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.

MODEL QUESTIONS-ANWERS

Group 'A

Attempt any Two questions **(2×10=20)**

1. A new computer program consists of two modules. The first module contains an error with probability 0.2. The second module is more complex; it has a probability of 0.4 to contain an error, independently of the first module. An error in the first module alone causes the program to crash with probability 0.5. For the second module, this probability is 0.8. If there are errors in both modules, the program crashes with probability 0.9. Suppose the program crashed. What is the probability of errors in both modules?

Solution:

Let A_1 = First module only
A_2 = Second module only
A_3 = Both module
A_4 = None of module
Let B = program crashes
Here, $P(A_3) = 0.2 \times 0.4 = 0.08$
$P(A_1) = 0.2 - 0.08 = 0.12$
$P(A_2) = 0.4 - 0.08 = 0.32$
$P(A_4) = 1 - 0.08 - 0.12 - 0.32 = 0.48$
Also,
$P(B/A_1) = 0.5$
$P(B/A_2) = 0.8$
$P(B/A_3) = 0.9$
$P(B/A_4) = 0$
$P(A_3/B) = ?$
Now,

$$P(A_3/B) = \frac{(A_3)\, P(B/A_3)}{\sum_{i=1}^{4} P(A_i)\, P(B/A_i)}$$

$$= \frac{0.08 \times 0.9}{0.12 \times 0.5 + 0.32 \times 0.8 + 0.08 \times 0.9 + 0.48 \times 0} = \frac{0.072}{0.388}$$

$$= 0.185$$

Hence when program is crashed probability of errors in both modules is 0.185.

2. **Explain how box-plot is helpful to know the shape of the data distribution. The following data set represents the number of new computer accounts registered during ten consecutive days.**

43 37 50 51 58 105 52 45 45 10

a) **Compute the mean, median, quartiles, and sample standard deviation.**

b) **Check whether there are outliers or not.**

c) **If outliers are present, then delete the detected outliers and compute the mean, median, quartiles, and sample standard deviation again.**

d) **Make your conclusion about the effect of outliers on descriptive statistical analysis.**

Solution:

Box plot gives information about location, dispersion and shape of the distribution

Shape is positively skewed if

i. $Q_3 - Md > Md - Q_1$

ii. $X_{max} - Md > Md - X_{min}$

iii. $X_{max} - Q_3 > Q_1 - X_{min}$

Shape is negatively skewed if

i. $Q_3 - Md < Md - Q_1$

ii. $X_{max} - Md < Md - X_{min}$

iii. $X_{max} - Q_3 < Q_1 - X_{min}$

Shape is symmetrical if

i. $Q_3 - Md = Md - Q_1$

ii. $X_{max} - Md = Md - X_{min}$

iii. $X_{max} - Q_3 = Q_1 - X_{min}$

Number of computer accounts registered(x)	$2x^2$
10	100
37	1369
43	1849
45	2025
45	2025
50	2500
51	2601
52	2704
58	3364
105	11025
$\Sigma x = 496$	$\Sigma x^2 = 29562$

Here,

$\text{Mean} = \bar{x} = \frac{\Sigma x}{n} = \frac{496}{10} = 49.6$

$\text{Median} = Md = \left(\frac{n+1}{2}\right)^{th} \text{item } \frac{10+1}{2} = 5.5^{th} \text{ item} = \frac{45+50}{2} = 47.5$

$\text{First quartile} = Q_1 \left(\frac{n+1}{4}\right)^{th} \text{item} = 2.75^{th} \text{ item} = \text{second item} + (\text{third item} - \text{second item}) \times 0.75$

$= 37 + (43 - 37) \times 0.75 = 41.5$

$\text{Third quartile} = Q_3 = 3\left(\frac{n+1}{4}\right)^{th} \text{item} = \frac{3(10+1)}{4} = 3 \times 2.75^{th} \text{ item}$

$= 8.25^{th}$ item

$= \text{eighth item} + (9^{th} \text{ item} - \text{eighth item}) \times 0.25$

$= 52 + (58 - 52) \times 0.25 = 53.5$

$\text{Sample sd} = s = \sqrt{\frac{1}{n-1}\Sigma(x-\bar{x})^2}$

$= \sqrt{\frac{1}{n-1}\{\Sigma(x-\bar{x})^2\}} = \sqrt{\frac{1}{10-1}\{29562 - 10 \times (49.6)^2\}} = 23.476$

Hence,

mean = 49.6, median = 47.5, First quartile = 41.5, Third quartile = 53.5, sample sd = 23.47

Outliers = values outside 1.5 times inter-quartile range below lower quartile and above upper quartile

Inter quartile range (IR) = $Q_3 - Q_1 = 53.5 - 41.5 = 12$

$Q_1 - 1.5$ IR $= 41.5 - 1.5 \times 12 = 23.5$

$Q_3 - 1.5$ IR $= 53.5 + 1.5 \times 12 = 71.5$

Here 10 is below 23.5 and 105 is above 71.5

Hence 10 and 105 are outliers

After omitting outliers

Number of computer accounts registered(x)	x^2
37	1369
43	1849
45	2025
45	2025
50	2500
51	2601
52	2704
58	3364
$\Sigma x = 381$	$\Sigma x^2 = 18437$

$$\text{Mean} = \bar{x} = \frac{\Sigma x}{n} = \frac{381}{8} = 47.625$$

$$\text{Median} = Md = \left(\frac{n+1}{2}\right)^{th} \text{item} = \frac{8+1}{2} = 4.5^{th} \text{ item} = \frac{45+50}{2} = 47.5$$

$$\text{First quartile} = Q_1 = \left(\frac{n+1}{4}\right)^{th} \text{item} = \frac{8+1}{4} = 2.25^{th} \text{ item}$$

$$= \text{Second item} + (\text{third item} - \text{second item}) \times 0.25$$

$$= 43 + (45 - 43) \times 0.25 = 43.5$$

$$\text{Third quartile} = Q_3 = 3\left(\frac{n+1}{4}\right)^{th} \text{item} = \frac{3(8+1)}{4} = 3 \times 2.25^{th} \text{ item}$$

$$= 6.75^{th} \text{ item}$$

$$= \text{Sixth item} + (\text{seventh item} - \text{sixth item}) \times 0.75$$

$$= 51 + (52 - 51) \times 0.75 = 51.75$$

$$\text{Sample sd} = s = \sqrt{\frac{1}{n-1}\Sigma(x-\bar{x})^2} = \sqrt{\frac{1}{n-1}\{\Sigma x^2 - n\bar{x}^2\}}$$

$$= \sqrt{\frac{1}{8-1}\{18437 - 8 \times (47.625)^2\}} = 23.476 = 6.45$$

After deleting the detected outliers

Mean = 47.625, Median = 47.5, First quartile = 43.5, Third quartile = 51.75, sample sd = 6.45

Here, Mean with outliers = 49.6 > Mean without outliers = 47.624

Median with outliers = 47.5 = Median without outliers = 47.5

First quartile with outliers = 41.5 < First quartile without outliers = 43.5

Third quartile with outliers = 53.5>Third quartile without outliers = 51.75

Sample sd with outliers = 23.476 > Sample sd without outliers = 6.45

Hence there is change in descriptive statistics except median with and without outliers.

