_1
	P_0	q_0	P_1	q_1				
A	2	10	3	12	20	24	30	36
B	5	16	6.5	11	80	55	104	71.5
C	3.5	18	4	16	63	56	72	64
D	7	21	9	25	147	175	189	225
E	3	11	3.5	20	33	60	38.5	70
Total					**343**	**370**	**433.5**	**466.5**

i) Las Peyre's Index number = $\frac{\sum P_1 q_0}{\sum p_0 q_0} \times 100$

= $\frac{433.5}{343} \times 100 = 126.4$

ii) Paasche's index number $P_{I01} = \frac{\sum P_1 q_1}{\sum p_0 q_1} \times 100 = \frac{466.5}{370} \times 100 = 126.1$

iii) Fisher's Ideal index number = $\sqrt{\frac{\sum P_1 q_0}{\sum p_0 q_0} \times \frac{\sum P_1 q_1}{\sum p_0 q_1}} \times 100$

$$= \sqrt{\frac{433.5 \times 466.5}{343 \times 370}} \times 100$$

$$= 126.2$$

iv)Bowly's Index numbers = $\frac{L_{I01} + P_{I01}}{2} = \frac{126.4+126.1}{2}$

$$= 126.25$$

v) Edge – Worth's Index number = $\frac{\sum P_1 q_0 + \sum P_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$

$$= \frac{433.5 + 466.5}{343 + 370} \times 100$$

$$= \frac{900}{713} \times 100$$

$$= 126.2$$

II – B. Weighted Average of Price Relative Method :

In this method the index number is computed by taking the weighted Arithmetic mean of price relatives. Thus if P is the price relative and V is the value weights $P_0 q_0$ then the index number $P_{01} = \frac{\sum PV}{\sum V}$

Example :

Index number using weighted arithmetic mean of price relatives.

Commodity	Price in 1990 P_0	Price in 1992 P_1	Quantity in 1990 q_0	V $P_0 q_0$	P $\frac{P_1}{P_0} \times 100$	PV

Coconut oil	Rs. 50	Rs. 54	15 lit.	750	108	8100
Groundnut oil	Rs. 45	Rs. 48	25 lit.	1125	106.7	120037.5
Gingili oil	Rs. 43	Rs. 45	30 lit.	1290	104.7	135063
Rice	Rs. 7	Rs. 9	350 kg	2450	128.6	315070
Total				**5615**	**-**	**651170.5**

Weighted index number = $\frac{\sum PV}{\sum V}$

$$=\frac{\mathbf{651170.5}}{\mathbf{5615}} \cong \mathbf{116}$$

Ideal Index number :

An index number is said to be ideal index number if it is subjected to the following three test.

i) The time reversal test

ii) The factor reversal test

iii)The commodity reversal test

i) The time reversal test :

Let $I_{(01)}$ denote the index number of the current year $y_1$relative to the base year y_0 without considering percentage, and $I_{(01)}$ denotes the index number of the base year y_0 relative to the current year y_1 without considering the percentage. If $I_{(01)} \times I_{(10)} = 1$, then we say that the index number satisfies the time reversal test.

ii) The factor reversal test:

In this test the prices and quantities are interchanged, without considering the percentage, satisfying the following relation $I_{(pq)} \times \ I_{(qp)} = \frac{\sum P_1 q_1}{\sum p_0 q_0}$, where $I_{(pq)}$ is the price index of the current year relative to the base year and $I_{(qp)}$ is the quantity index of the current year relative to the base year.

iii) The Commodity reversal test:

The index number should be independent of the order in which different commodities are considered. This test is satisfied by almost all index numbers.

Problem:

Construct Fisher's index number and show that it statistics both the factor reversal test and time reversal test.

Commodity	**aA**	**bB**	**cC**	**dD**
Base year price in Rupees	55	66	44	33
Base year quantity in Quintals	50	40	120	30
Current year price in Rupees	7	8	5	4
Current year quantity in Quintals	60	50	110	35

Solution:

Commodity	**Base year**		**Current year**		P_0q_0	P_0q_1	P_1q_0	P_1q_1
	P_0	q_0	P_1	q_1				
A	5	50	7	60	250	300	350	420
B	6	40	8	50	240	300	320	400
C	4	120	5	110	480	440	600	550

