Tribhuvan University
Institute of Science and Technology

Bachelor Level/First Year/Second Semester/Science
Computer Science and Information Technology STA 164
(Statistics I)

Full Marks: 60
Pass Marks: 24
Time: 3 Hours

Candidates are required to give their answers in their own words as far as practicable.

All notations have the usual meanings.

TU QUESTIONS-ANSWERS 2075

Long answer questions

Group A

Attempt ant two questions. **(2×10=20)**

1. **Distinguish between absolute and relative measure of dispersion. Two computer manufacturers A and B compete for profitable and prestigious contract. In their rivalry, each claim that their computer a consistent. For this it was decided to start execution of the same program simultaneously on 50 computers of each company and recorded time as given below;**

Time(sec)	0–2	2-4	4-6	6-8	8-10	10-12
No. of computers manufactured by company A	5	16	13	7	5	4
No. of computers manufactured by company B	2	7	12	19	9	1

2. ***Which company's computer is more consistent?***

Solution:

Difference between absolute and relative measure of dispersion

Absolute measure of dispersion	Relative measure of dispersion
1. It gives idea about amount of dispersion in a set of observations	1. It gives comparison of dispersion in two or more than two set of observations
2. It has same unit as set of original observation	2. It has no unit
3. Measures are range, quartile deviation, mean deviation, standard deviation etc.	3. Measures are coefficient of range, coefficient of quartile deviation, coefficient of mean deviation, coefficient of standard deviation

Time	No. of A (f_A)	No. of B (f_B)	Mid time(x)	f_Ax	f_Ax^2	f_Bx	f_Bx^2
0 – 2	5	2	1	5	5	2	2
2 - 4	16	7	3	48	144	21	63
4 - 6	13	12	5	65	325	60	300
6 – 8	7	19	7	49	343	133	931

8 – 10	5	9	9	45	405	81	729
10 – 12	4	1	11	44	484	11	121
	$N_A = \Sigma f_A =$ 50	$N_B = \Sigma f_B =$ 50		$\Sigma f_A x =$ 256	$\Sigma f_A x^2 =$ 1706	$\Sigma f_B x =$ 308	$\Sigma f_B x^2 =$ 2146

Now,

$$\bar{x} = \frac{\Sigma fA^x}{NA} = 256/50 = 5.12$$

$$\sigma_A = \sqrt{\frac{\Sigma f_A x^2}{N_A} - \left(\frac{\Sigma f_A x}{N_A}\right)^2} = \sqrt{\frac{1706}{50} - (5.12)^2} = \sqrt{34.2 - 26.214} = \sqrt{7.986}$$

$$= \sqrt{7.986} = 2.825$$

$$\bar{x}_B = \frac{\Sigma f_B x}{N_B} = 308/50 = 6.16$$

$$\sigma_B = \sqrt{\frac{\Sigma f_B x^2}{N_B} - \left(\frac{\Sigma f_B x}{N_B}\right)^2} = \sqrt{\frac{1706}{50} - (6.16)^2} = \sqrt{42.92 - 37.945} = 2.23$$

$$CV_A = \frac{\sigma_A}{\bar{x}_A} \times 100\% = \frac{2.825}{5.12} \times 100\% = 55.17\%$$

$$CV_B = \frac{\sigma_B}{\bar{x}_B} \times 100\% = \frac{2.23}{6.16} \times 100\% = 36.2\%$$

Here, $CV_B = 36.2\% < CV_A = 55.17\%$

Hence computer of company B is more consistent.

3. **In a certain type of metal test specimen, the effects of normal stress on a specimen is known to be functionally related to shear resistance. The following table gives the data on the two variables**

Normal stress	26	25	28	23	27	23	24	28	26
Shear resistance	22	27	24	27	23	25	26	22	21

(i) Identify which one is response variable and fit a simple regression line assuming that the relationship between them is linear.

(ii) Interpret the regression coefficient with reference to your problem.

(iii) Obtain coefficient of determination and interpret this

(iv) Based on fitted model, predict the shear resistance for a normal stress of 30 kilogram per square centimeter

Solution:

Here Sher resistance depends upon Normal stress hence Sher resistance is response variable. To fit linear regression line

Normal stress(x)	Shear resistance (y)	xy	x^2	Y^2
26	22	572	676	484
25	27	675	625	729
28	24	672	784	576
23	27	621	529	729
27	23	621	729	529
23	25	575	529	625
24	26	624	576	676
28	22	616	784	484
26	21	546	676	441
$\Sigma x = 230$	$\Sigma y = 217$	$\Sigma xy = 5522$	$\Sigma x^2 = 5908$	$\Sigma y^2 = 5273$

To fit, $y = a+bx$

$\Sigma y = na + b\Sigma x$

Or, $217 = 9a + 230b$ (i)

$\Sigma xy = a\Sigma x + b\Sigma x^2$

$5522 = 230a + 5908b$ (ii)

Solving (i) and (ii)

Coeff of a	Coeff of b	Constant
9	230	217
230	5908	5522

$$D = \begin{vmatrix} 9 & 230 \\ 230 & 5908 \end{vmatrix} = 9 \times 5908 - 230 \times 230 = 272$$

$$D_1 = \begin{vmatrix} 217 & 230 \\ 5522 & 5908 \end{vmatrix} = 217 \times 5908 - 5522 \times 230 = 11976$$

$$D_1 = \begin{vmatrix} 9 & 217 \\ 230 & 5522 \end{vmatrix} = 9 \times 5522 - 230 \times 217 = -212$$

Now,

$$a = \frac{D_1}{D} = \frac{11976}{272} = 44.029$$

$$b = \frac{D_2}{D} = \frac{-212}{272} = -0.7794$$

Hence regression equation is $y = a + bx$

or, $y = 44.029 - 0.779x$

Here regression coefficient is -0.779 it means shear resistance decreases by 0.779 for unit increase in normal stress.

$$\text{TSS (Total sum of square)} = \Sigma\left(y - \bar{y}\right)^2 = \Sigma y^2 - n\bar{y}^2 = 5273 - 9 \times \left(\frac{217}{9}\right)^2$$

$$= 5273 - 5232.11 = 40.88$$

$$\text{SSE (Sum of square due to error)} = \Sigma\left(y - \bar{y}\right)^2 = \Sigma y^2 - a\Sigma y - b\Sigma y$$

$= 5273 - 44.029 \times 217 (-0.7794) \times 5522\ \ 22.55$

SSR (Sum of square due to regression) = TSS – SSE = 40.88 – 22.55 = 18.32

$$\text{Coefficient of determination } (R^2) = \frac{SSR}{TSS} = \frac{18.32}{40} = 0.448 = 44.8\%$$

It means 44.8% variation in shear resistance is explained by normal stress.

4. (a) **What do you understand by binomial distribution? What are its main features?**

(b) **What do you mean by marginal probability distribution? Write down its properties**

Solution:

A discrete random variable following Binomial distribution is called Binomial variate. If X denotes the number of successes in n trials which can take the values 0, 1,2, …. n;

Random variable x is said to have binomial distribution if it's probability mass function is given by

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

Here n and p are parameters of the distribution. A random variable x following Binomial distribution is denoted by $x \sim B(n, p)$.

Features of binomial distribution

(i) It is a discrete distribution assuming nonnegative values of a random variable.

(ii) The parameters of Binomial distribution are n and p. so this distribution is also known as bi parametric i.e., having two parameters.

(iii) Mean = E(X) = np

(iv) Variance = V(X) = npq with maximum value of variance = $\frac{n}{4}$ when p = q.

(v) Mean ≥ Variance

(vi) Coefficient Skewness = $\beta_1 = \frac{(q-p)^2}{npq} = \frac{(1-2p)^2}{npq}$ and $\gamma_1 = \sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}}$

(vii) Coefficient of Kurtosis = $\beta_2 = 3 + \frac{1-6pq}{npq}$

(viii) When probability of success are same for both binomial variate the sum of two binomial variate is also a binomial variate. i.e., $X_1 \sim B(n_1, p)$ and $X_2 \sim B(n_2, p)$ then $X_1 \pm X_2 \sim B(n_1 \pm n_2, p)$.

Solution (b):

Let (X,Y) be two dimensional discrete random variable taking values (x_i, y_j) with probability mass function $p_{ij} = P(x_i, y_j)$, i = 1, 2, 3,.... n and j = 1, 2, 3,.... m. then the probability mass function of one discrete random variable obtained by summing the joint pmf over other discrete random variable is called marginal probability mass function.

The probability of discrete random variable X denoted by $p_i = P(x_i) = P(X=x_i) = P(X = x_i \text{ and } Y = y_1) + P(X=x_i \text{ and } Y = y_2) + P(X=x_i \text{ and } Y = y_3) + \ldots + P(X = x_i \text{ and } Y = y_m)$

= $p_{i1} + p_{i2} + p_{i3} + \ldots p_{im} = \sum_{j=1}^{m} p_{ij} = p_i$. is called marginal probability mass function of random variable X if it satisfies

i. $p_i \geq 0$

ii. $\sum_{i=1}^{n} p_i = 1$

Similarly,

The probability of discrete random variable Y denoted by $p_j = P(y_j) = P(Y=y_j)$

= $P(X=x_1 \text{ and } Y = y_j) + P(X=x_2 \text{ and } Y = y_j) + P(X=x_3 \text{ and } Y = y_j) + \ldots + P(X = x_n \text{ and } Y = y_j)$

= $P_{1j} + P_{2j} + P_{3j} + \ldots P_{nj} = \sum_{i=1}^{n} p_{ij} = p.j$ is called marginal probability mass function of random variable Y if it satisfies.

i. $p.j \geq 0$

ii. $\sum_{j=1}^{m} p.j = 1$

Let (X,Y) be two dimensional continuous random variable taking values $(-\infty \leq X \leq \infty, -\infty \leq Y \leq$ with joint probability density function f(x, y) then probability function of only one continuous random variable obtained by integrating the joint pdf with respect to other continuous random variable is marginal pdf.

The probability function of continuous random variable x denoted by $f(x) = \int_{-\infty}^{\infty}$ f(x, y)dy is called marginal pdf of X if it satisfies

i. $f(x) \geq 0$

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

Similarly,

The probability function of continuous random variable y denoted by $f(y) = \int_{-\infty}^{\infty}$ f(x, y)dx is called marginal pdf of Y if it satisfies

i. $f(y) \geq 0$

ii. $\int_{-\infty}^{\infty} f(y)\,dy = 1$

Marginalprobability of random variable x such that $a \le x \le b$ and marginal probability of random variable y such that $c \le y \le d$ is given by

$P(a \le x \le b) = \sum_{i=a}^{b} p_i$ for discrete random variables x

$P(c \le x \le d) = \sum_{j=c}^{d} P_{.j}$ for discrete random variables y

$P(a \le x \le b) = \int_a^b f(x)\,dx$ for continuous random variables x

$P(c \le y \le d) = \int_c^d f(y)\,dy$ for continuous random variables y

Short answer questions

Group B

Attempt any eight questions **(8×5=40)**

5. Measurement of computer chip's thickness (in nanometer) is recorded below

Thickness of chips (in nanometers)	34-39	39 - 44	44 - 49	49 - 54	54 - 59	Total
Number of computers	3	11	16	25	5	60

Find mode of thickness of computer chips and interpret the result.