3. **A computer manager interested to know how efficiency of his/her new computer program which depends on the size of incoming data. Efficiency will be measured by the number of processed requests per hour. In general, larger data sets require more computer time, and therefore, fewer requests are processed within 1 hour. Applying the program to data sets of different sizes, the following data were gathered.**

Data size (gigabytes)	6	7	7	8	10	10	15
Processed requests	40	55	50	41	17	26	16

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

b) **Interpret the regression coefficient with reference to your problem.**

c) **Obtain coefficient of determination, and interpret this.**

d) **Based on the fitted model in (a), predict the efficiency of new computer for data size 12 (gigabytes). Does it possible to predict efficiency for data size of 30 gigabytes? Discuss.**

Solution:

Here efficiency (Number of processes requested) is response variable (dependent variable)

To fit liner regression between processed requested and data size

Data size (x)	Processed requested (y)	xy	x^2	y^2
6	40	240	36	1600
7	55	385	49	3025
7	50	350	49	2500
8	41	328	64	1681
10	17	170	100	289
10	26	260	100	676
15	16	240	225	256
$\Sigma x = 63$	**$\Sigma y = 245$**	**$\Sigma xy = 1973$**	**$\Sigma x^2 = 623$**	**$\Sigma y^2 = 10027$**

To fit, $y = a+bx$

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

Or, $245 = 7a + 63b$ ....(i)

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

$1973 = 63a + 623b$ ... (ii)

Solving (i) and (ii)

Coeff of a	Coeff of b	Constant
7	63	245
63	623	1973

$$D = \begin{vmatrix} 7 & 63 \\ 63 & 623 \end{vmatrix} = 7 \times 623 - 63 \times 63 = 392$$

$$D_1 = \begin{vmatrix} 245 & 63 \\ 1973 & 623 \end{vmatrix} = 245 \times 623 - 1973 \times 63 = 28336$$

$$D_2 = \begin{vmatrix} 7 & 245 \\ 63 & 1973 \end{vmatrix} = 7 \times 1973 - 63 \times 245 = -1624$$

Now,

$$\hat{a} = \frac{D_1}{D} = \frac{28336}{392} = 72.285$$

$$b = \frac{D_2}{D} = -\frac{1624}{392} = -4.142$$

Hence regression equation is $y = a + bx$

or, $y = 72.285 - 4.142x$

Here regression coefficient is -4.142 it means process requested decreases by 4.142 for unit increase in data size.

TSS (Total sum of square) $= \Sigma (y - \bar{y})^2 = \Sigma y^2 - n\bar{y}^2$

$$= 10027 - 7 \times \left(\frac{245}{7}\right)^2 = 10027 - 8575$$

$$= 1452$$

SSE (Sum of square due to error) $= \Sigma (y - \hat{y})^2 = \Sigma y^2 - a\Sigma y - b\Sigma xy$
$= 10027 - 72.285 \times 245 - (-4.142) \times 19723 = 489.341$
SSR (Sum of square due to regression) $= TSS - SSE$
$= 1452 - 489.341 = 962.659$

Coefficient of determination $= R^2 = \frac{SSR}{TSS} = \frac{962.659}{1452} = 0.6629 = 66.29\%$

It means 66.29% variation in process requested is explained by data size
When $x = 12$, $y = 72.285 - 4.142\ x = 72.285 - 4.142 \times 12 = 22.58$
When $x = 30$, $y = 72.285 - 4.142\ x = 72.285 - 4.142 \times 30 = -51.97$
It is not possible because process requested is not negative.

Group B

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

4. Explain the role of statistics in computer science and information technology.

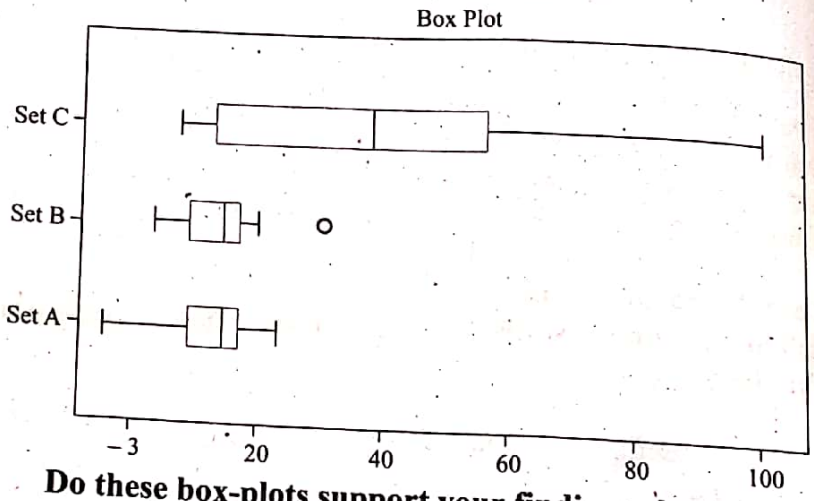
Solution:

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.

5. Following table presents some descriptive statistics computed from three different independent sample dataset (X).

Data	Sample size (n)	$\sum_{i=1}^{30} X_i$	Minimum	Q_1	Median	Q_3	Maximum	$\sum_{i=1}^{30} (X_i - \overline{X})^2$
Data set A	30	439	1	10	15	20	26	1348.97
Data set B	30	625	11	18	21	23	33	540.17
Data set C	30	1239	13	24	39.5	53	94	12836.3

a) **Compare sample mean and median, and explain about the shape of the data dist for each dataset. Compare the variability of the three set of dataset. Box-plots have been generated through SPSS for each dataset as follows.**



b) **Do these box-plots support your findings obtained in a) about the shape of the distribution? Explain.**

Solution:

For data set A

$$\bar{x} = \sum \frac{x_i}{n} = \frac{439}{30} = 14.63$$

$Md = 15$

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n} = 1348.97/30 = 44.965$$

Hence $\sigma = \sqrt{44.965} = 6.7$

$$CV = \frac{\sigma}{\bar{x}} \times 100\% = \frac{6.7}{14.63} \times 100\% = 45.79\%$$