D	3	30	4	35	90	105	120	140
Total					**1060**	**1145**	**1390**	**1510**

Fisher's Index number is $I_{01} = \sqrt{\frac{\sum P_1q_0}{\sum p_0q_0} \times \frac{\sum P_1q_1}{\sum p_0q_1}} \times 100$

$$= \sqrt{\frac{1390}{1060} \times \frac{1510}{1145}} \times 100$$

Time reversal test

Now, $I_{01} = \sqrt{\frac{\sum P_1q_0}{\sum p_0q_0} \times \frac{\sum P_1q_1}{\sum p_0q_1}}$

$$= \sqrt{\frac{1390}{1510} \times \frac{1060}{1390}}$$

Now, ,$I_{(01)} \times I_{(10)} = \sqrt{\frac{1390}{1060} \times \frac{1510}{1145} \times \frac{1060}{1390}} = 1$

Factor reversal test

$$I_{(Pq)} = \sqrt{\frac{\sum P_1q_0}{\sum p_0q_0} \times \frac{\sum P_1q_1}{\sum p_0q_1}} = \sqrt{\frac{1390}{1060} \times \frac{1510}{1145}}$$

Interchanging the factors

$$I_{(Pq)} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum P_1q_1}{\sum P_0q_1}}$$

$$= \sqrt{\frac{1145}{1060} \times \frac{1510}{1390}}$$

Now, $I_{(pq)} \times I_{(qp)} = \sqrt{\frac{1390}{1060} \times \frac{1510}{1145} \times \frac{1145 \times 1510}{1060 \times 1390}}$

$$=\frac{1510}{1060} = \frac{\sum P_1q_1}{\sum p_0q_0}$$

Hence the factor reversal test is also satisfied.

Problem:

Find the missing price in the following data if the ratio between Laspeyre's and Paasche's index numbers is 25:24.

Commodities	Base Year		Current year	
	Price	**Quantity**	**Price**	**Quantity**
A	1	15	2	15
B	2	15	-	30

Solution :

Let the missing price be x

Commodities	Base year		Current year		$P_0 q_0$	$P_0 q_1$	$P_1 q_0$	$P_1 q_1$
	P_0	q_0	P_1	q_1				
A	1	15	2	15	15	15	30	30
B	2	15	X	30	30	60	15x	30x
Total					**45**	**75**	**30+15x**	**30+30x**

Laspeyre's index number

$$L_{I_{(01)}} = \frac{\sum P_1 q_0}{\sum p_0 q_0} \times 100$$

$$= \frac{30+15x}{45} \times 100$$

Paache's index number,

$$P_{I_{(01)}} = \frac{\sum P_1 q_1}{\sum p_0 q_1} \times 100 = \frac{30+30x}{75} \times 100$$

Given $L_{I_{(01)}} : P_{I_{(01)}} = 25 : 24$

$$\therefore \left(\frac{30+15x}{45} \times 100\right) : \left(\frac{30+30x}{75} \times 100\right)$$

= 25:24

$$\therefore \; 24\left(\frac{30+15x}{45}\right) = 25\left(\frac{30+30x}{75}\right)$$

$$72(30+15x) = 45(30+30x)$$

$$8(30+15x) = 5(30+30x)$$

$$8 \times 15\,(2+x) = 5 \times 30(1+x)$$

$$4(2+x) = 5(1+x)$$

$$\therefore 8+4x = 5+5x$$

$$5x - 4x = 8 - 5$$

$$x = 3$$

Hence the missing price is Rs. 3

Interpolation:

Definition:

Interpolation is the process of finding the most appropriate estimate for missing data. It is the art of reading between the lines of a table.

It is also possible that we may require information for future in which case the process of estimating the most appropriate value is known as extrapolation. There are two methods in interpolation.

i) Graphic method

ii) Algebraic method

i) Graphic method is a simple method in which we just plot the available data on a graph sheet and read off the value for the missing period from the graph itself.

ii) Algebraic method:

There are several methods used for interpolation of which we deal with the following.

i). Finite differences

ii) Gregory – Newton's formula

iii) Lagrange's formula

Finite Differences:

Definition:

We define an operator Δ which is known as the first order difference on $U_x\ as$

$\Delta U_x = U_{x+h} - U_x\ where\ x = a, a+h, a+2h, \dots\dots$ in particular.

(i) =$\Delta U_a = U_{a+h} - U_a (ii)\ \Delta U_x = 0$ if U_x is constant.

Example :

The difference table for the following data is given below.

x	**0**	**1**	**2**	**3**	**4**
$\boldsymbol{U_x}$	**8**	**11**	**9**	**15**	**6**

Solution :

X	U_x	ΔU_x	$\Delta^2 U_x$	$\Delta^3 U_x$	$\Delta^4 U_x$
	8				
		3			
0					
	11		-5		
1					
		-2		13	
2					
	9		8		-36
3					
		6		-23	
4					
	15		-15		
		-9			
	6				

The differences $\Delta U_x, \Delta^2 U_x$ $etc.$are called forward differences. In contrast to the forward differences we have another kind of differences known as backward differences.