Solution:

Here maximum frequency is 25 for which corresponding class is (49 – 54)

Hence modal class is (49 – 54)

$L = 49$, $h = 5$, $f_0 = 16$, $f_1 = 25$, $f_2 = 5$

$\Delta_1 = f_1 - f_0 = 25 - 16 = 9$, $\Delta_2 = f_1 - f_2 = 25 - 5 = 20$

$$\text{Mode } (M_0) = L + \frac{\Delta_1}{\Delta_1\ \Delta_2} \times \frac{\Delta_1}{\Delta_1\ \Delta_2} \times h = 49 + \frac{9}{9+20} \times 5 = 50.55$$

6. Calculate Q_1, D_5 and P_{70} from the following data and interpret the results;

Respiratory rate	10	15	20	25	30	35	40	45	50
No of persons	8	12	36	25	28	18	9	12	6

Solution:

Respiratory rate (x)	No. of persons (f)	cf
10	8	8
15	12	20
20	36	56
25	25	81
30	28	109
35	18	127
40	9	136
45	12	148
50	6	154
	$N = \Sigma f = 154$	

Here, $N = 154$

To find Q_1

$$\frac{N+1}{4} = \frac{154+1}{4} = 38.75$$

Cf just greater than 38.75 is 56 for which corresponding value is 20. Hence $Q_1 = 20$

Hence 25% persons have respiratory rate below 20 and 75% persons have respiratory rate above 20

$$\text{To find } D_5 = 5\left(\frac{N+1}{10}\right) = 5\left(\frac{154+1}{10}\right) = 77.5$$

Cf just greater than 77.5 is 81 for which corresponding value is 25. Hence $D_5 = 25$

Hence 50% persons have respiratory rate below 25 and 50% persons have respiratory rate above 25.

To find P_{70}

$$70\left(\frac{N+1}{100}\right) = 70\left(\frac{154+1}{100}\right) = 108.5$$

Cf just greater than 108.5 is 109 for which corresponding value is 30. Hence $P_{70} = 30$

Hence, 70% persons have respiratory rate below 30 and 30% persons have respiratory rate above 30.

7. Define a random variable. For the following bivariate probability distribution of X and Y find (i) Marginal probability mass function of X and Y (ii) P(X ≤ 1, y = 2) (iii) P(X ≤ 1)

y \ x	1	2	3	4	5	6
0	0	0	1/32	2/32	2/32	3/32
1	1/16	1/16	1/8	1/8	1/8	1/8
2	1/32	1/32	1/64	1/64	1/64	1/64

Solution:

It is a rule which assigns one and only one real value to each outcome of a random experiment. It is also called a real valued function defined on a sample space of the random experiment. Random variables are denoted by capital letters X, Y, Z etc. and value taken by the random variables are denoted by small letters x, y, z and so on.

For example,

In tossing a coin, the number of heads is a rule which assigns 1 to the outcome head and 0 to the outcome tail. Hence the number of head is a random variable X taking value 1 and 0 with probabilities $\frac{1}{2}$ and $\frac{1}{2}$ respectively.

Random variable is divided into two types:

(i) **Discrete random variable:** A random variable is called discrete if it takes integer values. It is also called real valued function defined on discrete sample space. e.g. number of printing mistakes in a page of a book, number of students enrolled in a college, number of defective items in a sample of certain size, number of patients admitted in a hospital , number of files in folders etc.

(ii) **Continuous random variable:** A random variable is called continuous if it takes all possible values within a certain interval. e.g. amount of rainfall in rainy season, height of an individual, temperature recorded in a particular day, etc.

y \ x	1	2	3	4	5	6	P(X)
0	0	0	1/32	2/32	2/32	3/32	8/32
1	1/16	1/16	1/8	1/8	1/8	1/8	10/16
2	1/32	1/32	1/64	1/64	1/64	1/64	8/64
P(y)	3/32	3/32	11/64	13/64	13/64	15/64	1

Here,

Marginal probability mass function of X

$P(X = 0) = 8/32$, $P(X = 1) = 10/16$, $P(X = 2) = 8/64$

Marginal probability mass function of Y

$P(y = 1) = 3/32$, $P(y = 2) = 3/32$, $P(Y = 3) = 11/64$, $P(Y = 4) = 13/64$, $P(Y = 5) = 13/64$,

$P(y = 6) = 15/64$

$P(X \le 1, y = 2) = P(X = 0, Y = 2) + P(X = 1, Y = 2) = 0 + 1/16 = 1/16$

$P(X \le 1) = P(X = 0) + P(X = 1) = 8/32 + 10/16 = 28/32$

8. **If two random variables have the joint probability density function**
$f(x,y) = ke^{(-x+y)}$, $0 < x < \infty$, $0 < y < \infty$
0, otherwise
Find (i) k (ii) conditional probability density function of X given Y (iii) Var (3X+2Y)

Solution:

We know,

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

or, $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{ke^{-(x+y)}dy\}\,dx = 1$

or, $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ke^{-x}e^{-y}dy\}\,dx = 1$

or, $k\int_0^{\infty} e^{-x}\left\{\int_0^{\infty} e^{-y}y^{1-1}dy\right\}dx = 1$

or, $k\int_0^{\infty} e^{-x}x^{1-1}\,dx \times = 1$

or, $k = 1$

Hence, $f(x,y) = e^{-(x+y)}$

$f(x) = \int_{-\infty}^{\infty} f(x, y)\,dy$

$= \int_0^{\infty} e^{-(x+y)}\,dy = \int_0^{\infty} e^{-x}e^{-y}\,dy = e^{-x}\int_0^{\infty} e^{-y}\,y^{1-1}\,dy = e^{-x}$

$f(y) = \int_0^{\infty} f(x, y)\,dx$

$= \int_0^{\infty} e^{-(x+y)}\,dx = \int_0^{\infty} e^{-x}e^{-y}\,dx = e^{-y}\int_0^{\infty} e^{-x}x^{1-1}\,dx = e^{-y}$

Now, conditional pdf ox x given $y = f(x/y) = \frac{f(x, y)}{f(y)} = \frac{e^{-(x+y)}}{e^{-y}} = e^{-x}$

$E(x) = \int_{-\infty}^{\infty} x\,f(x)\,dx = \int_0^{\infty} xe^{-x}dx = \int_0^{\infty} e^{-x}x^{2-1}\,dx = \Gamma 2 = (2-1)! = 1! = 1$

$E(x^2) = \int_{-\infty}^{\infty} x^2\,f(x)\,dx = \int_0^{\infty} x^2e^{-x}dx = \int_0^{\infty} e^{-x}x^{3-1}\,dx = \Gamma 3 = (3-1)! = 2! = 2$

$E(y^2) = \int_{-\infty}^{\infty} y\,f(y)\,dx = \int_0^{\infty} ye^{-y}\,dx = \int_0^{\infty} e^{-y}y^{2-1}\,dx = \Gamma 2 = (2-1)! = 1! = 1$

$E(y^2) = \int_{-\infty}^{\infty} y^2\,f(y)\,dx = \int_0^{\infty} y^2e^{-y}dx = \int_0^{\infty} e^{-y}y^{3-1}\,dx = \Gamma 3 = (3-1)! = 2! = 2$

$E(xy) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\,f(x, y)\,dy\,dx = \int_0^{\infty}\int_0^{\infty} xye^{-(x+y)}\,dydx$

$= \int_0^{\infty}\left\{xe^{-x}\int_0^{\infty} ye^{-y}\,dy\right\}dx = \int_0^{\infty} xe^{-x}\,dx = 1$

$V(x) = E(x^2) - [E(x)]^2 = 2 - 1^2 = 1$

$V(y) = E(y^2) - [E(y)]^2 = 2 - 1^2 = 1$

$Cov(x, y) = E(xy) - E(x)\,E(y) = 1 - 1 \times 1 = 0$

Now, $V(3x+2y) = 9V(x) + 4\,V(y) + 12\,Cov(x, y) = 9 \times 1 + 4 \times 1 + 12 \times 0 = 13$

9. **A certain machine makes electrical resistors having a mean resistance of 40 ohms and standard deviations of 2 ohms. Assuming that the resistance follows a normal distribution**

(i) What percentage of resistors will have resistance exceeding 43 ohm?

(ii) What percentage of resistors will have resistance between 30 ohms to 45 ohms?

Solution:

Let X = resistance of resistor

Mean = $\mu = 40$ ohm

Standard deviation = $\sigma = 2$ ohm

$X \sim N(\mu, \sigma^2)$

Define $Z = \frac{x-\mu}{\sigma} = \frac{x-40}{2}$

$P(X > 43) = ?$

When $X = 43$, $Z = \frac{43-40}{2} = 1.5$

$P(X > 43) = P(Z > 1.5) = 0.5 - P(0 < Z < 1.5)$

$= 0.5 - 0.4332 = 0.0668 = 6.68\%$

$P(30 < X < 45) = ?$

When $X = 30$, $Z = \frac{30-40}{2} = -5$

When $X = 45$, $Z = \frac{45-40}{2} = 2.5$

$P(30 < X < 45) = P(-5 < Z < 2.5) = P(-5 < Z < 0) + P(0 < Z < 2.5)$

$= P(0 < Z < 5) + P(0 < Z < 2.5)$

$= 0.5 + P(0 < Z < 2.5)$

$= 0.5 + 0.4938 = 0.9938 = 99.38\%$

10. As part of study of the psychological correlates of success in athletes, the following measurements are obtained from members of Nepal national football team

Anger	6	7	5	21	13	5	13	14
Vigor	30	23	29	22	19	19	28	19

Calculate Spearman's rank correlation coefficient.