Here mean = 14.63 < Median = 15, hence negatively skewed

For data set B

$$\bar{x} = \sum \frac{x_i}{n} = \frac{625}{30} = 20.83$$

$Md = 21$

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{540.17}{30} = 18.005$$

Hence $\sigma = \sqrt{18.005} = 4.24$

$$CV = \frac{\sigma}{\bar{x}} \times 100\% = \frac{4.24}{20.83} \times 100\% = 20.37\%$$

Here, mean = 20.83 < Median = 21, hence negatively skewed

For data set C

$$\bar{x} = \sum \frac{x_i}{n} = \frac{1239}{30} = 41.3$$

Md = 39.5

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{12836.3}{30} = 427.876$$

Hence $\sigma = \sqrt{427.876} = 20.685$

$$CV = \frac{\sigma}{\bar{x}} \times 100 = \frac{20.685}{41.3} \times 100\% = 50.08$$

Here, mean = 41.3 > Median = 39.5, hence positively skewed
Also, CV of data set A = 45.79% > CV of data set B = 20.37% < CV of data set C = 50.08%
Minimum variability in data set B and maximum variability in data set C
From box plot of data set A, it is negatively skewed.
From box plot of data set B, it is negatively skewed
From box plot of data set C, it is positively skewed.
Hence box plot support finding about shape of distribution with comparison of mean and median.

6. **A large chain retailer purchases a certain kind of electronic device from a manufacturer. The manufacturer indicates that the defective rate of the device is 3%.**
   a) **The inspector randomly picks 20 items from a shipment. What is the probability that there will be at least one defective item among these 20?**
   b) **Suppose that the retailer receives 10 shipments in a month and the inspector randomly tests 20 devices per shipment. What is the probability that there will be exactly 3 shipments each containing at least one defective device among the 20 that are selected and tested from the shipment?**

Solution:

Let X = number of defective items
Probability of defective item (p) = 3% = 0.03
q = 1 – p = 1 – 0.03 = 0.97
Number of items selected (n) = 20
Probability of at least one defective $P(X \geq 1) = ?$
$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - c(20, 0)\ (0.03)^0\ (0.97)^{20-0}$
= 1 – 0.543 = 0.457
Number of shipment (n) = 10
Let Y = number of shipment containing at least one defective
Probability that shipment containing at least on defective in 20 (p) = 0.457
q = 1-p = 1 – 0.457 = 0.543
Probability of 3 shipment containing at least one defective P(y = 3) = ?
$P(y=3) = c(10, 3)\ (0.457)^3\ (0.543)^{10-3} = 120 \times 0.0954 \times 0.0139 = 0.159$

7. **Messages arrive at an electronic message center at random times, with an average of 9 messages per hour.**
   a) **What is the probability of receiving at least five messages during the next hour?**
   b) **What is the probability of receiving exactly seven messages during the next hour?**

Solution:

Let y = number of messages

Average number of messages (λ) = 9 per hour

Probability of receiving at least five messages during next hour $P(y \geq 5)$

$= 1 - P(y < 5)$

$= 1 - \{P(y=0) + P(y=1) + P(y=2) + P(y=3) + P(y=4)\}$

$= 1 - \left\{\frac{e^{-9}9^0}{0!} + \frac{e^{-9}9^1}{1!} + \frac{e^{-9}9^2}{2!} + \frac{e^{-9}9^3}{3!} + \frac{e^{-9}9^4}{4!}\right\}$

$= 1 - e^{-9}\{1 + 9 + 40.5 + 121.5 + 273.375\}$

$= 1 - 445.375 \times e^{-9} = 1 - 0.054 = 0.946$

Probability of receiving exactly seven messages during next hour $P(y = 7)$

$= \frac{e^{-9}9^7}{7!} = 0.117$

8. The time, in minutes, it takes to reboot a certain system is a continuous variable with the density function:

$$f(x) = \begin{cases} C(10-x)^2, & 0 < x < 10 \\ 0, & \text{otherwise} \end{cases}$$

Compute C, and then compute the probability that it takes between 1 and 2 minutes to reboot.

Solution:

To find C

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

or, $\int_0^{10} C(10-x)^2\,dx = 1$

or, $C\int_0^{10} (100 - 20x + x^2) = 1$

or, $C\left[100x - \frac{20x^2}{2} + \frac{x^3}{3}\right]_0^{10} = 1$

or, $C\left[100 \times 10 - 10 \times (10)^2 + \frac{10^3}{3}\right] = 1$

or, $C\left[1000 - 1000 + \frac{1000}{3}\right] = 1$

or, $C = \frac{3}{1000}$

Hence $f(x) = \frac{3}{1000}(10-x)^2$

Probability that it takes between 1 and 2 minutes to reboot $P(1 \leq x \leq 2)$

$= \int_1^2 f(x)\,dx$

$= \int_0^{10} \frac{3}{1000}(10-x)^2\,dx = \frac{3}{1000}\int_1^2 (100 - 2x + x^2)\,dx$

$= \frac{3}{1000}\left[100x - \frac{20x^2}{2} + \frac{x^3}{2}\right]_1^2$

$$= \frac{3}{1000}\left[100 \times 2 - 10\,(2)^2 + \frac{2^3}{3} - 100 \times 1 + 10 \times 1^2 - \frac{1^3}{3}\right]$$

$$= \frac{3}{1000}\left[200 - 40 + \frac{8}{3} - 100 + 10 - \frac{1}{3}\right] = \frac{3}{1000}\,[70 + 7/3]$$

$$= 0.217$$

9. **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.

Solution:

Preference of student on computer DELL (R_1)	Preference of student on computer HP (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
		$\Sigma d = 0$	$\Sigma d^2 = 216$

Here, $n = 10$

$\Sigma d^2 = 216$

Correlation of brand preference (R) $= 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 216}{10 \times 99}$

$= 1 - 1.309 = -0.309$

10. **Define exponential distribution with parameter λ. The time required to reach to the printer after ordering in the computer follows exponential distribution at an average rate of 3 jobs per hour.**
    a) **What is the expected time between jobs?**
    b) **What is the probability that the next job is sent within 5 minutes?**

Solution:

A continuous random variable X assuming non negative values is said to follow an exponential distribution with parameter $\lambda > 0$, if its probability density function is given by

$f(x) = \lambda e^{-\lambda x}$, $x \geq 0$

0, otherwise

A random variable following exponential distribution with parameter λ is denoted by $X \sim \exp(\lambda)$