Theorem : [Fundamental Theorem for finite differences]

If U_x is a polynomial of degree n hence

$$\Delta^r U_x = \begin{cases} \text{constant if r=n} \\ 0 \qquad\quad \text{if r=n} \end{cases}$$

i.e.) the n^{th} order difference of a polynomial of degree n is constant and differences of order higher than n are zero.

Proof:

Let $U_x = a_0x^n + a_1x^{n-1} + \cdots \ldots.. + a_{n-1}x + a_{n.}$ Where $a_0, a_1 \ldots \ldots., a_n$ *are constants and* $a_n \neq 0$.

$\Delta U_x = U_{x+h} - U_x$

$= [a_0(x+h)^{n-1} + \cdots \ldots.. + a_{n-1}(x+h) + a_n] - [a_0x^n + a_1x^{n-1} + \cdots.. + a_{n-1}x + a_n]$

$= [a_0(x^n + nc_1x^{n-1}h + \cdots \ldots. + h^n) + a_1(x^{n-1} + n - 1c_1x^{n-2}h + \cdots. + h^{n-1} + \cdots + a_n]$

$- [a_0x^n + a_1x^{n-1} + \cdots. + a_n]$

$= a_0nhx^{n-1} + b_2x^{n-2} + \cdots \ldots. + b_{n-1}x + b_n$,where $b_2,\ b_3 \ldots.. b_n$ are constants independent of x and $a_0nh \neq 0$.

$\therefore$ $\Delta U_x$is a polynomial of degree n-1 continuing this process we get $\Delta^2 U_x$ is a polynomial of degree n-2.

$\Delta^3 U_x$is a polynomial of degree n-3 etc.

$$\begin{aligned}\therefore\ \Delta^n U_x &= a_0n(n-1)(n-2) \ldots..2.1\ h^n x^0 \\ &= a_0 n!\ h^n \\ &= constant.\ and\ \Delta^r U_x = 0\ for\ r > n.\end{aligned}$$

Problem:

Find first and second differences for

$i) U_x = ab^{cx} (ii) U_x = \frac{x}{x^2+7x+12}$ taking interval of differencing as h.

Solution :

i) $\Delta U_x = U_{x+h} - U_x$

$= ab^{c(x+h)} - ab^{cx}$

$= ab^{cx} ab^{ch} - ab^{cx}$

$= ab^{cx}(b^{ch} - 1)$

$\Delta^2 U_x = (b^{ch} - 1)\Delta\,(ab^{cx})$

$= (b^{ch} - 1)^2 (ab^{cx})$

$ii)\ U_x = \dfrac{x}{x^2 + 7x + 12}$

$= \frac{x}{x+4} - \frac{3}{x+3}$ (by partial fraction)

$\Delta U_x = \left[\dfrac{4}{(x+1)+4} - \dfrac{3}{(x+1)+3}\right] - \left[\dfrac{4}{x+4} - \dfrac{3}{x+3}\right]$

$= \frac{4}{x+5} - \frac{3}{x+4} - \frac{4}{x+4} + \frac{3}{x+3}$

$= \frac{4}{x+5} - \frac{7}{x+4} + \frac{3}{x+3}$

$Similarly, \Delta^2 U_x = \dfrac{4}{x+6} - \dfrac{11x}{x+5} + \dfrac{10}{x+4} - \dfrac{3}{x+3}$

Problem:

If $U_0 = 1, U = 5, U_2 = 8, U_3 = 3, U_4 = 7, U_5 = 0\ find\ \Delta^5 U_0$.

Solution :

	U_x	ΔU_x	$\Delta^2 U_x$	$\Delta^3 U_x$	$\Delta^4 U_x$	$\Delta^5 U_x$
0	1					
		4				
1	5		-1			
				7		
		3				
2	8		-8		24	
				7		
		-5				-61
3	3		9		-37	
				-20		
		4				
4	7		-11			
		-7				
5	0					

Hence , $\Delta^5 U_0 = -61$

Problem:

Estimate the missing term in the following table.

x	0	1	2	3	4
U_x	1	3	9	-	81

Explain why the resulting value from 3^3

Solution :

Let the missing term in U_x $be\ a$.