Solution:

Anger (x)	Vigor (y)	Rx	Ry	d = Rx - Ry	d^2
6	30	3	8	-5	25
7	23	4	5	-1	1
5	29	1.5	7	-5.5	30.25
21	22	8	4	4	16
13	19	5.5	2	3.5	12.25
5	19	1.5	2	-0.5	0.25
13	28	5.5	6	-0.5	0.25
14	19	7	2	5	25
				$\Sigma d = 0$	$\Sigma d^2 = 110$

Here,

$n = 8$, $\Sigma d^2 = 110$, $m_1 = 2$, $m_2 = 2$, $m_3 = 3$

Spearman's rank correlation coefficient = $R = 1 -$

$$= \frac{6\left[\Sigma d^2 + \frac{m_1(m_1^2-1)}{12}\right] + \frac{m_2(m_2^2-1)}{12} + \frac{m_3(m_3^2-1)}{12}}{n(n^2-1)}$$

$$= 1 - \frac{6\left[110 + \frac{2(4-1)}{12}\right] + \frac{2(4-1)}{12} + \frac{3(9-1)}{12}}{8(64-1)}$$

$$= 1 - \frac{6[110 + 0.5 + 0.5 + 2]}{504} = 1 - \frac{678}{504} = 1 - 1.345 = -0.345$$

11. **Compute percentile coefficient of kurtosis from the following data and interpret the result**

Hourly wage(Rs)	23 - 27	28 - 32	33 - 37	38 - 42	43 - 47	48 - 52
Number of workers	22	16	9	4	3	1

Solution:

Hourly wage	Number of workers (f)	cf
23 – 27	22	22
28 – 32	16	38
33 – 37	9	47
38 – 42	4	51
43 – 47	3	54
48 – 52	1	55
	N = Σf = 55	

Now,

To find P_{10},

$$\frac{10N}{100} = \frac{10 \times 55}{100} = 5.5$$

Cf just greater than 5.5 is 22 for which corresponding class is (23 – 27) . Hence P_{10} class is (23 – 27)

It is inclusive class. Hence adjusted class is (22.5 – 27.5)

L = 22.5, h = 5 , f = 22, cf = 0

$$P_{10} = L + \frac{\frac{10N}{100} - cf}{f} \times h = 22.5 + \frac{5.5 - 0}{22} \times 5 = 23.75$$

To find P_{25},

$$\frac{25N}{100} = \frac{25 \times 55}{100} = 13.75$$

Cf just greater than 13.75 is 22 for which corresponding class is (23 – 27). Hence P_{25} class is (23 – 27)

It is inclusive class. Hence adjusted class is (22.5 – 27.5)

L = 22.5, h = 5, f = 22, cf = 0

$$P_{25} = L\frac{\frac{25N}{100} - Cf}{f} \times h = 22.5 + \frac{13.75 - 0}{22} \times 5 = 25.625$$

To find P_{75}, $\frac{75N}{100} = \frac{75 \times 55}{100} = 41.25$

Cf just greater than 41.25 is 47 for which corresponding class is (33 – 37) . Hence P_{90} class is (33 – 37)

It is inclusive class. Hence adjusted class is (32.5 – 37.5)

L = 32.5, h = 5, f = 9, cf = 38

$$P_{75} = L\frac{\frac{75N}{100} - Cf}{f} \times h$$

$$= 32.5 + \frac{41.25 - 38}{9} \times 5 = 34.305$$

To find P_{92}

$$\frac{90N}{100} = \frac{90 \times 55}{100} = 49.5$$

Cfjusi greater than 49.5 is 51 for which corresponding class is (38 – 42) . Hence P_{90} class is (38 – 42)

It is inclusive class. Hence adjusted class is (37.5 – 42.5)

L = 37.5 , h = 5 , f = 4, cf = 47

$$P_{90} = L + \frac{\frac{90N}{100} - cf}{f} \times h = 37.5 + \frac{49.5 - 47}{4} \times 5 = 40.625$$

Percentile coefficient of kurtosis =

$$k = \frac{P_{75} - P_{25}}{2(P_{90} - P_{10})} = \frac{34.305 - 25.625}{2(40.625 - 23.75)} = 8.6805/33.75 = 0.2572$$

Here K = 0.2572 < 0.263 . Hence the distribution is Platykurtic.

2. Write the properties of Poisson distribution. Fit Poisson distribution and find expected frequencies

x	0	1	2	3	4	5	6	7
f	71	112	117	57	27	11	3	1

Solution:

Properties of Poisson distribution;

i) Poisson distribution is discrete distribution.

ii) It has only one parameter λ, hence it is also known as uni-parametric.

iii) Mean = λ

iv) Variance = λ

v) Mean = variance

vi) Coefficient of Skewness $\beta_1 = 1/\lambda$, $\gamma_1 = 1/\sqrt{\lambda}$.

vii) Coefficient of kurtosis $\beta_2 = 1/\lambda$, $\gamma_2 = 1/\sqrt{\lambda}$.

viii) Sum of two Poisson variate is Poisson variate i.e. if $X \sim P(\lambda_1)$, $X_2 \sim P(\lambda_2)$ then $X_1 \pm X_2 \sim P(\lambda_1 \pm \lambda_2)$.

ix) It is used in the case of waiting time analysis where np < 5 and probability of success is very low i.e. $p \to 0$ and $n \to \infty$.

To fit poisson distribution

x	f	fx	$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$	Expected frequency = N P(x)
0	71	0	$\frac{e^{-3.01}3.01^0}{0!} = 0.0492$	19.63 ≈ 20
1	112	112	$\frac{e^{-3.01}3.01^1}{1!} = 0.1483$	59.17 ≈ 59
2	117	234	$\frac{e^{-3.01}3.01^2}{2!} = 0.2232$	89.05 ≈ 89
3	57	171	$\frac{e^{-3.01}3.01^3}{3!} = 0.224$	89.37 ≈ 89
4	27	108	$\frac{e^{-3.01}3.01^4}{4!} = 0.1685$	67.23 ≈ 67
5	11	551	$\frac{e^{-3.01}3.01^5}{5!} = 0.1014$	40.45 ≈ 40
6	3	18	$\frac{e^{-3.01}3.01^6}{6!} = 0.0509$	20.3 ≈ 20
7	1	7	$\frac{e^{-3.01}3.01^7}{7!} = 0.0218$	8.73 ≈ 9
	N = Σf = 399	Σfx = 1201		

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{1201}{399} = 3.01$$

Hence, $\lambda = \bar{x} = 3.01$

13. Define primary data and secondary data and explain the difference between them.

Solution:

Primary Data

The data which are originally collected by investigator or researcher for the first time for the purpose of statistical enquiry is called primary data. It is collected by government, an individual, institution and research bodies.

Secondary Data

The data that has been already collected for a particular purpose and used for next purpose is called secondary data. Hence one purpose primary data is another purpose secondary data. It is not new and original data.

Difference between primary data and secondary data

Primary data	Secondary data
1. It is first hand data	1. It is second hand data
2. It needs more fund, time and manpower	2. It saves fund, time and manpower
3. It is more reliable and accurate	3. It is less reliable and accurate
4. It is collected using different methods such as direct personal interview, indirect oral interview, mailed questionnaire, information through correspondents, schedule sent through enumerator, observation method etc.	4. It is obtained through different sources such as published source, unpublished source
5. It is specific to researcher's need.	5. It may or may not be specific to researcher's need

14. What do you mean by sampling? Explain non probability sampling with merits and demerits.

Solution:

When one by one study of all units of a population is not possible due to some factors like time, cost, manpower, recourses and destructive nature of study, we take a small representative part from the population for the study. This small representative part selected for the study from the population is called sample.

The process of selecting a sample from a population is called *sampling*. For example;

A housewife takes only two or three grains of rice as a sample from the cooking pan to know whether the rice is property cooked or not

A pathologist takes a syringe of blood as a sample to find out a disease.

Sampling are (i) probability sampling (ii) Non probability sampling

Non-probability Sampling:

It is defined as the method of sampling technique in which each unit in a sample is selected on the basis of personal judgment. There are several non-random sampling methods for selecting samples from a population. These are Judgement sampling, convenience sampling, Quota sampling, Snow ball samplingetc

Judgement Sampling

The sampling method sample is selected according to personal judgement of researcher or investigator is called judgement sampling. The investigator includes only those units in sample from population which they think most appropriate for the study.

Merits: (i) Simple method of sampling (ii) Practical method of quick decision on urgent need (iii) Better for small sample

Demerits: May not be representative of population

Convenience sampling:

The sampling method in which sample units are selected which are convenient to obtain is called convenience sampling. The representative units are selected because of availability and easy access.

Merits: (i) Quick method of data selection (ii)Can be used when population is not clearly defined

Demerits: (i) Sample may not represent the population as a whole (ii) It may be biased

Quota sampling

It is special case of stratified sampling without use of probability. It is judgement sampling with stratification. In this sampling quota are set up according to personal judgement of investigator. The size of quota for each stratum is proportional to size of stratum in population. Sampling is continue until pre-determined sample size obtained from each stratum.

Merits: (i) It is cheap method (ii) It is effective method

Demerits: (i) It may be biased (ii) It may not be representative of population

Snowball sampling:

The method of sampling in which sample are selected on referral basis. It is used by researcher to identify potential subjects of studies where subjects are hard to locate. A respondent is identified according to objective of study and other respondents are identified according to referral from the respondent. It is used for hidden population which are difficult for researcher to access.

Merits: (i) It is efficient method (ii) It can be used in hidden population

Demerits: (i) It is time consuming (ii) It has lack of representativeness

TU QUESTIONS-ANWERS 2076

1. What are the roles of measure of dispersion in descriptive statistics? Following table gives the frequency distribution of thickness of computer chips (in nanometer) manufactured by two companies.

Thickness of computer chips		5	10	15	20	25	30
Number of chips	Company A	10	15	24	20	18	13
	Company B	12	18	20	22	24	4

Which company many be considered more consistent in terms of thickness of computer chips? Apply appropriate descriptive statistics.

Ans: Dispersion is the spread of data from central tendency or average. It gives idea about how far data are scattered from average so that along with average value correct decision can be made. Let us consider two set of data have same average value but different in variation or dispersion then decision on the set of data can be made considering dispersion value.

Thickness of chips (x)	Number of chips of company A (f_A)	Number of chips of company B (f_B)	f_Ax	f_Bx	f_Ax^2	f_Bx^2
5	10	12	50	60	250	300
10	15	18	150	180	1500	1800
15	24	20	360	300	5400	4500
20	20	22	400	440	8000	8800
25	18	24	450	600	11250	15000
30	13	4	390	120	11700	3600
	$N_A=\Sigma f_A$ 100	$N_A=\Sigma f_A$ 100	$\Sigma f_Ax=1800$	$\Sigma f_Bx=1700$	$\Sigma f_Ax^2=38100$	$\Sigma f_Bx^2=34000$

Now,

$$\bar{x} f_Ax/N_A = \frac{1800}{100} = 18$$

$$\sigma_A = \sqrt{\frac{\Sigma f Ax^2}{N_A} - \left(\frac{\Sigma fAx}{N_A}\right)^2} = \sqrt{\frac{38100}{100} - (18)^2} = \sqrt{381 - 324} = \sqrt{57} = 7.54$$

$$\bar{x}_B = \frac{\Sigma f_Bx}{N_B} = 1700/100 = 17$$

$$\sigma_B = \sqrt{\frac{\Sigma f Ax^2}{N_B} - \left(\frac{\Sigma fBx}{N_B}\right)^2} = \sqrt{\frac{34000}{100} - (17)^2} = \sqrt{340 - 289} = \sqrt{51} = 7.14$$

$$CV_A = \frac{\sigma_A}{\bar{x}_A} \times 100\% = \frac{7.54}{18} \times 100\% = 41.88\%$$

$$CV_B = \frac{\sigma_B}{\bar{x}_B} \times 100\% = \frac{7.14}{17} \times 100\% = 42\%$$

Here, $CV_A = 41.88\% < CV_B = 42\%$, hence company a is more consistent in term of thickness of computer chip.