Let X = time of job

Average rate of job (λ) = 3 /hr

E(X) = $1/\lambda = 1/3$ hours = 20 minutes between jobs

Time (t) = 5 min = (1/12) hrs

$$P\left\{X \leq \frac{1}{12}\text{hrs}\right\} = \int_0^{1/12} 3e^{-3x}\,dx = \left[\frac{3e^{-3x}}{-3}\right]_0^{1/12} = \left[-e^{-3x}\right]_0^{1/12}$$

$$= -e^{-3/12} + 1 = 1 - e^{-1/4} = 0.22$$

11. 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 5012 hours

b) Lifetime of components between 4000 to 6000 hours

c) Lifetime of components more than 7000 hours

Solution:

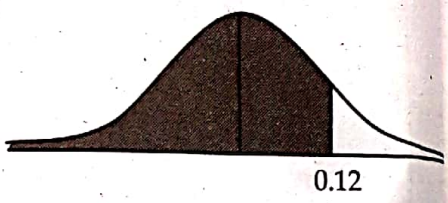
Let X = life time of certain electronic component

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

Expectation (μ) = 5000 hrs

Standard deviation (σ) = 100

Define $Z = \frac{X - \mu}{\sigma} = \frac{x - 5000}{100}$



$P(X < 5012) = ?$

When X = 5012, $Z = \frac{5012 - 5000}{100} = 0.12$

$P(X < 5012) = P(Z < 0.12)$

$= 0.5 + P(0 < Z < 0.12)$

$= 0.5 + 0.0478 = 0.5478$

$P(4000 < X < 6000) = ?$

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

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

$P(4000 < X < 6000) = P(-10 < Z < 10) = P(-10 < Z < 0) + P(0 < Z < 10)$

$= P(0 < Z < 10) + P(0 < Z < 10)$

$= 2P(0 < Z < 10) = 2 \times 0.5 = 1$

$P(X > 7000) = ?$

When X = 7000, $Z = \frac{7000 - 5000}{100} = 20$

$P(X > 7000) = P(Z > 20) = 0$

12. Write short notes on the following.

a) Sampling error and non-sampling error

b) Conditional probability

Ans: (a) Difference between sampling error and non sampling error

Sampling error	Non sampling error
1. The difference between the values of the sample statistic obtained from a sample and the value of corresponding population parameter obtained from the population is called the sampling error.	1. Non sampling error is occurred in research from other sources than the sample, which can be minimize by checking the process, preparing the questionnaire properly, performing pilot survey, fixing procedure, by using competent manpower and experts etc.
2. It is random	2. It is random or non random
3. It occurs in sample.	3. It occurs in sample and census.
4. It decreases with increase in sample size	4. It has no effect on sample size.

Ans (b) Conditional probability

The probability of happening of one event given that other had already happened is called conditional probability.

Let A and B be any two events then probability of event A given that B had already happened is called conditional probability of A given B and is given by

$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$

Similarly probability of event B given that A had already happened is called conditional probability of B given A and is given by

$$P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0$$

Eg. A card is drawn from pack of 52 cards then probability of getting king given that card is spade is the conditional probability.

Tribhuvan University
Institute of Science and Technology

Bachelor Level/First Year/Second Semester/Science — Full Marks: 60
Computer Science and Information Technology STA 164 — Pass Marks: 24
(Statistics I) — 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.

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

$= 5273 - 5232.11 = 40.88$

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

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

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

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 $p_i = P(x_i) = P(X=x_i) = P(X=x_i \text{ and } Y = y_1) + P(X=x_i \text{ and } Y = y_2) + P(X=x_i \text{ and } Y = y_3) + \ldots + P(X=x_i \text{ and } Y = y_m)$

$= p_{i1} + p_{i2} + p_{i3} + \ldots p_{im} = \sum_{j=1}^{m} p_{ij} = p_{i\cdot}$ 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 $a \leq x \leq b$ and marginal probability of random variable y such that $c \leq y \leq d$ is given by

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

$P(c \le x \le d) = \Sigma_{j=c}^{d} P.j$ 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 = Σf = 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

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 $y = f(x/y) = \frac{f(x,y)}{f(y)} = \frac{e^{-(x+y)}}{e^{-y}} = e^{-x}$

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

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

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

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

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

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

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

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

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

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

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
	$N = \Sigma f = 55$	

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}

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

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.

12. 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

Model Questions Sets For Practice

Bachelor Level/First Year/Second Semester/Science **Full Marks: 60**
Computer Science and Information Technology STA 164 (Statistics I) **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.

MODEL SET 1

Group A

Attempt any Two questions **(2×10=20)**

1. What do you mean by conditional probability? What is advantage of Baye's theorem over conditional probability. Suppose that assembly plant receives it's voltage regulator from three different suppliers 60% from B_1, 30% from B_2 and 10% from B_3. If 90% of voltage regulators from B_1, 80% of voltage regulators from B_2 and 60% from B_3 perform according to specification. Compute probability that any one voltage regulator received by the plant perform according to specification. Also find probability that a voltage regulator which is known to perform according to the specification comes from plant B_3.
Ans: 0.071

2. What do you mean by correlation? For what purpose Karl Pearson's correlation coefficient is used in statistical analysis? State it's major properties.
It has been realized that production of coal in certain coal factory has been affected to certain extent by number of workers involved. The following table shows the production of coal and number of workers in a certain time during which capital equipment remained constant.

Output in tons(y)	21	21	20	18	17	17	14	13
Number of workers(x)	70	68	65	50	47	47	43	42

Using above data fit regression equation of y on x and predict y for x=60.
Ans: 19.038

3. Compute the measure of central tendency, dispersion and skewness from following distribution of wages (Rs per hour) of employees in a software company.

Wage (Rs)	*100-110*	*110-120*	*120-130*	*130-140*	*140-150*	*150-160*	*160-170*	*170-180*
No of employees	2	5	12	17	14	6	3	1

Ans: 136.83, 14.66, 0.039

Group B

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

4. The following are number of minutes that a person had wait for the bus to work in a local Bus park of Kathmandu on 15 working days; 10 , 1 , 13 , 9 , 5 , 9 , 2 , 10 , 3 , 8 , 6 , 17 , 4 , 10 and 15. Compute mean , median , mode and explain about shape of distribution.
Ans: 8.13, 9, 10

5. Compute appropriate measure of kurtosis for download speed of data given below:

Download speed(mbps)	Below 100	100- 150	150- 200	200- 250	250-300	300 & above	Total
Time in minutes	10	25	145	220	70	30	500

Ans: 0.2423

6. What do you mean by sampling? Describe different steps of sampling.

7. Define binomial probability distribution. If mean of binomial distribution is 3 and standard deviation is 2 explain whether information is correct or not for the binomial distribution.
Ans: Incorrect