X	$\boldsymbol{U_x}$	$\boldsymbol{\Delta U_x}$	$\boldsymbol{\Delta^2 U_x}$	$\boldsymbol{\Delta^3 U_x}$	$\boldsymbol{\Delta^4 U_x}$
0	1				
		2			
1	3		4		
		6		a-19	
2	9		a-15		124 – 4a
		a-9		105-3a	
3	Q		90-2a		
		81-a			
4	81				

Since 4 values of U_x are given it is a polynomial of degree 3.

Hence by fundamental theorem of finite differences $\Delta^4 U_x = 0\ for\ all\ x$.

In particular $\Delta^4 U_0 = 0$

Hence 124 – 4a =0

a = 31

Problem:

Give an estimate of the population in 1971 from the following table.

Year	1941	1951	1961	1971	1981	1991
Population in lakhs	363	391	421	?	467	501

Solution :

Let the population in 1971 be ‘a’.

$U_0 = 363, U_1 = 391, U_2 = 421, U_3 = a, U_4 = 467, U_5 = 501$

Since five values are given

$$\Delta^5 U_x = 0, \text{ for all } x$$

In particular, $\Delta^5 U_0 = 0$

$$(E-1)^5 U_0 = 0$$
$$(E^5 - 5E^4 + 10E^3 - 10E^2 + 5E - 1)U_0 = 0$$
$$\therefore\ U_5 - 5U_4 + 10U_3 - 10U_2 + 5U_1 - U_0 = 0$$
$$\therefore\ 501 - 5 \times 497 + 10a - 10 \times 421 + 5 \times 391 - 363 = 0$$
$$\therefore\ 501 - 2335 + 10a - 4210 + 1955 - 363 = 0$$
$$10a - 4452 = 0$$
$$a = 445.2\ lakhs$$

Hence the estimated population in 1971 is 445.2 lakhs.

Problem:

Find the missing figures in the following table

x	0	5	10	15	20	25
U_x	7	11	?	18	?	32

Solution :

Here ,two values are missing. Let the missing values be a and b.

$U_0 = 7, U_1 = 11, U_2 = a, U_3 = 18, U_4 = b, U_5 = 32.$

Since four values are given, we have $\Delta^4 U_x = 0 \ for\ all\ x.$

In particular, $\Delta^4 U_0 = 0 \ and\ \Delta^4 U_1 = 0$

$$(E-1)^4 U_0 = 0$$

$$\therefore \ (E^4 - 4E^3 + 6E^2 - 4E + 1)U_0 = 0$$

$$U_4 - 4U_3 + 6U_2 - 4U_1 + U_0 = 0$$

$$b - 72 + 6a - 44 + 7 = 0$$

$$ie)\ 6a + b = 10q \ \ldots\ldots\ldots..(1)$$

Taking $\Delta^4 U_1 = 0$

$$(E^4 - 4E^3 + 6E^2 - 4E + 1)U_1 = 0$$

$$U_5 - 4U_4 + 6U_3 - 4U_2 + U_1 = 0$$

$$32 - 4b + 108 - 4a + 11 = 0$$

$$4a + 4b = 151 \ \ldots\ldots\ldots..(2)$$

Solving (1) and (2) we get,

A = 14.25, b=23.5

Hence the missing values are 14.25 and 23.5.

Problem:

Given that $U_0\ to\ U_8 = 80;\ U_1 + U_7 = 10;$

$U_2 + U_6 = 5;\ \ U_3 + U_5 = 10\ find\ U_{4.}$

Solution :

Since 4 values are given

$\therefore \;\; \Delta^n U_x = 0 \, for \; all \; n \geq 4 \; for \; all \; x \, .$

$$In \; particular \; \Delta^8 U_0 = 0$$

$$Hence \; (E-1)^8 U_0 = 0$$

$$U_8 - 8U_7 + 28U_6 - 56U_5 + 70U_4 - 56U_3 + 28U_2 - 8U_1 + U_0 = 0$$

$$(U_0 + U_8) - 8\,(U_1 + U_7) + 28(U_2 + U_6) - 56\,(U_3 + U_5) + 70U_4 = 0$$

$$\therefore 80 - 80 + 140 - 560 + 70U_4 = 0$$

$$\therefore \;\; 70U_4 = 420$$

$$\therefore \;\; U_4 = 6.$$

Problem:

Given that $U_1 + \; U_2 + U_3 = 25, U_4 \; = 29, \; U_5 \; + U_6 = 113.$ Find the polynomial U_x and hence find $U_{10.}$

Solution :

Since three values are given U_x is a polynomial of degree 2.