2. **A study was done to study the effect of ambient temperature on the electric power consumed by a chemical plant. Following table gives the data which were collected from an experimental pilot plant.**

Temperature (°F)	27	45	72	58	31	60	34	74
Electric power (BTU)	250	285	320	295	265	298	267	321

i. Identify which one is response variable, and fit a simple regression line, assuming that the relationship between them is linear.

ii. Interpret the regression coefficient with reference to your problem.

iii. Obtain coefficient of determination, and it interpret this.

Based on the fitted model in (a), predict the power consumption for an ambient temperature of 65°F.

Ans: Response variable is result of explanatory variable. It is also called dependent variable. Here electric power consumption is response variable.

To fit linear relationship between electric power consumption and temperature

Temperature (x)	Electric power (y)	xy	x^2
27	250	6750	729
45	285	12825	2025
72	320	23040	5184
58	295	17110	3364
31	265	8215	961
60	298	17880	3600
34	267	9078	1156
74	321	23754	5476
$\sum x = 401$	$\sum y = 2301$	$\sum xy = 118652$	$\sum x^2 = 22495$

To fit, $y = a+bx$

$\Sigma y = na + b\Sigma x$

Or, $2301 = 8a + 401b$ (i)

$\Sigma xy = a\Sigma x + b\Sigma x^2$

$11852 = 401a + 22495b$ (ii)

Solving (i) and (ii)

Coeff of a	Coeff of b	Constant
48	401	2301
401	22495	11852

$$D = \begin{vmatrix} 8 & 401 \\ 401 & 22495 \end{vmatrix} = 8 \times 22495 - 401 \times 401 = 19159$$

$$D_1 = \begin{vmatrix} 8 & 401 \\ 401 & 22495 \end{vmatrix} = 2301 \times 22495 - 11852 \times 401 = 47008343$$

$$D_2 = \begin{vmatrix} 8 & 2301 \\ 401 & 11852 \end{vmatrix} = 8 \times 11852 - 401 \times 2301 = -827885$$

Now, $a = \frac{D_1}{D} = \frac{47008343}{19159} = 2453.59$

$b = \frac{D_2}{D} = -\frac{827885}{19159} = -43.21$

Hence regression equation is $y = a + bx$

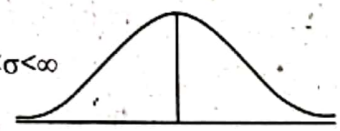
Or, $y = 2453.59 - 43.21x$

Here regression coefficient is -43.21 means per unit increase in temperature (F) decrease electric power by 43.21 (BTU).

3. (a) **Define Normal distribution. What are the main characteristic of a Normal distribution?**

Ans: A continuous random variable X is said to follow normal distribution if its probability density function is given by

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



With parameters μ and σ^2. It is written as $X \sim N(\mu, \sigma^2)$

Characteristics of normal distribution are

1. The distribution is continuous with two parameters μ and σ.
2. The curve is bell shaped and symmetrical about its mean.
3. Mean =median = mode
4. The distribution is unimodal
5. Maximum amplitude of curve is $\frac{1}{(\sigma\sqrt{2\pi})}$ and occurs at $x=\mu$.
6. Coefficient of skewness = $\beta_1 = 0$, $\gamma_1 = 0$.
7. Coefficient of Kurtosis = $\beta_2 = 3$, $\gamma_2 = 0$.
8. Quartile deviation is $\frac{2}{3}\sigma$.
9. Mean deviation is $\sqrt{\frac{2}{\pi}}\sigma \simeq \sigma$.
10. Q.D.:M.D.:S.D. = 10: 12: 15.

11. The curve is asymptotic to x axis.
12. Odd ordered moments are zero i.e. $\mu_{2n+1} = 0$.
13. Even ordered moments are given by relation $\mu_{2n} = (2n-1)\sigma^2\mu_{2n-2}$.
14. The theoretical range is $-\infty$ to ∞but the workable right is -3σ to 3σ.
15. The linear combination of independent normal variate is also a normal variate.
16. The point of inflexion of normal curve is $\mu \pm \sigma$.
17. The area under normal curve is unity.
18. $P(\mu - \sigma \le X \le \mu + \sigma) = 0.6826 \Rightarrow P(|Z| \le 1) = 0.6826$

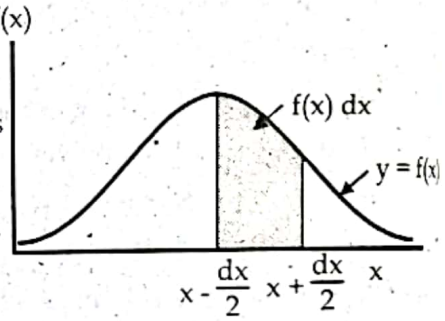
$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0.9544 \Rightarrow P(|Z| \le 2) = 0.9544$

$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0.9973 \Rightarrow P(|Z| \le 3) = 0.9973$

(b) What do you mean by probability density function? Write down its properties.

Ans: Let x be a continuous random variable and f(x) is continuous function of x, the probability that x lies in interval $x - \frac{dx}{2}$ and $x + \frac{dx}{2}$ of width dx is given by

$$f(x)dx = P[x - \frac{dx}{2} \le x \le x + \frac{dx}{2}]$$



f(x)dx represents the probability that x lies in the infinitesimal small interval $x - \frac{dx}{2}$ and $x + \frac{dx}{2}$.

The probability that x lies in interval [a, b] is given by $P[a \le x \le b] = \int_a^b f(x)\,dx$

. Then the function f(x) is called probability density function(pdf)

Properties of pdf are

i. $f(x) \ge 0$, for every x in the given range

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

Group B

Short Answer Questions

Attempt any EIGHT questions **(8×5=40)**

4. **The following table gives the installation time (in minutes) for hardware on 50 different computers.**

Installation time	0-10	10-20	20-30	30-40	40-50	Total
Number of computers	4	-	10	-	10	50

If the average installation times 30.2 minutes, find missing frequencies.

Ans:

Installation time	Number of computers (f)	Mid installation time (x)	fx
0 – 10	4	5	20
10 – 20	a=6	15	15a
20- 30	10	25	250
30 – 40	26 – a = 20	35	910 – 35a
40 – 50	10	45	450
	N=∑f= 50		∑fx = 1630 – 20a

Let us suppose number of computers for installation time 10 – 20 = a

Number of computers for installation time 30 – 40 = 50 – 4 – a – 10 – 10 = 26 – a

$$\bar{x} = \frac{\sum fx}{N}$$

or, $30.2 = (1630-20a)/50$

or, $30.2 \times 50 = 1630 - 20a$

or, $20a = 1630 - 1510$

or, $a = 6$

Also $20 - a = 26 - 6 = 20$

Hence missing frequency for installation time 10 – 20 is 6 and for installation time 30 – 40 is 20.

5. **The length of power failure in minute are recorded in the following table.**

Power failure time	22	23	24	25	26	27	28	Total
Frequency	2	5	7	10	4	3	2	33

Find Q_3, D_2 and P_{40} and interpret the results.

Ans:

Power failure time (x)	Frequency (f)	Cumulative frequency (cf)
22	2	2
23	5	7
24	7	14
25	10	24
26	4	28
27	3	31
28	2	33
	N = ∑f = 33	

Now,

To find Q_3

$$\frac{3(N+1)}{4} = \frac{3(33+1)}{4} = 25.5$$

Cf just greater than 25.5 is 28 for which corresponding value is 26

Hence $Q_3 = 26$

It means $Q_3 = 26$ divides 75% data in left side and 25% data in right side of the data arranged in ascending order

To find D_2

$$\frac{2(N+1)}{10} = \frac{2(33+1)}{10} = 6.8$$

Cf just greater than 6.8 is 7 for which corresponding value is 23

Hence $D_2 = 23$

It means $D_2 = 26$ divides 20% data in left side and 80% data in right side of the data arranged in ascending order

To find P_{40}

$$\frac{40(N+1)}{100} = \frac{40(33+1)}{100} = 13.6$$

Cf just greater than 13.6 is 14 for which corresponding value is 24

Hence $P_{40} = 24$

It means $P_{40} = 24$ divides 40% data in left side and 60% data in right side of the data arranged in ascending order

6. **A manufacturing company employs three analytical plans for the design and development of a particular product. For cost reasons, all three are used at varying times. In facts, plan 1, 2 and 3 are used for 30%, 20% and 50% of the products respectively. The defect rate in different procedures is as follows: $P(D/P_1) = 0.01$, $P(D/P_2) = 0.03$, $P(D/P_3) = 0.02$, here $P(D/P_j)$ is the probability of a defective product, given plan j. If a random product was observed and found to be defective, which plan was most likely used and thus responsible?**

Ans: Let P_1, P_2 and P_3 be plan 1,2 and 3 respectively

Here $P(P_1) = 30\% = 0.3$

$P(P_2) = 20\% = 0.2$

$P(P_3) = 50\% = 0.5$

$P(D/P_1) = 0.01$

$P(D/P_2) = 0.03$

$P(D/P_3) = 0.02$

$$P(D) = \sum_{i=1}^{3} P(P_i)\, P\left(\frac{D}{P_i}\right) = 0.3 \times 0.01 + 0.2 \times 0.03 + 0.5 + 0.02 = 0.019$$

$$P\left(\frac{P_1}{D}\right) = \frac{P(P_2)P\left(\frac{D}{P_2}\right)}{P(D)} = \frac{0.2 \times 0.03}{0.19} = 0.315$$

$$P(P_3/D) = \frac{P(P_3)P\left(\frac{D}{P_3}\right)}{P(D)} = \frac{0.5 \times 0..02}{0.019} = 0.526$$

Hence if a random item is defective then plan 3 is most responsible

7. **The random variable X has following probability distribution:**

X	0	1	2	3	4	5	6
P(X = x)	0.03	0.15	0.4	0.2	0.1	.07	.05

Find (i) E(X) and var(X) (ii) Calculate E(Y) if Y = 3X + 5.

Ans:

x	P(x)	xP(x)	x^2 P(x)
0	0.03	0	0
1	0.15	0.15	0.15
2	0.4	0.8	1.6
3	0.2	0.6	1.8
4	0.1	0.4	1.6
5	0.07	0.35	1.75
6	0.05	0.3	1.8
		$\sum xP(x) = 2.6$	$\sum x^2 P(x) = 8.7$

$E(x) = \sum xP(x) = 2.6$

$V(x) = E(x^2) - (E(x)^2 = \sum x^2 P(x) - (2.6)^2 = 8.7 - 6.76 = 1.94$

If $y = 3x + 5$

$E(y) = E(3x + 5) = 3E(x) + 5 = 3 \times 2.6 + 5 = 12.8$

8. **If two random variable have the joint probability density function**

$$f(x, y) = \begin{cases} k(2x + 3y), & \text{for } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find (i) constant k (ii) Conditional probability density function of X given Y (iii) Identify whether X and Y are independent.