8. Let (x,y) be two dimensional random variable having probability density function $f(x,y) = \frac{2}{3}(x + y); 0 < x < 1, 0 < y < 1$

   0 ; otherwise

   Find i) condition probability density function of x given y ii) conditional probability density function of y given x iii) E(x)
Ans: $\frac{2(x + y)}{2y + 1}, \frac{2(x + y)}{2x + 1}$, 0.38

9. What is Poisson variate? Service calls come to a service center according to poisson process on average of 2.7 calls per minute. Find probability of (i)no more than four calls come in one minute (ii) 5 calls in 2 minutes (iii) 1 call in 30 second.
Ans: 0.89, 0.17, 0.349

10. Given a random variable having normal distribution with mean 16.2 and variance 1.56 find probability that it will take value (i) greater than 16.8 (ii) less than 14.9 (iii) between 13.6 and 18.8
Ans: 0.31, 0.14, 0.96

11. What do you mean by expectation? An importer is offered a shipment of machine tools for Rs 140000 and probability that he will be able to sell them for Rs 180000, Rs 170000 and Rs 150000 are 0.32, 0.55 and 0.13 respectively . What is importer expected gross profit?
Ans: 1300

12. Following data represents marks secured by students of in exam of C programming and Digital Logic;

Student	A	B	C	D	E	F	G	H
Marks in C programming	32	51	44	35	44	38	45	49
Marks in Digital Logic	41	33	28	40	51	37	24	39

Apply Spearman's measure to determine association between marks on those two subjects.

Ans: -0.5

MODEL SET 2

Group A

Attempt any Two questions **(2×10=20)**

1. Define independent and mutually exclusive events. Can two events be mutually exclusive and independent simultaneously? Support your answer with an example.
A factory has three machines A,B and C producing large number of a certain item of the total daily production of the items 50% are produced on A, 30% on B and 20% on C. Record show that 2% of the items produced on A are defective, 3% of items produce on B are defective and 4% of items produced on C are defective. The occurrence of defective item is independent of all other items. One item is chosen at random from a day's total output, (i) show that the probability of it being defective item is 0.027 (ii) Given that is defective, find the probability that it was produced on machine A.
Ans: 0.37

2. Distinguish between correlation and regression. Also point out properties of regression coefficients. The following sample observations were randomly selected;

x	5	3	6	3	4	4	6	8
y	7	6	8	4	5	6	9	10

Determine coefficient of correlation and coefficient of determination. Interpret the association between X and Y. Find the regression equation of Y on X. **Ans:** 0.93, 0.86, y = 1.536 + 1.09X

3. Define the following three measures of dispersion; Range, Standard deviation and semi-inter-quartile range. Describe a situation under which semi interquartile range is preferred than standard deviation . Score obtained by 10 students in a test are given below. Compute range, and standard deviation.

42	55	35	60	55	55	65	40	45	35

Ans: 30, 10.159

Group B

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

4. What is meant by measurement scale? Describe various types of scales.

5. Determine five number summary and construct box plot from following data. Also describe shape of the distribution.

Marks	35	38	45	60	72	80	85
No. of students	3	5	10	8	4	2	1

Ans: Positive Skewed

6. A continuous random variable X has the probability density function given by

$f(x) = K(1+x^2)$; $-1 \leq x \leq 1$

0 ; elsewhere

Where k is a constant

(I) Find value of k (ii) Find $P(0.3 \leq x \leq 0.6)$ (iii) Find variance of x

Ans: $\frac{3}{8}$, 0.136, 0, 0.4

7. The marks of 500 candidates in an examination are normally distributed with mean of 45 marks and standard deviation of 20 marks;

(i) Given the pass mark is 40, estimate the number of candidate who passed the examination

(ii) If 5% of the candidate obtain a distinction by scoring x marks or more, estimate the value of x.

Ans: 299, 77.9

8. The joint density function of w and z is given by

$f(w,z) = bwz$; $1 \leq w \leq 2,\ 2 \leq z \leq 4$

0 ; otherwise

Find marginal density function of w and z. **Ans:** $\frac{2}{3}w, \frac{z}{6}$

9. Mr X recorded number of emails received over a period of 150 days with the following results

Number of emails	0	1	2	3	4
Number of days	51	54	36	6	3

i) Find mean number of emails per day.

ii) Calculate the frequencies of the poisson distribution having the same mean.

Ans: 1.04, 53,55, 29, 10, 3

10. A random variable X has the probability distribution shown below;

x	0	1	2
P(X=x)	0.2	0.3	0.5

i) Find E(X) and V(X)

ii) Calculate E(Y) if Y = 3X +2

Ans: 1.3, 0.61, 5.9

11. Define probability mass function. Is there any consistency in the statement that the mean of binomial distribution is 20 and standard deviation is 4. If no inconsistency is found then find value of p,q and n. **Ans:** 100

12. Find the skewness of following set of data pertaining to kilowatt hours of electricity consumed by 100 persons in a city.

Consumption(in KWH)	Below 10	10-20	20-30	30-40	40 and above
Number of consumers	10	20	40	20	10

Also interpret the result. **Ans:** 0

MODEL SET 3

Group A

Attempt any Two questions **(2×10=20)**

1. Write the algebraic computation expressions for mean and standard deviation based on a given sample $x_1, x_2 \ldots, x_n$. Why they are important in statistics? Write down their properties. Compute the measure of dispersion and skewness from the following scores of 10 samplestudents. **Ans:** 11.31, – 0.574

45	55	30	60	55	55	65	40	45	35

2. Explain the terms - sample space and an event of a random experiment. State the classical and the statistical definition of probability. Which of the two definitions is most useful in statistics and why? If A,B and C are events of a sample space such that
$P(A)=0.5$, $P(A \cap C)=0.2$, $P(A \cap B^c \cap C^c)=0.1$, and $P(A \cap B \cap C)=0.05$.
find $P(A \cap B)$. **Ans:** 0.25

3. A large company wants to measure the effectiveness of newspaper advertising media on sale promotion of its products. A sample of 22 cities with approximately equal populations is selected for study. The sales of the product (Y) in thousand Rs and the level of newspaper advertising expenditure (X) in thousand Rs are recorded for each of the 22 cities (n) and the recorded sum, sum of square, and sum of cross product of X and Y are summarized below.
$\sum Y = 26953,\ \sum X = 660,\ \sum Y^2 = 35528893,\ \sum X^2 = 22700,\ \& \sum YX = 851410$
Using the above summary results:
   a) Compute correlation coefficient r between X and Y, and coefficient of determination.
   b) Fit a simple linear regression model of Y on X using least square method and interpret the estimated slope regression coefficient. **Ans:** 0.52, 0.27, y = 782.174 + 14.765x

Group B

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

4. The number of runs scored by two group of cricket players in a test match are

Group A	10	25	85	72	115	80	52	45	30	10
Group B	120	15	30	35	42	65	80	34	25	15

Test which group is more consistent.