Let $U_x = ax^2 + bx + c$

$$U_1 = a + b + c, U_2 = 4a + 2b + c, U_3 = 9a + 3b + c$$

Given, $U_1 + U_2 + U_3 = 25$

$$\therefore \;\; 14a + 6b + 3c = 25 \; \ldots \ldots .. (1)$$

$$Now, U_4 = 24$$

$$\Rightarrow 16a + 4b + c = 24 \ldots \ldots \ldots \ldots \ldots (2)$$

$U_5 + U_6 = 113$

$$61a + 11b + 2c = 113$$

Solving (1), (2) and (3) we get

A = 2, b=-1, c=1

$$\therefore U_x = 2x^2 - x + 1$$

$$\begin{aligned} Now, U_{10} &= 100a + 10b + c \\ &= 200 - 10 + 1 \\ &= 191 \end{aligned}$$

Problem:

If $U_1 = (12 - x)(4 + x);$

$$U_2 = (5 - x)(4 - x),$$

$$U_3 = (x + 18)(x + 6) and\ U_4 = 9.$$

Obtain a value of x assuming second differences constant.

Solution :

$$\Delta^3 U_1 = 0$$

$$Hence\ (E - 1)^3 U_1 = 0$$

$$\therefore (E^3 - 3E^2 + 3E - 1)U_1 = 0$$

$$U_4 - 3U_3 + 3U_2 - U_1 = 0$$

$$\therefore \ \ 94 - 3(x + 18)(x + 6) + 3(5 - x)(4 - x) - ?\,(12 - x)(4 + x) = 0$$

$$ie)\ \ x^2 - 107x - 218 = 0$$

$$(x - 109)(x + 2) = 0$$

$$x = 109\ \ or - 2$$

Newton's Formula:

Consider the function y=f(x). Let f$(x_0) = y_0, f(x_1) = y_1 \dots\dots\dots f(x_n) = y_n$.

Newton-Gregory formula for forward interpolation:

$$U_x = U_a + (x-a)\frac{\Delta U_a}{1!\,h} + (x-a)(x-a-h)\frac{\Delta^2 U_a}{2!\,h^2}$$

$$+ \cdots \ldots.(x-a)(x-a-h)\ldots\ldots(x-a-\overline{n-1}\,h)\frac{\Delta^n U_a}{n!\,h^n}$$

Newton's formula for backward interpolation

$$U_x = U_{a+nh} + \frac{\nabla U_a + nh}{1!h}\left(x-\overline{a+nh}\right) + \frac{\nabla^2 U_{a+nh}}{2!h^2}$$

$$\left(x-\overline{a+nh}\right)(x-\overline{a+(n-1)h} + \cdots \ldots. + \frac{\nabla^n U_{a+nh}}{n!\ h^n}\left(x-\overline{a+nh}\right)\ldots$$

$$\ldots(x-\overline{a+h})$$

Problem:

$$U_{75} = 246;\ \ U_{80} = 202;\ \ U_{85} = 118\ and\ U_{90} = 40\ find\ U_{79.}$$

Solution :

Here, a= 75, h=5

To find $U_{a+rh} = U_{79}$

$$\therefore\ \ a + rh = 79$$

$$75 + 5r = 79$$

$$r = \frac{4}{5} = 0.8$$

By Newton-Gregory formula for equal intervals,

$$U_{a+rh} = U_a + \frac{r}{1!}\Delta U_a + \frac{r(r-1)}{2!}\Delta^2 U_a + \cdots \ldots$$

X	U_x	ΔU_x	$\Delta^2 U_x$	$\Delta^3 U$
75	246			
		-44		
80	202		-40	
		-84		46
85	118		6	
		-78		
90	40			

$$\therefore\ U_{79} = 246 + \frac{0.8\,(-44)}{1} + \frac{0.8\,(0.8-1)}{1.2}(-40) + \frac{0.8\,(0.8-1)(0.8-2)}{1.2.3}(46)$$

$$= 246 - 35.2 + 3.2 + 1.472$$

$$= \ 215.472$$

Problem:

By using Gregory – Newton's formula find U_x for the following data. Hence estimate (i) $U_{1.5}(ii)U_9$

U_0	U_1	U_2	U_3	U_4
1	11	21	28	29

Solution :

X	U_x	$\Delta \boldsymbol{U_x}$	$\Delta^2 \boldsymbol{U_x}$	$\Delta^3 \boldsymbol{U_x}$	$\Delta^4 \boldsymbol{U_x}$
		10			
0	1				
			0		
1	11				
		10		-3	
2	21				
			-3		0
3	28				
		7		-3	
4	29				
			-6		
		1			

Here the third order difference are constant and hence required function is a polynomial of degree 3.