Ans:

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

or, $\int_0^1\int_0^1 k(2x + 3y)\, dx\, dy = 1$

or, $k\int_0^1\int_0^1 (2x + 3y)\, dy\, dx = 1$

or, $k\int_0^1\int_0^1 (2xdy + 3ydy)\, dx = 1$

or, $k\int_0^1 \left[2xy + \frac{3y^2}{2}\right]_0^1 dx = 1$

or, $k\int_0^1 \left(2x+\frac{3}{2}\right) dx = 1$

or, $k\left[2\frac{x^2}{2}+\frac{3}{2}x\right]_0^1 = 1$

or, $k\,(1+3/2) = 1$

or, $k\frac{5}{2} = 1$

or, $k = \frac{2}{5}$

Hence, $f(x,y) = \frac{2}{5}(2x+3y)$

Now, $f(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$

$$= \int_0^1 \frac{2}{5}(2x + 3y)\, dy = \frac{2}{5}\left[2xy + \frac{3y^2}{2}\right]_0^1$$

$$= \frac{2}{5}\left(\frac{2x+3}{2}\right) = \frac{2}{5}\frac{(4x + 3)}{2} = \frac{(4x + 3)}{5}$$

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

$$= \int_0^1 2/5(2x + 3y)\, dx = 2/5\left[2\frac{x^2}{2} + 3xy\right]_0^1 = 2/5(1+3y)$$

$$f(x/y) = \frac{f(x, y)}{f(y)} = \frac{\frac{2}{5}(2x + 3y)}{\frac{2}{5}(1 + 3y)}) = \frac{2x + 3y}{1 + 3y}$$

To test independence

$$f(x)\, f(y) = \frac{(4x + 3)}{5} \times \frac{2(1 + 3y)}{5} = \frac{2(4x + 12xy + 9y + 3)}{25} \neq f(x, y) = \frac{2}{5}(2x+3y)$$

Hence x and y are not independent.

9. **A large chain retailer purchases a certain kind of electronic device from manufacturer. The manufacturer indicates that the defective rate of th device is 15%. The inspector randomly picks 10 items from a shipmen What is the probability that there will be at least one defective item amon these 10?**

Ans:

Let x = number of defective item

Probability of defective (p) = 15% = 0.15

Number of items (n) = 10

$P(x \geq 1) = ?$

$P(x \geq 1) = 1 - P(x < 1) = 1 - P(x = 0) = 1 - C(n, x)\, p^x q^{n-x}$

$= 1 - c(10, 0)\ (0,15)^0\ (1 - 0.15)^{10-0} = 1 - (0.85)^{10} = 0.803$

10. **Message arrive at an electronic message center at random times, an average of 9 messages per hour.**

(a) **What is the probability of receiving at least four messages during the next hour?**

(b) **What is the probability of receiving at most three message during the next hour?**

Ans: Let x = Number of message arriving at an electronic message center

Average message (λ) = 9 per hour

(a) $P(x \geq 4) = 1 - (x < 4) = 1 - \{P(x = 0) + P(x = 1) + P(x + 2) + P(x = 3)$

$= 1 - \{e^{-9}9^0/0! + e^{-9}9^1/1! + e^{-9}9^2/2! + e^{-9}9^3/3!\}$

$= 1 - e^{-9}\left(\frac{1 + 9 + 81}{\frac{2 + 729}{6}}\right)$

$= 1 - e^{-9}\,172 = 0.978$

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

$= \left\{\frac{e^{-9}9^0}{0!} + \frac{e^{-9}9^1}{1!} + \frac{e^{-9}9^2}{2!} + \frac{e^{-9}9^3}{3!}\right\}$

$= e^{-9}\left(\frac{1 + 9 + 81}{2} + \frac{729}{6}\right) = 172e^{-9} = 0.022$

Or, $P(X \leq 3) = 1 - P(X > 3) = 1 - P(x \geq 4) = 1 - 0.978 = 0.022$

11. **Following data represent the preference of 10 students studying B.Sc (CSIT) towards two brands of computers namely DELL and HP.**

Computer	Student preference									
DELL	5	2	9	8	1	10	3	4	6	7
HP	10	5	1	3	8	6	2	7	9	4

Apply appropriate statistical tool to measure whether the brand preference is correlated. Also interpret your result.

Ans:

Student preference of DELL laptop (R_1)	Student preference of HP laptop (R_2)	$d = R_1 - R_2$	d^2
5	10	-5	25
2	5	-3	9
9	1	8	64
8	3	5	25
1	8	-7	49
10	6	4	16
3	2	1	1

4	7	-3	9
6	9	-3	9
7	4	3	9
		$\sum d=0$	$\sum d^2 = 216$

It is preference (i.e rank) data. Hence Spearman's rank correlation coefficient is used to determine correlation

$$\text{Spearman's rank correlation (R)} = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 216}{10(100-1)}$$

$$= 1 - 1.35 = -0.35$$

There is low degree of negative correlation between student preference in two brands of laptop namely DELL and HP.

12. **What do you mean by measurement scale? Describe the different types of measurement scales used in statistics.**

Ans: Measurement consists of counting the number of units or parts of units displayed by objects and phenomena. Measurement is a process of assigning numbers or symbols to any facts or objects or products or items according to some rule.It is a tool by which individuals are distinguished on the variables of area under study. Scale is simply a range of levels or numbers used for measuring something. It is a set of all the different levels of symbols or numerals or something so constructed, from the lowest to highest, that these can be assigned by rule to objects or to items or to the individuals or to their behavior to whom it is applied.

Different measurement scales are used on the basis of nature of data. These measurement scales of variables under study are (a) Nominal scale (b) Ordinal scale (c) Interval scale (d) Interval scale

Nominal scale

It is the simplest type of scale, also known as categorical scale. It is lowest level of measurement scale. It is simply a system of assigning number or the symbols to objects or events to distinguish one from another or in order to label them. The symbols or the numbers have no numerical meaning. The arithmetic operations cannot be used for these numerals. The orders of the symbol have no mathematical meaning.

For example, gender, occupation, religion are measured in nominal scale. If we use scale 1 for male programmer and 2 for female programmer for measuring the gender of programmer, then 1 and 2 have no numeric meaning. It is used to distinguish male and female. In this case the number of a set of objects is not comparable to the other set. Mathematical operations such as addition, subtractions, multiplication, and division cannot be performed for the variable having nominal scale. Frequency count is possible in this case so that percentage, mode can be obtained, and chi square test can be performed for test of significance for variable having nominal scale.

Ordinal scale

The second and the lowest level of ordered scale is the ordinal scale. It is the quantification of items by *ranking*. In this scale, the numerals are arranged in some order but the gaps between the positions of the numerals are not made equal. It is used to rate preference of respondents. It indicates relative extent to which object possess certain characteristic. It represents qualitative values in ascending or descending order. The rank orders represent ordinal scale and mostly useful in scaling the qualitative phenomena.

For example, qualification levels, preference to different statistical software, preference to different made of Laptops are measured in ordinal scale. If we are going to study the computer literacy and we took the qualification level of people as primary, lower secondary, secondary , higher secondary , bachelor , master and Ph. D then we can use 1 , 2 , 3, , 4, , 5, 6 , 7 to represent primary , lower secondary , secondary , higher secondary , bachelor , master and Ph.D. respectively in ascending order. Mathematical operations such as addition, subtraction, multiplication, division cannot be performed for variable having ordinal scale. Frequency count is possible and due to ranking partition values can be determined for variable having ordinal scale. In this case median, mode, rank correlation also can be obtained.

Interval scale

In addition to ordering the data, this scale uses equidistant units to measure the difference between scores. It assumes data have equal intervals. This scale does not have absolute zero but only arbitrary zero. Interval scale is the developed form of the ordinal scale. The intervals between the ordered numerals are adjusted in terms of some rule.

For example, scale of temperature is an example of ordinal scale. In an increase in temperature from 320F to 420F and from 640F to 740F, we can say the increases are equal of 100F; but one cannot say that the temperature of 640F is twice as warm as the temperature of 320F. The 00C or 320F are the arbitrarily set points for the freezing point of water so the temperatures 320F and 640F are not viable to express in ratio because the zero is not a true zero but it is an arbitrary point. Arithmetic mean, standard deviation, correlation can be obtained for variable having ordinal scale. To test significance t-test and F test can be used.

Ratio scale

Ratio scale is the ideal scale and an extended form of Interval Scale. It is most powerful scale of measurement. It possesses the characteristics of nominal, ordinal and interval scale. Ratio scale has an absolute zero or true zero or natural zero of measurement. The true zero point or the initial point indicates the completely absence of that property of an object what is being measured. It can be expressed as relative level of measurement.

For example, the absolute zero or the natural zero in the centimeter scale indicates the absence of the length. Numbers on the scale indicates the actual amount of property being measured. The ratio involved in the ratio scale possesses measure property and it facilitates comparison, which is not possible in interval scale. Mathematical operations like addition, subtraction, multiplication and division all can be performed.

It represents actual number of variables and it is used to measure the physical dimensions. Some examples of ratio scale variables are disk space,

amount expenses in IT department in different year, age of programmers, etc. Geometric mean, Harmonic mean, coefficient of variation along with all other measures can be obtained for variable having ratio scale. All statistical techniques can be applied to the variable present in ratio scale.

The following is the properties of categories, rank, equal interval and true zero point of four scales.

Level of measurement	Property			
	Categories	Ranks	Equal intervals	True zero points
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Ratio	Yes	Yes	Yes	Yes

13. What is sampling? Discuss various probability sampling techniques with merits and demerits.

Ans: When one by one study of all units of a population is not possible due to some factors like time, cost, manpower, recourses and destructive nature of study, we take a small representative part from the population for the study. This small representative part selected for the study from the population is called sample.

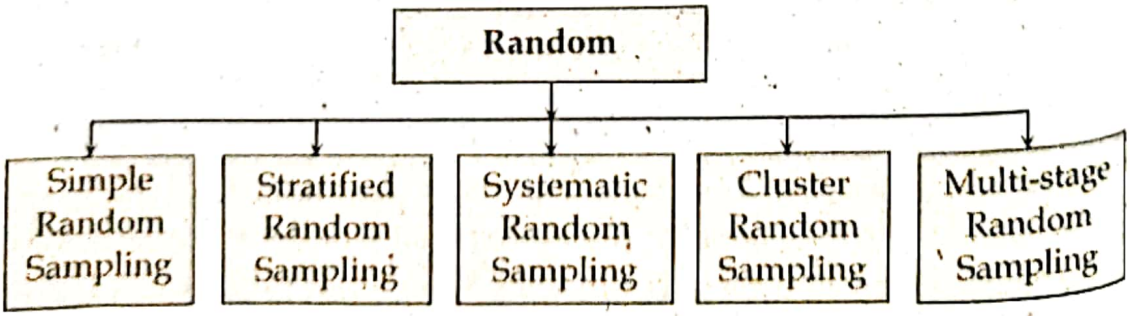
The process of selecting a sample from a population is called *sampling*. For example,

A housewife takes only two or three grains of rice as a sample from the cooking pan to know whether the rice is property cooked or not.