Ans: Group A is consistent

5. Determine kurtosis from following data and comment upon the shape of the distribution.

Customer service time(min)	0-5	5-10	10-15	15-20	20-25	25-30
No. of customers	2	8	26	50	28	6

Ans: 59.865

6. What is exponential distribution? The life time of a large number of components of an equipment has exponential distribution with mean 1/3 per year. Find the probability that life time of component selected at random is at most 6 months.

Ans: 0.776

7. Explain discrete and continuous random variables with suitable examples. Suppose a continuous random variable X has the density function.

$$f(x) = \begin{cases} kx^2 & if \ -1 < x < 2 \\ 0 & otherwise \end{cases}$$

Find (a) value of the constant k, and (b) E(X).

Ans: $\frac{1}{3}$, 1.25

8. Suppose that X and Y have joint density function

$$f(x,y) = \begin{cases} 4xy & If\ 0 < x < 1\ and\ 0 < y < 1 \\ 0 & otherwise \end{cases}$$

Find (a) marginal density functions of X and Y, (b) $P(X \leq 0.3)$ and (c) $P(Y \geq 0.5)$.

Ans: 2x, 2y, 0.09, 0.75

9. Three are three traffic lights on your way home. As you arrive at each light assume that it is either red (R) or green (G) and that it is green with probability 0.7. Construct the sample space by listing all possible eight simple events. Assign probability to each simple event. Are the events equally likely? What is the probability that you stop no more than one time. **Ans:** No, 0.784

10. What do you mean by survey? Differentiate between sample survey and census survey.

11. If a random variable X is normally distributed with a mean of 120 and a standard deviation of 12. Compute the following probabilities: (a) P (X >130), (b) P(X<115), and (c) P(110 <X<130)

Ans: 0.2030.34, 0.593

12. The marks distribution of 100 students of a college was as follows.

Marks	Below 20	20-39	40-59	60-79	80 and above
No. of students	15	20	30	15	10

a. Find the highest mark of the weakest 30% students
b. Find the lowest mark of top 40% students
c. Find limit of marks of middle 50% students

Ans: 31.5, 52.166, 62.83

MODEL SET 4

Group A

Attempt any Two questions **(2×10=20)**

1. Differentiate between central tendency and dispersion? Score obtained by 48 students in a test are given in the following table.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
Number of students	3	8	12	17	6	2

Compute mean , median , mode , quartiles

Ans: 29.37, 30.58, 33.125, 20.83, 37.64

2. Distribution of successful project by two different streams of students in computer are given below:

Successful projects conducted	20	22	23	25	26
Nos. of B Sc CSIT students	5	7	10	8	5
Nos. of BCA students	3	7	15	8	2

Find out which stream students are consistent and why?.

Ans: BCA students are consistent

3. What do you mean by Joint probability distribution? Write down its properties.

Given the following bivariate probability distribution of X & Y.

Y \ X	-1	0	1
0	1/15	2/15	1/15
1	3/15	2/15	1/15
2	2/15	1/15	2/15

Find:(i) P(X=1,Y ≤ 1) (ii)P (Y ≤ 1) (iii) P(X= -1) (iv) P(X = 1/ Y =2)

Ans: $\frac{2}{15}, \frac{2}{3}, \frac{2}{5}, \frac{2}{5}$

Group B

Attempt any Eight questions (8×5 = 40)

4. A marital arithmetic test for 8 questions given to a class of 32 pupils. The result were summarized in the following table

Number of correct answers	0	1	2	3	4	5	6	7	8
Number of pupils	1	1	2	3	4	8	5	4	3

Determine Karl Pearson's coefficient of skewness. Describe the shape of the distribution. **Ans:** - 0.104

5. The pdf of a continuous random variable is given by

$$f(x) = \begin{cases} 12\,x(1-x)^2, & 0 \le x \le 1 \\ 0, & \text{elsewhere} \end{cases}$$

Find E(x), V(x)

Ans: $\frac{2}{5}, \frac{1}{25}$

6. If two random variables X_1 and X_2 have the joint probability density.

$$f(x_1, x_2) = \frac{2}{3}(x_1 + 2x_2), \text{ for } 0 < x_1 < 1, 0 < x_2 < 1$$

0, elsewhere

Find the conditional density of (i) X_1 given $X_2 = x_2$.(ii) X_2 given $X_1 = x_1$

Ans: $\frac{2(x_1 + 2x_2)}{1 + 4x_2}$

7. A jewelry dealer is interested in purchasing gold necklace for which probabilities are 0.18, 0.22, 0.33 and 0.27 respectively that it will be able to sell it for a profit of Rs 5000, Rs 8000, breakeven and sell for a loss of Rs 3000. Find (i) expected profit (ii) Standard deviation of profit. **Ans:** 1850, 4193.745

8. Define normal probability distribution. Explain important properties of normal distribution.

9. In a photographic process, the developing time of prints may be looked upon as a random variable having the normal distribution with a mean of 16.28 seconds and a standard deviation of 0.12 second. Find the probability that it will take (i) anywhere from 16.00 to 16.50 seconds to develop one of the prints, (ii) at least 16.20 seconds to develop one of the prints. **Ans:** 0.956, 0.7454

10. Find the average profit percentage per shop from the selling of Dell computers from following data.

Profit (%)	4–8	8 – 12	12–16	16–20	20–24	24–28	28–32	32–36	36–40
No. of shop	6	10	18	30	15	12	10	6	2

Ans: 18.119

11. What is sampling frame? Differentiate between probability sampling and non probability sampling.

12. The following data gives the experience of machine operators in years and their performance as given by the number of good parts turned out per 100 pieces.

Operator	I	II	III	IV	V	VI	VII	VIII
Experience	16	12	18	4	3	10	5	12
Performance	87	88	89	68	78	80	75	83

Calculate the regression equation of performance on experience and hence estimate the probable performance if an operator has 8 years experiences. Interpret the regression coefficient.

Ans: y = 71.04 + 0.995x.79

MODEL SET 5

Group A

Attempt any Two questions **(2×10=20)**

1. Marks secured by two students in exam of B.Sc. CSIT second semester are given below;

	Discrete structure	Object oriented programming	Microprocessor	Mathematics II	Statistics I
Student A	46	39	50	30	39
Student B	41	33	44	38	32

Determine who is (i) better (ii) intelligent (iii) consistent

Ans: A, A, B

2. What is probability? Differentiate between addition and multiplication law of probability. The following table shows the survey result regarding the purchase behavior of TV's and DVD players in the last six months of 300 house hold.