$$U_x = U_a + (x-a)\frac{\nabla U_a}{1!} + (x-a)(x-a-h)\frac{\Delta^2 U_0 h^2}{2!} + \cdots ....$$

Here, a=0 and h=1

$$\therefore\ U_x = 1 + (x-0)\times\frac{10}{1!} + (x-1)\times\frac{0}{2!} + x(x-1)(x-2)\times\frac{(-3)}{3!}$$

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

$$= \frac{1}{2}[2 + 20x - x^3 + 3x^2 - 2x]$$

$$U_x = \frac{1}{2}(-x^3 + 3x^2 + 18x + 2)$$

i) $\therefore\ U_{1.5} = \frac{1}{2}\,[-(1.5)^3 + 3(1.5)^2 + 18(1.5) + 2]$

$$= \frac{1}{2}[-3.375 + 6.75 + 27 + 2]$$

$$= 16.188$$

ii) $U_9 = \frac{1}{2}[-9)^3 + 3 \times 9^2 + 18 \times 9 + 2]$

$$= \frac{1}{2}[-729 + 243 + 162 + 2] = -161$$

Problem:

Population was recorded as follows in a villages.

Year	1941	1951	1961	1971	1981	1991
Population	2500	2800	3200	3700	4350	5225

Estimate the population for the year 1945.

Solution :

Year x	Population U_x	$\Delta \boldsymbol{U_x}$	$\Delta^2 \boldsymbol{U_x}$	$\Delta^3 \boldsymbol{U_x}$	$\Delta^4 \boldsymbol{U_x}$	$\Delta^5 \boldsymbol{U_x}$
		300				
1941	2500		100			
1951	2800	400		0		
1961	3200		100		50	
1971	3700	500		50		-25
1981	4350		150		25	
1991	5225	650		75		
			225			
		875				

We have to find U_{1945}

Here a = 1941 and h=10

$$\therefore U_{a+rh} = U_{1945}$$

$$Hence\ 1941 + 10 = 1945$$

$$r = 0.4$$

Applying Newton – Gregory formula

$$U_{1945} = 2500 + 0.4 \times \frac{300}{1} + 4(0.4-1) \times \frac{100}{2!} + 0.4(0.4-1)(0.4-2) \times \frac{0}{3!} +$$

$$(0.4)(0.4-1)(0.4-3) \times \frac{50}{4!} + 0.4(0.4-1)(0.4-2)(0.4-3)(0.4-4) \times \frac{25}{5!}$$

$$= 2500 + 120 - 12 - 2.08 + 0.75$$

$$= 2606.67 \cong 2607$$

4. From the following data estimate the number of persons whose daily wage is between Rs. 40-50.

Daily wage in Rs.	No. of persons
0-20	120
20-40	145
40-60	200
60-80	250
80-100	150

Solution :

The less than cumulative frequency table of the above data is given by

Wage less than x	**No. of persons (c.f)**
20	120
40	265
60	465
80	715
100	865

The difference table

X	U_x	$\boldsymbol{\Delta U_x}$	$\boldsymbol{\Delta^2 U_x}$	$\boldsymbol{\Delta^3 U_x}$	$\boldsymbol{\Delta^4 U_x}$
20	120				
		145	55		
40	265				
		200		-5	
60	465				
		250	50		-145
80	715				
		150		-150	
100	865				
			-100		

Number of persons whose earnings is between Rs. 40-50 is $U_{50} - U_{40}$

We have $U_{40} = 265$

Find U_{50}

$$U_{50} = U_{a+rh}$$

$$a = 20, h = 20$$

$$50 = 20 + 20r$$

$$r = 1.5$$

By Newton-Gergory formula,

$$U_{50} = 120 + 1.5 \times \frac{145}{1} + 1.5(1.5-1) \times \frac{55}{2!} + 1.5\,(1.5-1)(1.5-2) \times \frac{-5}{3!} + (1.5)(1.5-1)(1.5-2)(1.5-3)\frac{-145}{4!}$$

$$= 120 + 217.5 + 20.625 + 0.3125 - 3.3984$$

$$= 355$$

Number of persons whose earnings is between Rs. 40-50 is

$$U_{50} - U_{40} = 355 - 265$$

$$= 90$$

Problem:

The following data gives the melting point of an alloy of lead and zinc. $\theta$is the temperature in degrees centigrade and x is the temperature of lead.