A pathologist takes a syringe of blood as a sample to find out a disease.

Probability Sampling Technique:

It is defined as the method of sampling technique in which each unit of the population has some fixed probability of being selected in the sample. There are several random sampling techniques for selecting samples from a population:



(a) **Simple random sampling:** It is the most common and the simplest method of sampling in which each sample unit is selected from a population with equal probability. There are two types simple random sampling:

(i) **Simple Random Sampling with Replacement:** A simple random sampling process; in which a unit is selected from a population, noted it and then returned back to the population before selecting a next unit from the population is called a simple random sampling with replacement (SRSWR). This process is repeated 'n' times to get a sample of 'n' units.

(ii) **Simple Random Sampling without Replacement:** A simple random sampling process; in which a unit is selected from a population, noted it and not returned back to the population before selecting a next unit from the population is called a simple random sampling without replacement (SRSWOR). This process is repeated 'n' times to get a sample of 'n' units.

Merits

1. Since the sample unit are selected at random giving each unit an equal chance of being selected, the element of subjectivity or personal bias is completely eliminated. As such a simple random sample is more representative of population as compared to the judgment or purposive sampling.
2. The statistician can ascertain the efficiency of the estimate of the parameters by considering the sampling distribution of the statistics (estimate) e.g., $\bar{Y}n$ as an estimate of $\bar{Y}N$ becomes more efficient as sample size n increases.

Demerits:

1. The selection of SRS requires an up to date frame, i.e. a completely catalogued population from which samples are to be drawn. Frequently it is virtually impossible to identify the units in the population before the sample is drawn and this restricts the use of simple random sampling technique.
2. Administrative inconvenient: A simple random sample may result in the selection of the sampling units which are widely spread geographically and in such case the cost of collecting the data may be much in terms of time and money.
3. At times, a simple random sample might give most non-random looking results. For example if we draw a random sample of size 13 from a pack of card, we may get all the cards of same suit. However the probability of such an outcome is extremely small.
4. For a given precision, simple random sampling usually requires larger sample size as compared to stratified random sampling.

(b) **Stratified Random Sampling:** When units in the population are not similar i.e. population under study is heterogeneous in nature, the population is divided into sub groups called strata before the sample is drawn. Then a simple random sample is drawn from each stratum in proportion to its size. The stratification of the population should be done as follows:

- The strata should be non-overlapping and should together comprise the whole population.
- The strata should be as possible as homogeneous within groups and heterogeneous between the groups.

Merits

1. The process of stratification makes the selected sample representative of the population because it includes all types of unit in the population.
2. The estimation of the population parameters are more precise.
3. In case of unsymmetrical distribution stratified sampling is the best method to use.

Demerits

1. This method consumes considerable amount of time and cost.

(c) **Systematic Random Sampling:** A sampling technique in which only first unit is selected with the help of random numbers and the rest get selected automatically according to some pre-designed pattern is known as systematic random sampling. This technique of selecting samples is usually recommended if the complete and up-to-date list of the sampling units is available and the units are arranged in some systematic order such as alphabetical, chronological, geographical order etc. Suppose N units of the population are numbered from 1 to Nin order and we want to draw a sample of size n such that $N = nk \Rightarrow k = \frac{N}{n}$ where k is called sample interval. Systematic sampling consists in selecting any unit at random from the first kunits numbered from 1 to k and then selecting every k^{th} unit in succession subsequently. For example suppose that we want to select 10 families from a list of families containing 80 families arranged systematically. Here, n = 10 and N = 80 then $k = \frac{N}{n} = \frac{80}{10} = 8$. We select any number from 1 to 8 at random, suppose selected number is 2. Then systematic sample of will consists of 10 families in the list as: 2, 10, 18, 26, 34, 42, 50, 58, 66, and 74. Here rest of the sample is obtained on adding k = 8 in succession for subsequent samples.

Merits

1. .Systematic sampling is operationally more convenient than simple random sampling or stratified random sampling. Time and work involved is also relatively much less. More over systematic sampling yields a sample which is evenly spread over the entire population.
2. .Systematic sampling may be more efficient than simple random sampling provided the frame (the list from which sample units are drawn) is arranged wholly at random. The most common approach to randomness is provided by alphabetic list.

Demerits

1. The main disadvantages of systematic sampling is that systematic samples are not in general wholly random.
2. If N is not a multiple of n then (i) the sample size is different from the required and (ii) sample mean is not an unbiased estimate of population mean.

3. Systematic sampling may yield highly biased estimate if there are periodic features associated with the sampling interval. i.e. if the frame (list) has a periodic feature and k is equals to or multiple of the period.

(d) **Cluster random sampling:** In *cluster random sampling*, we divide the population into discrete groups prior to sampling. The groups are termed as clusters and then we select some clusters as sample from all clusters using simple random sampling. After selecting the clusters in the sample, all the elements within the clusters are enumerated for the study. Clusters differ from strata in a manner that elements within a cluster are relatively heterogeneous and between clusters are relatively homogeneous. Thus, when dividing the population into different clusters, it should be kept in mind that there is greater variation within the clusters and less variation between the clusters.

This method of sampling is useful when the population consists of vary large number of similar groups which are geographically distant.

Merits

1. It is easier, cheaper and faster.
2. It is useful when sampling frame of elements may not be readily available.

Demerits

1. The efficiency decrease with increase in size of cluster.
2. Enumeration of sampling units within cluster is difficult when population size is large.

(e) **Multi-stage random sampling:** Multi-stage random sampling, sometimes called multi-stage cluster sampling, is a development of cluster sampling. In cluster sampling, clusters are considered as sampling units and all the elements in the selected clusters are enumerated completely. Instead of selecting all units of the selected clusters, a sample of desirable size can be drawn from selected clusters which may increase the efficiency is called multi-stage random sampling. In this technique, clusters which form the units of sampling at the first stage are called the *first stage units (fsu) or primary sampling units (psu)* and the elements within clusters are called *second-stage units (ssu)*. For example, in crop surveys for estimating yield of a crop in a district, a block may be considered a primary sampling unit, the villages the second stage units, the crop fields the third stage units, and a plot of fixed size the ultimate unit of sampling.

Merits

1. Flexible method of sampling.
2. Sample size reduces in each time, hence saves time and cost.

Demerits

1. In different stage sample should be taken carefully.
2. Less accurate.

TU QUESTIONS-ANWERS 2078

Long answer questions

Group A

Attempt ant two questions; (2X10 = 20)

1. **What are the different methods of measuring dispersion. Sample of polythene bags from two manufactures, A, B are tested by a prospective buyer for bursting pressure and the results are as follows:**

Bursting pressure		5-10	10-15	15-20	20-25	25-30	30-35
Number of bags manufactured by	A	2	9	29	54	11	5
	B	9	11	18	32	27	13

Which set of bags has more uniform pressure? If price are the same, which manufacture's bags would be preferred by buyer? Use appropriate statistical tool.

Ans: Dispersion is the scatterness of data from average. Different methods of measuring dispersion are

(i) Range (ii) Quartile deviation (iii) Mean deviation (iv) standard deviation

Range = Difference between largest and smallest observation. It is absolute measure

Coefficient of range = Difference between largest and smallest observation / sum of largest and smallest observation. It is relative measure

Quartile deviation= Half of the difference between upper quartile and lower quartile.. It is absolute measure

Coefficient of quartile deviation = Difference of upper and lower quartile/ sum of upper and lower quartile.. It is relative measure

Mean deviation = Average value of absolute value of deviation in which deviation is from mean or median or mode. It is absolute measure

Coefficient of mean deviation = Mean deviation/mean or median or mode from which deviation is taken. It is relative measure

Standard deviation = Positive square root of average of square of deviation in which deviation is from mean. It is absolute measure

Coefficient of standard deviation = Standard deviation/mean. It is relative measure

Coefficient of variation = Coefficient of standard deviation expressed in percentage.

Bursting pressure	Number of bags by A(fA)	Number of bags by B (fB)	Mid pressure (x)	$d=\frac{(x-22.5)}{5}$	fAd	fBd	fAd²	fBd²
5 - 10	2	9	7.5	-3	-6	-27	18	81
10 - 15	9	11	12.5	-2	-18	-22	36	44
15 - 20	29	18	17.5	-1	-29	-18	29	29
20 - 25	54	32	22.5	0	0	0	0	0
25 - 30	11	27	27.5	1	11	27	11	11
30 - 35	5	13	32.5	2	10	26	20	40
	$N_A=110$	$N_B=110$			$\sum f_Ad=-26$	$\sum f_Bd=-14$	$\sum f_Ad^2=114$	$\sum f_Bd^2=205$

$$\bar{x}_A = A + \sum\frac{f_Ad}{N_A} \times h = 22.5 + \left(-\frac{26}{110}\right) \times 5 = 22.5 - 1.81 = 21,319$$

$$\sigma_A = \sqrt{\frac{\sum f_Ad^2}{N_A} - \left(\frac{\sum f_Ad}{N_A}\right)^2} \times h = \sqrt{\frac{114}{110} - \left(\frac{-26}{110}\right)^2} \times 5$$

$$= \sqrt{1.036 - 0.055} \times 5 = 4.952$$

$$\bar{x}_B = A + \sum\sum\frac{f_Bd}{N_B} \times h = 22.5 + \left(-\frac{14}{110}\right) \times 5 = 22.5 - 0.636 = 21.864$$

$$\sigma_B = \sqrt{\frac{\sum f_Bd^2}{N_B} - \left(\frac{\sum f_Bd}{N_B}\right)^2} \times h = \sqrt{\frac{205}{110} - \left(\frac{-14}{110}\right)^2} \times 5$$

$$= \sqrt{1.863 - 0.016} \times 5 = 6.795$$

$$CV_A = \frac{\sigma_A}{\bar{x}_A} \times 100\% = \frac{4.952}{21.319} \times 100\% = 23.22\%$$

$$CV_B = \frac{\sigma_B}{\bar{x}_B} \times 100\% = \frac{6.795}{21.864} \times 100\% = 31.07\%$$

Here, $CV_A = 23.22\% < CV_B = 31.07\%$.

Hence bags of manufacturer A have uniform pressure.

If price is same bags manufactured by A will be preferred by buyers.