Purchase TV	Purchase DVD		
	Yes	No	Total
Yes	38	42	80
No	70	150	220
Total	108	192	300

a. Find the probability that a randomly selected household that purchased a TV.

b. Find the probability that a randomly selected household that purchased a TV and a DVD player.

c. What is the probability that he/she purchased a TV or DVD player?

d. What is the probability that a household has purchased a DVD player given that household purchased a TV?

e. What is probability that a household purchased TV given that household has not purchased DVD player?

Ans: 0.266, 0.126, 0.5, 0.43, 0.218

3. The following table gives the distribution of items and also defective items among them according to size groups. Find the correlation coefficient between size and defect in quality. Estimate the defect in quality for size of 44.

Size group	16 – 20	20 – 24	24 – 28	28 – 32	32 – 36	36 – 40
No of items	200	270	340	360	400	300
No of defective items	150	162	170	180	180	120

Ans: – 0.939, 28

Group B

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

4. Determine modal I Q score of 100 CSIT students from the score distribution of students given below;

IQ score	50-59	60- 69	70- 79	80-89	90-99	100-109	110-119	120- 129	130- 139
No. of students	2	4	8	15	21	26	20	3	1

Ans: 104.045

5. Calculate first four moments about mean from following distribution;

Marks	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No of students	6	12	16	10	6

Ans: 0, 139.84, 127.87, 42870.32

6. An examination of 10 applicants was taken by a company on skill and ability of candidates. From the marks obtained by the applicant in Skill and Ability Papers. Calculate the rank correlation coefficient.

Applicant	A	B	C	D	E	F	G	H	I	J
Marks in Skill	38	41	68	41	38	55	85	81	28	41
Marks in Ability	48	39	38	36	58	61	72	83	61	82

Ans: 0.279

7. Calculate coefficient of skewness and kurtosis from following frequency distribution and interpret the result.

Hourly Remuneration(Rs)	No of workers
Below 150	8
Below 200	22
Below 250	40
Below 300	64
Below 350	80
Below 400	92
Below 450	100

Ans: 0.105, 0.235

8. Three computer viruses arrived as an email attachment. Virus A damages the system with probability 0.4. Independently of it, virus B damages the system with probability 0.5. Independently of A and B, virus C damages the system with probability 0.2. What is probability that (i) system gets damaged (ii) System works properly.
Ans: (i) 0.76, (ii) 0.24

9. Two continuous random variable X and Y have the joint density function
$f(x,y) = c(x^2+y)$; $-1 \le x \le 1, 0 \le y \le 1$
0 ; otherwise
   (i) Compute constant c
   (ii) Find marginal density functions of X and Y. Are these two variables independent?
   (iii) Compute probabilities P(Y< 0.6) and P(Y<0.6/X<0.5)
Ans: $\frac{3(x^2+2)}{10}, \frac{2(3y+1)}{5}$, No, 0.456, 0.44

10. Fit Binomial distribution to following data and find expected frequencies

x	0	1	2	3	4	5	6
f	7	6	19	35	23	7	1

Ans: 2, 11, 25, 30, 21, 8, 1

11. Installation of some software package requires downloading 82 files. On average it takes 15 seconds to download one file with variance of 16 second^2. What is probability that software is installed in less than 20 minutes?
Ans: 0.2033

12. Write short notes on
   (a) Kurtosis and its measure
   (b) Gamma distribution

MODEL SET 6

Group A

Attempt any Two questions **(2×10=20)**

1. What is mathematical definition of probability? How it differs from Statistical definition of probability? A sample of 500 respondents was selected in a large IT company to determine various information concerning consumer behavior. Among the question asked was "Do you enjoy shopping for gadgets?" O 240 males 136 answered yes. Of 260 females, 224 answered yes What is the probability that a respondent chosen at random;
   i) Enjoys shopping for gadgets?
   ii) Is a female and enjoy shopping for gadgets?
   iii) Is a female or enjoy shopping for gadgets?

iv) Suppose the respondent chosen enjoys shopping for gadgets, what then is the probability that the individual is a male? **Ans:** 0.72, 0.44, 0.79, 0.37

2. What do you mean by correlation? Following information represents marks secured by 10 students in Mathematics and Statistics;

Student	A	B	C	D	E	F	G	H	I	J
Marks in Mathematics	35	42	51	28	44	35	32	41	37	44
Marks in Statistics	42	28	36	29	50	40	50	34	45	50

Find rank correlation coefficient between marks secured by students in Mathematics and Statistics. **Ans:** 0.109

3. The following frequency distribution represents the weight of 200 laptops.

Weight in lbs	Frequency	Weight in lbs	Frequency
4-5	20	8-9	32
5-6	24	9-10	24
6-7	35	10-11	8
7-8	48	11-12	2

Compute the first three quartiles and quartile deviation.
Ans: 6.12, 7.36, 8.55, 1.22

Group B

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

4. The average marks secured in CSIT exam in the year 2072 by College A and College B are 78 and 80 with variances 100 and 81 respectively. The number of students appeared in the CSIT exam from College A is 100 and College B is 150. Compute the combined mean and combined variance of marks secured by the two groups. **Ans:** 79.2, 89.5

5. From the following frequency distribution, calculate Bowley's coefficient of skewness.

Hourly wages(Rs)	230- 279	280- 329	330- 379	380- 429	430- 479	480- 529
No of workers	12	26	9	4	3	1

Ans: 0.25

6. What do you mean by sampling? Discuss different types of non-probability sampling.

7. What do you mean by expectation? Mention its properties.
A jewelry dealer is interested in purchasing gold necklace for which probabilities are 0.18, 0.22, 0.33 and 0.27 respectively that it will be able to sell it for a profit of Rs 5000, Rs 8000, breakeven and sell for a loss of Rs 3000. Find expected profit? **Ans:** 1850

8. Calculate mean and variance of a Poisson variable X, if $P(x = 4) = P(x = 5)$. **Ans:** 5,5

9. Suppose that life of battery used in laptop of a software company is normally distributed with a mean of 40 months and a standard deviation of 5 months. If at a time, 1000 batteries are issued, how many will need the replacement after 35 months? **Ans:** 841

10. Let (X,Y) be two dimensional random variable with joint pmf $P(x,y)=\frac{x-y+3}{48}$; x=0,1,2,3; y=0,1,2,3. Find: marginal pmf of X and Y. **Ans:** (2x+3)/24, (9y-2)/24