X	40	50	60	70	80	90
θ	184	204	226	250	276	304

Find $\theta\ when\ (i)x = 42\ \ (ii)x = 38$

Solution :

To find U_{42}, a = 40, h=10

Hence $U_{a+rh} = U_{42}$

$$a + rh = 42$$

$$40 + 10r = 42$$

$$r = 0.2$$

$$\therefore\ U_{42} = 184 + 0.2 \times \frac{20}{1!} + 0.2(0.2-1) \times \frac{2}{2!}$$

$$= 184 + 4 - 0.16 = \ 187.84$$

X	θ	$\Delta\theta$	$\Delta^2\theta$	$\Delta^3\theta$
40	184			
		20	2	
50	204			0
		22	2	
60	226			0
		24	2	
70	250			0
		26	2	
80	276			
		28		
90	304			

(ii) To find U_{38}

$$U_{40} \ + 10\,r = \ U_{38}$$

$$Hence, 40 + 10r = 38$$

$$r = -0.2$$

$$\therefore \ U_{38} == \ 184 \ + \ (-0.2)\text{x}\frac{20}{1!} + (-0.2)(-0.2-1)\text{x} \ \frac{2}{2!}$$

$$= \ \ 184 - 4 + 0.24$$

$$= 180.24$$

Problem:

The following table gives the census population of a town for the years 1931 – 1971. Estimate the population (i) for the year 1965, (ii) for the year 1933 by using an appropriate interpolation formula.

Year	1931	1941	1951	1961	1971
Population in lakhs	36	66	81	93	101

Solution :

i) To find the population in the year 1965

Year	Population U_x	∇U_x	$\nabla^2 U_x$	$\nabla^3 U_x$	$\nabla^4 U_x$
1931	36				
		30			
1941	66		-15		
		15		12	
1951	81		-3		-13
		12		-1	
1961	93		-4		
		8			
1971	101				

To find U_{1965}

$$Here\ a + nh = 1971, h = 10$$

$$U_{a+nh+rh} \ = \ U_{1965}$$

$$a + nh + rh = 1965$$

$$\therefore 1971 + rh = 1965$$

$$r = \ -0.6$$

Applying Newton's backward difference formula for interpolation we get,

$$U_{1965} = 101 + \frac{(-0.6)8}{1!} + \frac{(-0.6)(-0.6+1)}{2!}(-4) + \frac{-0.6\ (-0.6+1)(-0.6+2)}{3!}(-1) + \frac{-0.6\ (-0.6+1)(-0.6+2)(-0.6+3)}{4!}(-13)$$

$$= 101 - 4.8 + 0.48 + 0.056 + 0.4368$$

$$= 97.1728.$$

ii) To find the population in 1993

Year	Population U_x	$\Delta \boldsymbol{U_x}$	$\Delta^2 \boldsymbol{U_x}$	$\Delta^3 \boldsymbol{U_x}$	$\Delta^4 \boldsymbol{U_x}$
1931	36				
		30	-15		
1941	66				
		15		12	
1951	81				
		12	-3		-13
1961	93				
		8		-1	
1971	101				
			-4		

To find U_{1933}

$$a = 1931\ and\ h = 10$$

$$U_{a+rh} = U_{1931}$$

$$Hence\ a + rh = 1933$$

$$\therefore 1931 + 10r = 1933$$

$$r = 0.2$$

Applying Newton –Gregory's formula we get,

$$U_{1933} = 36 + 0.2 \times\ 30 + \frac{0.2(0.2-1)}{2!}(-15) + \frac{0.2(0.2-1)(0.2-2)}{3!} \times 15 + \frac{0.2(0.2-1)(0.2-2)(0.2-3)}{4!} \times (-13)$$

$$= 36 + 6 + 1.2 + 0.576 + 0.4368$$

$$= 44.2128$$

Lagrange's formula

The Lagrange's formula becomes

$$U_x \; = \; \frac{(x-a_2)(x-a_3)\,\ldots\ldots(x-a_n)}{(a_1-a_2)(a_1-a_3)\,\ldots\ldots(a_1-a_n)}U_{a_1} + \frac{(x-a_1)(x-a_3)\,\ldots\ldots(x-a_n)}{(a_2-a_1)(a_2-a_3)\,\ldots\ldots(a_2-a_n)}U_{a_2}$$

$$+\cdots+ \; \frac{(x-a_1)(x-a_2)\,\ldots\ldots(x-a_{n-1})}{(a_n-a_1)(a_n-a_2)\,\ldots\ldots(a_n-a_{n-1})}U_{a_n}$$

Problem:

Find U_5 given that $U_1 = 4;\;\; U_2 = 7;\; U_4 = 13;\;\; and\; U_7 = 30$.