2. **Write the properties of correlation coefficient. The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files, with the average size of 126 Kbytes and the standard deviation of 35 Kbytes. The average transmitted time was 0.04 seconds with the standard deviation 0.01 seconds. The correlation coefficient between the time and size was 0.86. Based on these data, fit a linear regression model and predict the time it will take to transmit a 400Kbyte file.**

Ans: Following are the important properties of Karl Pearson's Correlation Coefficient

i. Correlation coefficient lies between – 1 to 1 i.e., $-1 \le r \le +1$.

ii. Correlation coefficient is symmetrical i.e. $r_{xy} = r_{yx} = r$.

iii. Correlation coefficient is independent of change of origin and scale.

iv. Correlation coefficient is the geometric mean of two regression coefficient

$$r = \pm\sqrt{b_{yx} \times b_{xy}}$$

v. Correlation coefficient has no unit because of relative measure.

Let us consider ,time taken to transmit file = y and x= size of file

Here, $n = 30$, $\bar{x} = 126$, $\sigma_x = 35$, $\bar{y} = 0.04$, $\sigma_y = 0.01$ $r = 0.86$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.86 \times \frac{0.01}{35} = 0.000245$$

Now regression equation of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

or, $y - 0.04 = 0.000245(x - 126)$

or, $y = 0.000245x - 126 \times 0.000245 + 0.04$

or, $y = 0.00913 + 0.000245x$

When $x = 400$

$$y = 0.00913 + 0.000245 \times 400 = 0.107$$

Hence it takes 0.107 seconds to transmit 400 Kbyte file.

3. (a) What do you understand by Poisson distribution? What are its main features?

Ans: The discrete random variable X is said to follow Poisson distribution if its probability mass function is given by $P(X = x) = P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$; $X = 0, 1, 2, ...$

Here λ is the parameter of Poisson distribution and a random variable following Poisson distribution is denoted by $X \sim P(\lambda)$.

Characteristics of Poisson distribution are

i. Poisson distribution is discrete distribution.

ii. It has only one parameter λ, hence it is also known as uni-parametric.

iii. Mean = λ

iv. Variance = λ

v. Mean = variance

vi. For non-integer λ, it is the largest integer less than λ. For integerλ, $x = \lambda$ and $x = \lambda - 1$ are the two modes.

vii. Coefficient of Skewness$\beta_1 = 1/\lambda$, $\gamma_1 = 1/\sqrt{\lambda}$.

viii. Coefficient of kurtosis$\beta_2 = 3 + 1/\lambda$, $\gamma_2 = 1/\lambda$,

ix. Sum of two Poisson variate is Poisson variate i.e. if $X \sim P(\lambda_1)$, $X_2 \sim P(\lambda_2)$ then $X_1 \pm X_2 \sim P(\lambda_1 \pm \lambda_2)$.

x. It is used in the case of waiting time analysis where np< 5 and probability of success is very low i.e. $p \to 0$ and $n \to \infty$.

(b) **What do you mean by joint probability distribution function? Write down its properties.**

Ans: Let (X,Y) be two dimensional random variable then the cumulative distribution function or simply distribution function of the two dimensional random variable is the probability that random variable X takes value less than or equal to x and random variable Y takes value less than or equal to y. The joint probability distribution function is denoted by F(X,Y) and is given by

$F(X,Y) = P(X \le x, Y \le y) = \sum_x \sum_y P(x_i, y_j)$ for discrete random variable having joint pmf $P(x_i, y_j)$

$= \int_{-\infty}^{x} \left\{ \int_{-\infty}^{y} f(x, y)\, dy \right\} dx$ for continuous random variable having pdf f(x,y)

Properties of joint probability distribution function

i. The distribution function is defined for every pair of real number(X,Y)

ii. $0 \le F(x, y) \le 1$

iii. $F(-\infty, -\infty) = 0$

iv. $F(+\infty, +\infty) = 1$

v. $\frac{\partial^2 F(x, y)}{\partial x\, \partial y} = f(x, y)$

Group B

Short Answer Questions

Attempt any EIGHT questions (8×5=40)

4. **If 50 image of your website, 10 have black and white image, and their average scanned image occupies with 2.5 megabytes of memory. The total image occupies by the entire work 281 megabytes. Find the average occupies megabytes of those colour images.**

Ans: Total number of image (n) = 50

Number of black and white image (n_1) = 10

Average of black and white image ($\bar{x}_1$) = 2.5

Total image ($\sum x$) = 281

Average of color image ($\bar{x}_2$) = ?

Now average of all image $(\bar{x}_{12}) = \frac{\sum x}{n} = \frac{281}{50} = 5.62$

Number of color image (n_2) = $n - n_1 = 50 - 10 = 40$

Now,

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$\text{or, } 5.62 = \frac{10 \times 2.5 + 40 \times \bar{x}_2}{50}$$

$$\text{or, } 5.62 \times 50 = 25 + 40\,\bar{x}_2$$

$$\text{or, } 281 - 25 = 40\, n_1\bar{x}_2$$

$$\text{or, } \bar{x}_2 = \frac{(281 - 25)}{40} = 6.4$$

Hence, on average color image occupies 6.4 megabytes memory.

5. Calculate Q_1, D_2 and P_{58} from the following data and interpret the results.

Weight	0-10	10-15	15-25	25-30	30-35	35-40	40-45	45-50	50-60
No. of person	4	8	30	15	13	6	4	4	1

Ans:

Weight	No. of person (f)	cf
0 – 10	4	4
10 – 15	8	12
15 – 25	30	42
25 – 30	15	57
30 - 35	13	70
35 – 40	6	76
40 – 45	4	80
45 – 50	4	84
50 – 60	1	85
	$N = \sum f = 85$	

To find Q_1

$$\frac{N}{4} = \frac{85}{4} = 21.25$$

Cf just greater than 21.25 is 42 for which corresponding class is 15 – 25. Hence Q_1 class is 15 – 25

Here, L = 15, f = 30, cf = 12, h = 10

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times h$$

$$= 15 + \frac{21.25 - 12}{30} \times 10 = 18.083$$

Hence 18.083 divides 25% data in left side and 75% in right side of the data arranged in ascending order

To find D_2

$$\frac{2N}{10} = \frac{2 \times 85}{10} = 17$$

Cf just greater than 17 is 42 for which corresponding class is 15 – 25. Hence D_2 class is 15 – 25

Here, L = 15, f = 30, cf = 12, h = 10

$$D_2 = L + \frac{\frac{2N}{10} - cf}{f} \times h$$

$$= 15 + \frac{17 - 12}{30} \times 10 = 16.666$$

Hence 16.666 divides 20% data in left side and 80% in right side of the data arranged in ascending order

To find P_{58}

$$\frac{58N}{100} = \frac{58 \times 85}{100} = 49.3$$

Cf just greater than 49.3 is 57 for which corresponding class is 25 – 30.

Hence P_{58} class is 25 – 30

Here, L = 25, f = 15, cf = 42, h = 5

$$P_{58} = L + \frac{\frac{58N}{100} - cf}{f} \times h$$

$$= 25 + \frac{49.3 - 42}{15} \times 5 = 27.433$$

Hence 27.433 divides 58% data in left side and 42% in right side of the data arranged in ascending order

6. **The following joint probability data apply to fatigue test to be run on bronze strips. X represent to failure (in 10^5) when alternate strips are bent at a high level of deflection. Y represent the same at a lower deflection level.**

X \ Y	20	30	40	50
4	0.01	0.03	0.05	0.02
5	0.03	0.1	0.08	0.04
6	0.02	0.08	0.12	0.11
7	0.02	0.04	0.07	0.18

(a) **Find the marginal probability distribution for X and for Y.**

(b) **Determine the conditional probability distribution of Y given X = 5.**

(c) **Are X and Y independent?**

Ans:

X \ Y	20	30	40	50	P(X)
4	0.01	0.03	0.05	0.02	0.11
5	0.03	0.1	0.08	0.04	0.25
6	0.02	0.08	0.12	0.11	0.33
7	0.02	0.04	0.07	0.18	0.31
P(Y)	0.08	0.25	0.32	0.35	1

Marginal probability distribution for x = 4

$P(x = 4) = P(x = 4, y = 20) + P(X = 4, y = 30) + P(X = 4, y = 40) + P(X = 4, y = 50)$

$= 0.01 + 0.03 + 0.05 + 0.02 = 0.11$

Marginal probability distribution for y= 50

$P(y = 50) = P(x = 4, y = 50) + P(x = 5, y = 90) + P(x = 6, y = 50) + P(x = 7, y = 50) = 0.02 + 0.04 + 0.11 + 0.18 = 0.35$

Conditional probability distribution of y given x = 5

$$P\left(\frac{y}{X = 5}\right) = \left(\frac{P(x = 5, y)}{P(x = 5)}\right)$$

$$P\left(\frac{y = 20}{x = 5}\right) = \left(\frac{P(x = 5, y = 20)}{P(x = 5)}\right) = \frac{0.03}{0.25} = 0.12$$

$$P\left(\frac{y = 30}{x = 5}\right) = P\left(\frac{P(x = 5, y = 30)}{P(x = 5)}\right) = \frac{0.1}{0.25} = 0.4$$

$$P\left(\frac{y = 40}{x = 5}\right) = \left(\frac{P(x = 5, y = 40)}{P(x = 5)}\right) = \frac{0.08}{0.25} = 0.32$$

$$P\left(\frac{y = 40}{x = 5}\right) = \left(\frac{P(x = 5, y = 50)}{P(x = 5)}\right) = \frac{0.04}{0.25} = 0.16$$

For independence of x and y

$P(x,y) = P(x)\,P(y)$

Here $P(x = 4, y = 20) = 0.01$

$P(x = 4) = 0.11$

$P(y = 20) = 0.08$

$P(x = 4)\,P(y = 20) = 0.11 * 0.08 = 0.0088$

$P(x = 4, y = 20) = 0.01 \neq P(x = 4)\,P(y = 20) = 0.0088$

Hence X and Y are not independent.

7. Fit a binomial distribution to the following data:

X	0	1	2	3	4	5	6
f	5	8	15	14	10	6	2

Ans:

x	f	fx	$P(x) = C(n, x)\,p^x q^{n-x}$
0	5	0	$C(6,0)\,(0.45)^0\,(0.55)^5 = 0.027$
1	8	8	$C(6,1)\,(0.45)^1\,(0.55)^5 = 0.135$
2	15	30	$C(6,2)\,(0.45)^2\,(0.55)^4 = 0.277$
3	14	42	$C(6,3)\,(0.45)^3\,(0.55)^3 = 0.303$
4	10	40	$C(6,4)\,(0.45)^4\,(0.55)^2 = 0.186$
5	6	30	$C(6,5)\,(0.45)5\,(0.55)^1 = 0.06$
6	2	12	$C(6,6)\,(0.45)^6\,(0.55)^0 = 0.008$
	$N = \sum f = 60$	$\sum fx = 162$	

Now, $\bar{x} = \frac{\sum fx}{N} = \frac{162}{60} = 2.7$

Here n = 6

In binomial distribution

$\bar{x} = np$

or, $2.7 = 6p$

or, $p = 0.45$

$q = 1 - p = 1 - 0.45 = 0.55$

$P(x = 0) = C(6,0)\ (0.45)^0\ (0.55)^{6-0} = 0.027$

8. **If two random variables have the joint probability density function**

$$f(x, y) = \begin{cases} K(2x + 3y), \text{ for } 0 \le x \le 1, 0 \le y \le 1 \\ 0, \text{ otherwise} \end{cases}$$

Find (i) constant k (ii) Conditional probability density function of X given Y(iii) identify whether X and Y are independent.