11. Suppose that waiting time (hrs) for bus in a bus station has a negative exponential distribution with parameters $\theta = 5$ hours. What is the probability that a man has to wait at least 15 minutes? Also find expected waiting time for bus. **Ans:** 0.28, 1/5hr

12. The following measurements show the respective height in inches of 10 fathers and their eldest sons

Father	67	63	66	71	69	65	62	70	61	72
Son	68	66	65	70	67	67	64	71	62	63

Find the regression line of son's height on father's height and estimate the height of son for the given height of father as 70 inches. Also determine coefficient of determination and interpret. **Ans:** y=40.43+0.388x, 67.62, 0.266

MODEL SET 7

Group A

Attempt any Two questions **(2×10=20)**

1. From 20 pairs of X and Y variables the following results obtained

$\Sigma X = 127, \Sigma Y = 100, \Sigma X^2 = 860, \Sigma Y^2 = 549, \Sigma XY = 674$

At the time of verification, the following wrong values of X and Y were taken as (10, 14) and (8,6) instead of correct values (8,12)and (6, 8). Find correct value of(i) correlation coefficient (ii) coefficient of determination and interpret. **Ans:** 0.47, 0.22

2. Fit Poisson distribution and find the expected frequencies.

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

Ans: 68, 120, 106, 63, 28, 10, 3

3. Calculate the first four moments about the mean from following distribution and then findthe measure of skewness and kurtosis and comment upon the distribution.

X	0	1	2	3	4	5	6	7	8	9	10
f	5	10	30	70	140	200	140	70	30	10	5

Ans: 0, 3.49

Group B

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

4. The number of telephone calls received at an exchange for 200 successive one-minute intervals are given below.

No of calls	0	1	2	3	4	5	6	Total
Frequency	15	22	28	35	42	34	24	200

Compute the mean, median and mode. **Ans:** 3.325, 3, 4

5. From following grouped frequency distribution compute Karl Pearson's coefficient of skewness.

Mid value of income in '00' Rs.	150	250	350	450	550	650	750	850
No of staff in IT company	80	105	120	165	100	90	60	40

Ans: 0.08

6. An examination of 10 applicants was taken by a company on skill and ability of candidates. From the marks obtained by the applicant in Skill and Ability Papers. Calculate the rank correlation coefficient.

Applicant	A	B	C	D	E	F	G	H	I	J
Marks in Skill	38	41	68	41	38	55	85	81	28	41
Marks in Ability	48	39	38	36	58	61	72	83	61	82

Ans: 0.278

7. Describe role of Statistics in Computer science and information technology.

8. A book case contains 6 digital logic books and 9 microprocessor books. Four books are selected at a time. Find the probability for the first selection to give 4 digital logic and second to give 4 microprocessor books if (i) the books are replaced before the second selection. (ii) the books are not replaced before the second selection. **Ans:** 6/5915 , 3/715

9. Three lab contain 4 Dell and 3 Acer, 4 Dell and 5 Acer, 3 Dell and 4 Acer computers respectively. One computer is selected from each lab. Find the expected number of Dell computer selected. **Ans:**1.44

10. The life time of Lenovo cell phone has gamma distribution with parameter 2. Find the probability that cell phone has life (i) more than 2 years (ii) Between 3 years to 5 years.
Ans: 0.405, 0.159

11. A dangerous computer virus attacks a folder consisting of 250 files. Files are affected by the virus independently of one another. Each file is affected with the probability 0.032. What is the probability that more than 5 files are affected by this virus?
Ans: 0.808

12. Write short note on
    (i) Sample survey and census survey
    (ii) Joint probability mass function and joint probability density function.

MODEL SET 8

Group A

Attempt any Two questions **(2×10=20)**

1. Drive D of a computer having 25 folders contains the following number of files.
4,0,5,2,3,1,4,3,2,3,4,3,1,6,3,2,3,4,2,1,0,3,2,5,4
    a) List the five number summary
    b) Construct a box plot for the data
    c) Are the data skewed?
Ans: {0, 2, 3, 4, 6}, symmetrical

2. 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 interpretyour result.
Ans: - 0.309

3. Fit the Binomial distribution and find the expected frequencies for the following data.

X	0	1	2	3	4	5	6	total
f	7	6	19	35	23	7	1	98

Ans: 2, 11, 25, 30, 21, 8, 1

Group B

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

4. The frequency distribution of time required to open the operating system of 200 computers is given below.

Time in seconds	No. of Computers	Time in seconds	No. of Computers
0-4	2	20-24	48
5-9	20	25-29	32
10-14	35	30-34	18
15-19	40	35- 39	5

Compute the standard deviation. **Ans:** 7.87..

5. What do you mean by data? Describe different types of data.
6. National Planning Commission (NPC) is performing preliminary study to determine the relationship between certain economic indicator and annual percentage change in Gross National Product(GNP). The concern is to estimate the percentage change in GNP. One of such indicator being examined is government's deficit. Data on 6 years are given below;

Percentage change in GNP	3	1	4	1	2	3
Government deficit in lakh Rs	50	200	70	100	90	40

a) Develop the estimating equation to predict percentage change in GNP from government deficit.

b) Compute the coefficient of determination and interpret.

Ans: y = 3.725 – 0.015x, 0.524

7. Compute first four central moments from following series;

Class interval	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
frequency	5	7	9	7	5

Ans: 0, 163.63, 0, 52727.27]

8. Find mean and S.D. of a distribution having pdf $f(x) = ke^{-x/\sigma}$, $0 < x < \infty$ and $\sigma > 0$.

[Ans: σ, σ]

9. Let two dimensional random variable (X,Y) have joint pdf $f(x,y) = \begin{cases} k(6-x-y); 0 \le x \le 2, 2 \le y \le 4 \\ 0 \quad \text{elsewhere} \end{cases}$. Determine: (i) Constant k (ii) $P(X \le 1 \cap Y < 3)$ (iii)$P(Y \le 3)$ **Ans:** 1/8, 3/8, 5/8]

10. A source of liquid is known to contain bacteria with the mean no of bacteria per cubic centimeter equal to 3. Ten 1 cubic centimeter test tubes are filled with the liquid. Assuming the Poisson distribution is applicable. Calculate the probability that (i) All 10 test tube will show growth (i.e. at least 1 bacteria each) (ii)Exactly 7 test tubes will show growth. **Ans:** 0.6008, 0.215

11. Incomes of a group of 10,000 computer operators were found to be normally distributed with mean Rs. 15,200 and standard deviation Rs. 1600, find (i) highest income of poorest 2000 computer operators (ii) lowest income of richest 1000 computer operators. **Ans:** 1385.6, 1724.8

12. Write short notes on following:
    i. Gamma distribution and its characteristics
    ii. Advantages and disadvantages of sampling