Solution :

Take $a_1 = 1,\; a_2 = 2;\;\; a_3 = 4;\;\; a_4 = 7\; and\; x = 5$.

$Substituting\ in\ Lagrange's formula\ we\ get,$

U_5

$$= \left[\frac{(5-2)(5-4)(5-7)}{(1-2)(1-4)(1-7)}\right] \times\; 4 + \left[\frac{(5-1)(5-4)(5-7)}{(2-1)(2-4)(2-7)}\right]\times 7$$

$$+\; \left[\frac{(5-1)(5-2)(5-7)}{(4-1)(4-2)(4-7)}\right]\times 13 + \left[\frac{(5-1)(5-2)(5-4)}{(7-1)(7-2)(7-4)}\right]\times 30$$

$$= \; \left[\frac{3\;\times 1\times(-2)}{(-1)(-3)(-6)}\right]\times\; 4 + \left[\frac{4\times 1\times(-2)}{1(-2)(-5)}\right]\times 7 + \left[\frac{4\times 3\times(-2)}{3\times 2\times(-3)}\right]$$

$$\times 13\left[\frac{4\;\times\;3\times 1}{6\times 5\times 3}\right]\times 30$$

$$= \frac{4}{3} - \frac{28}{5} + \frac{52}{3} + 4$$

$= \;17.06$

Problem:

Find the form of the function U_x for the following data. Find U_3

X	0	1	2	5
U_x	2	3	12	147

Solution :

Here $a_1 = 0;\ \ a_2 = 1;\ \ a_3 = 2;\ a_4 = 5$

$\therefore\ U_{a1} = 2;\ Ua_2 = 1;\ \ Ua_3 = 12;\ \ Ua_4 = 147$

Applying Lagrange's formula, we get

$$U_x = \left[\frac{(x-1)(x-2)(x-5)}{(0-1)(0-2)(0-5)}\right] \times 2 + \left[\frac{(x-0)(x-2)(x-5)}{(1-0)(1-2)(1-5)}\right] \times 1 + \left[\frac{(x-0)(x-1)(x-5)}{(2-0)(2-1)(2-5)}\right] \times 12 + \left[\frac{(x-0)(x-1)(x-2)}{(5-0)(5-1)(5-2)}\right] \times\ 147$$

$$= \frac{-(x^3 - 8x^2 + 17x - 10)}{5} + \frac{3(x^3 - 7x^2 + 10x)}{4} - 2(x^3 - 6x^2 + 5x)$$

$$+ \frac{x^3 - 3x^2 + 2x}{60} \times 147$$

$$= \frac{1}{60}[x^3(-12 + 45 - 120 + 147) + x^2(96 - 315 + 720 - 441)$$

$$+ x(-204 + 450 - 600 + 294) + 120]$$

$$= \frac{1}{60}[60x^3 + 60x^2 - 60x + 120]$$

$$\therefore\ U_x = x^3 + x^2 - x + 2$$

$$\therefore\ U_3 = 3^3 + 3^2 - 3 + 2$$

$$= 35$$

Problem:

Determine by Lagrange's formula the percentage number of criminals under 35 years.

Age	% number of criminals
Under 25 years	52.0
Under 30 years	67.3
Under 40 years	84.1
Under 50 years	94.4

Solution :

To find U_{35}. $a_1 = 25;\ a_2 = 30;\ a_3 = 40;\ a_4 = 50$

$$Ua_1 = 52;\ Ua_2 = 67.3;\ Ua_3 = 84.1, Ua_4 = 94.4$$

Applying Lagrange's formula we get

$U_{35} =$

$$\left[\frac{(35-30)(35-40)(35-50)}{(25-30)(25-40)(25-40)}\right] \times 52.0 + \left[\frac{(35-25)(35-40)(35-50)}{(30-25)(30-40)(30-50)}\right] \times 67.3 +$$

$$\left[\frac{(35-25)(35-30)(35-50)}{(40-25)(40-30)(40-50)}\right] \times 84.1 + \left[\frac{(35-25)(35-30)(35-40)}{(50-25)(50-30)(50-40)}\right] \times 94.4$$

$$= \frac{(-1) \times 52.0}{5} + \frac{3 \times 67.3}{4} + \frac{1 \times 84.1}{2} - \frac{1 \times 94.4}{20}$$

$$= -10.40 + 50.38 + 42.05 - 4.72$$

$$= 77.31$$

Hence the estimated % number of criminals under 35 years is 77.31.

Prepared by

Ms. C. KANI

Assistant Professor of Mathematics, St. Jude's College,
Thoothoor - 629 176, Kanyakumari District.