Ans:

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

or $\int_0^1\int_0^1 k(2x + 3y)\, dx\, dy = 1$

or $k\int_0^1\int_0^1 (2x + 3y)\, dy\, dx = 1$

or $k\int_0^1\int_0^1 (2xdy + 3ydy)\, dx = 1$

or, $k\int_0^1 \left[2xy + \frac{3y^2}{2}\right]_0^1 dx = 1$

or, $\int_0^1 \left(2x + \frac{3}{2}\right) dx = 1$

or, $k\left[2\frac{x^2}{2} + \frac{3}{2}x\right]_0^1 = 1$

or, $k\left(\frac{1+3}{2}\right) = 1$

or, $k\frac{5}{2} = 1$

or, $k = \frac{2}{5}$

Hence, $f(x,y) = \frac{2}{5}(2x+3y)$

Now, $f(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$

$= \int_0^1 2/5.(2x + 3y)\, dy = \frac{2}{5}\left[2xy + \frac{3y^2}{2}\right]_0^1$

$= \frac{2}{5}\left(\frac{2x + 3}{2}\right) = \frac{2}{5}\frac{(4x + 3)}{2} = \frac{(4x + 3)}{5}$

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

$$= \int_0^1 2/5(2x + 3y)\, dx = \frac{2}{5}\left[2\frac{x^2}{2} + 3xy\right]_0^1 = \frac{2}{5}(1 + 3y)$$

$$f(x/y) = \frac{f(x, y)}{f(y)} = \frac{\frac{2}{5}(2x + 3y)}{\frac{2}{5}(1 + 3y)} = \frac{2x + 3y}{1 + 3y}$$

To test independence

$$f(x)\, f(y) = \frac{(4x + 3)}{5} \times \frac{2(1 + 3y)}{5} = \frac{2(4x + 12xy + 9y + 3)}{25} \neq f(x, y) = \frac{2}{5}(2x + 3y)$$

Hence x and y are not independent.

9. Compute first four moments about arbitrary point 4 fro following distribution and describe the characteristics of data.

X	2	3	4	5	6
f	1	3	7	2	1

Ans:

x	f	d=(x-4)	fd	fd^2	fd^3	fd^4
2	1	-2	-2	4	-8	16
3	3	-1	-3	3	-3	3
4	7	0	0	0	0	0
5	2	1	2	2	2	2
6	1	2	2	4	8	16
	$N = \sum f = 14$		$\sum fd = -1$	$\sum fd^2 = 13$	$\sum fd^3 = -1$	$\sum fd^4 = 37$

First four moments about arbitrary point 4 are

$$\mu_1' = \frac{\sum fd}{N} = -\frac{1}{14} = -0.071$$

$$\mu_2' = \frac{\sum fd^2}{N} = \frac{13}{14} = 0.928$$

$$\mu_3' = \frac{\sum fd^3}{N} = -\frac{1}{14} = 0.071$$

$$\mu_4' = \frac{\sum fd^4}{N} = \frac{37}{14} = 2.642$$

Now to describe characteristics of the distribution

$$\mu_2 = \mu_2' - \mu_1'^2 = 0.928 - (-0.071)^2 = 0.922$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -0.071 - 3 \times 0.928 \times (-0.071) + 2 \times (-0.071)^3$$

$$= 0.125$$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$

$= 2.642 - 4 \times (-0.071) \times -0.071) + 6 \times 0.928 \times (-0.071)^2 - 3 \times (-0.071)^4$

$= 10.949 = 2.649$

Again,

Measure of central tendency, $\bar{x} = A + \mu_1' = 4 + (-0.071) = 3.929$

Measure of dispersion, $\sigma = \sqrt{\mu_2} = \sqrt{0.922} = 0.96$

Measure of skewness, $\gamma_1 = \frac{\mu_3}{\frac{3}{2}} = \frac{0.125}{(0.922)^{3/2}} = 0.141$

Measure of kurtosis, $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{2.649}{(0.922)^2} = 3.11$

Hence average value of data is 3.929 with standard deviation 0.96. Data is positively skewed with leptokurtic distribution.

10. **The lifetime of a certain electronic component is a normal random variate with the expectation of 5000 hours and a standard deviation of 100 hours. Compute the probabilities under the following conditions:**

a) **Lifetime of components is less than 4000 hours**

b) **Lifetime of components between 3000 to 6500 hours**

c) **Lifetime of components more than 6000 hours**

Ans: Let X = life time of certain electronic component

$X \sim N(\mu, \sigma^2)$

Expectation (μ) = 5000 hrs

Standard deviation (σ) = 100

Define $Z = \frac{X - \mu}{\sigma} = \frac{x - 50000}{100}$

(a) $P(X < 4000) = ?$

When $X = 4000$, $Z = \frac{4000 - 5000}{100} = -10$

$P(X < 4000) = P(Z < -10) = 0$

(b) $P(3000 < X < 6500) = ?$

When $X = 3000$, $Z = \frac{3000 - 5000}{100} = -20$

When $X = 6500$, $Z = \frac{6500 - 5000}{100} = 15$

$P(3000 < X < 6500) = P(-20 < Z < 15) = P(-20 < Z < 0) + P(0 < Z < 15)$

$= 0.5 + 0.5 = 1$

(c) $P(X > 6000) = ?$

When $X = 6000$, $Z = \frac{6000 - 5000}{100} = 10$

$P(X > 6000) = P(Z > 10) = 0$

11. Calculate Sperman's rank correlation coefficient for the following ran given by three judges in a music contest.

1st Judge	2	1	4	6	5	8	9	10	7	3
2nd Judge	4	3	2	5	1	6	8	9	10	7
3rd Judge	5	8	4	7	10	2	1	6	9	3

Indicate which pair of judges has the nearest approach to music.

Ans:

Rank by 1st judge (R_1)	Rank by 2nd judge (R_2)	Rank by 3rd judge (R_3)	$d_{12} = R_1 - R_2$	d_{12}^2	$d_{13} = R_1 - R_3$	d_{13}^2	$d_{23} = R_2 - R_3$	d_{23}^2
2	4	5	-2	4	-3	9	-1	1
1	3	8	-2	4	-7	49	-5	25
4	2	4	2	4	0	0	-2	4
6	5	7	1	1	-1	1	-2	4
5	1	10	4	16	-5	25	-9	81
8	6	2	2	4	6	36	4	16
9	8	1	1	1	8	64	7	49
10	9	6	1	1	4	16	3	9
7	10	9	-3	9	-2	4	1	1
3	7	3	-4	16	0	0	4	16
			$\sum d_{12}=0$	$\sum d_{12}^2=60$	$\sum d_{13}=0$	$\sum d_{13}^2=204$	$\sum d_{23}=0$	$\sum d_{23}^2=206$

Now, Spearman's rank correlation coefficient between judge 1 and 2

$$R_{(1,2)} = 1 - \frac{6\sum d_{12}^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{10(100-1)} = 0.636$$

Spearman's rank correlation coefficient between judg 1 and 3

$$R_{(1,3)} = 1 - \frac{6\sum d_{13}^2}{n(n^2-1)} = 1 - \frac{6 \times 204}{10(100-1)} = -0.236$$

Spearman's rank correlation coefficient between judge 2 and 3

$$R_{(2,3)} = 1 - \frac{6\sum d_{23}^2}{n(n^2-1)} = 1 - \frac{6 \times 206}{10(100-1)} = -0.248$$

Here, $R_{(1,2)} = 0.636$ is positive correlation, hence judge 1 and 2 have near approach to music

12. What do you mean by sampling? Explain the differences betwe stratified sampling and cluster sampling.

Ans: When one by one study of all units of a population is not possible due some factors like time, cost, manpower, recourses and destructive nature study, we take a small representative part from the population for the stu This small representative part selected for the study from the population called sample.

The process of selecting a sample from a population is called *sampling.* I example,

- A housewife takes only two or three grains of rice as a sample from t cooking pan to know whether the rice is property cooked or not.
- A pathologist takes a syringe of blood as a sample to find out a disea

Difference between stratified sampling and cluster sampling

Startified sampling	Cluster sampling
1. Population is divided into different groups called strata such that population within strata is homogeneous and between strata is heterogeneous then sample is selected from each strata proportional to it's size using simple random sampling	1. Population is divided into different groups called cluster such that population within cluster is heterogeneous and between cluster is homogeneous then one or more cluster is selected at random using simple random sampling. Elements in selected cluster are taken as sample
2. It is homogeneous within group	2. It is homogeneous between group
3. It is heterogeneous between group	3. It is heterogeneous within group
4. Sample are selected individually from group	4. Sample are selected collectively from group
5. It increase precision and representation	5. It reduce cost and improve efficiency
6. Groups are formed by expert or researcher	6. Groups are formed naturally

14. State with suitable examples the role played by computer technology in applied statistics and the role of statistics in information technology.

Ans: Computer technology plays important role in applied statistics. Computer is used for numerical and graphical data analysis, symbolic computations, simulations, storing statistical knowledge, presentation of results. Statistical package such as EXCEL, SPSS, STATA, R, Pandas, SAS, Prism, Matlab, Minitab, Epi Info, Statistica,Systat, SigmaStat, GenStat, StatXactetc are used for the statistical analysis. Statistical package make statistical work easy and faster.

Statistics is used for various purpose in computer science. Statistics used for data mining, data compression, speech recognition, vision and image analysis, artificial intelligence and network, traffic modeling, quality management, software engineering, storage and retrieval processes, hardware engineering and manufacturing etc. Data mining is performed with help of statistics to find irregularities or inconsistencies in data. Data compression uses statistical algorithms to compress data. In speech recognition statistical model learn the pattern in audio that make sounds of speech. The models are used to automatically transcribe new speech. In vision and image analysis statistical learning techniques are used to recognize faces. In artificial and intelligence network statistics combines logical and probabilistic relation. In traffic modeling statistical techniques such as stochastic process and queuing theory are used to estimate and predict of flows in traffic network. Statistics provide solution of problem related to quality management such as quality planning, quality assurance, quality control and quality improvement. Statistical methods are used for controlling and improving the quality and productivity of practices used on creating software. In hardware engineering and manufacturing statistical tools such as quality control and process control are used to manage conformance to indicated specifications.