Confidential

G.C.E. (A/L) Examination - 2021 (2022)

Marking Scheme

31 - Business Statistics

This has been prepared for the use of marking examiners. Changes would be made according to the views presented at the Chief/Assistant Examiners' meeting.

Amendments to be included.

G.C.E. (A/L) Examination - 2021 (2022)

31 - Business Statistics

Allocation of Marks

I	Paper	-	1	×	50	=	50
II	Paper	-	20	×	05	=	100

$$\text{Final mark} = 50 + \frac{100}{2}$$

$$= 100$$

Common Techniques of Marking Answer Scripts.

It is compulsory to adhere to the following standard method in marking answer scripts and entering marks into the mark sheets.

1. Use a red color ball point pen for marking. (Only Chief/Additional Chief Examiner may use a mauve color pen.)

2. Note down Examiner's Code Number and initials on the front page of each answer script.

3. Write off any numerals written wrong with a clear single line and authenticate the alterations with Examiner's initials.

4. Write down marks of each subsection in a △ and write the final marks of each question as a rational number in a □ with the question number. Use the column assigned for Examiners to write down marks.

Example: **Question No. 03**

(i) ..... √ $\frac{4}{5}$

(ii) ..... √ $\frac{3}{5}$

(iii) ..... √ $\frac{3}{5}$

03 (i) $\frac{4}{5}$ + (ii) $\frac{3}{5}$ + (iii) $\frac{3}{5}$ = $\frac{10}{15}$

MCQ answer scripts: (Template)

1. Marking templets for G.C.E.(A/L) and GIT examination will be provided by the Department of Examinations itself. Marking examiners bear the responsibility of using correctly prepared and certified templates.

2. Then, check the answer scripts carefully. If there are more than one or no answers Marked to a certain question write off the options with a line. Sometimes candidates may have erased an option marked previously and selected another option. In such occasions, if the erasure is not clear write off those options too.

3. Place the template on the answer script correctly. Mark the right answers with a '√' and the wrong answers with a 'X' against the options column. Write down the number of correct answers inside the cage given under each column. Then, add those numbers and write the number of correct answers in the relevant cage.

Structured essay type and assay type answer scripts:

1. Cross off any pages left blank by candidates. Underline wrong or unsuitable answers. Show areas where marks can be offered with check marks.

2. Use the right margin of the overland paper to write down the marks.
3. Write down the marks given for each question against the question number in the relevant cage on the front page in two digits. Selection of questions should be in accordance with the instructions given in the question paper. Mark all answers and transfer the marks to the front page, and write off answers with lower marks if extra questions have been answered against instructions.

4. Add the total carefully and write in the relevant cage on the front page. Turn pages of answer script and add all the marks given for all answers again. Check whether that total tallies with the total marks written on the front page.

Preparation of Mark Sheets.

Except for the subjects with a single question paper, final marks of two papers will not be calculated within the evaluation board this time. Therefore, add separate mark sheets for each of the question paper. Write paper 01 marks in the paper 01 column of the mark sheet and write them in words too. Write paper II Marks in the paper II Column and wright the relevant details.

***

AL/2021(2022)/31/E-I

0647

අධ්‍යයන පොදු සහතික පත්‍ර (උසස් පෙළ) විභාගය, 2021(2022)
கல்விப் பொதுத் தராதரப் பத்திர (உயர் தர)ப் பரீட்சை, 2021(2022)
General Certificate of Education (Adv. Level) Examination, 2021(2022)

ව්‍යාපාර සංඛ්‍යානය	I
வணிகப் புள்ளிவிவரவியல்	I
Business Statistics	I

31 E I

පැය දෙකයි
இரண்டு மணித்தியாலம்
Two hours

Instructions:

- *Answer **all** questions.*
- *Write your **Index Number** in the space provided in the answer sheet.*
- *Statistical tables will be provided. Use of Calculator is **not allowed**.*
- *Instructions are given on the back of the answer sheet. Follow those carefully.*
- *In each of the questions **1** to **50**, pick one of the alternatives from (1), (2), (3), (4), (5) which is **correct** or **most appropriate** and **mark your response on the answer sheet with a cross** (×) on the number of the correct option in accordance with the instructions given at the back of the answer sheet.*

1. Consider the following statements on ordinal scale measurement.

A - The order is meaningful but the difference is not meaningful.

B - The magnitude of the interval and the value zero are meaningful.

C - The value itself is not meaningful.

The true statement/s from the above is/are,

(1) only A. (2) only A and B. (3) only A and C.
(4) only B and C. (5) all A, B and C.

2. Which of the following statements is true?

(1) Response errors in a completed questionnaire are removed during the editing stage.
(2) The response rate in the interview method is less than the response rate in the postal method.
(3) The way the questions are asked in a questionnaire sometimes affects the response.
(4) Data classification affects the reliability of the data.
(5) Questions with choices are not included in the questionnaire.

3. Which of the following statements is true?

(1) Class width is used to calculate the area of each rectangle of a histogram.
(2) Only the height of the columns should be considered when creating a multiple vertical bar chart.
(3) The values of a relative frequency distribution are given in integers only.
(4) It is not always possible to construct a histogram for a frequency distribution with unequal classes although it is always possible with equal classes.
(5) Data represented by a percentage component bar chart can also be represented by a pie chart.

[see page two

4. The marks obtained by 27 students are given in the following stem-and-leaf diagram. (Stem unit is 10 and leaf unit is 1)

3	2	3	4	5			
4	1	3	4	4	5	6	7
5	0	2	3	4	5	7	8
6	2	3	4	5	6		
7	4	5	5				
8	6						

What is the inter quartile range (IQR)?

(1) 15.5 (2) 16.0 (3) 20.0 (4) 32.0 (5) 54.0

5. Consider the following statements.

A - The width of a class interval is calculated by taking the difference between the upper and lower class boundaries.

B - Raw data cannot be accurately retrieved from a frequency polygon.

C - Relative frequency polygon cannot be used to obtain exact median of a data set.

The true statement/s from the above is/are

(1) only A. (2) only C. (3) only A and B.

(4) only B and C. (5) all A, B and C.

6. Which of the following statements is true?

(1) The median which is a measure of the central tendency does not represent all observations.

(2) The total frequency less than any upper class boundary is indicated by the percentage ogive with that class boundary.

(3) When there is a positive skewness in the distribution, there is a long tail on the right side than the left side of the frequency polygon.

(4) The kurtosis measurements represent the asymmetry of a distribution.

(5) If the standard deviation of a distribution is large, then the sample mean is a reliable estimator of the central tendency.

7. Which of the following statements is **false**?

(1) The geometric mean is appropriate when the data is in the form of ratios or percentages.

(2) The kurtosis coefficient is zero for a normal distribution.

(3) The most suitable dispersion measure to compare variability of any two data sets is the coefficient of variation.

(4) For a moderately skewed distribution, (mean-mode)= 3(mean-median).

(5) The dispersion measurements provide information about the reliability of the sample mean.

8. A person travels from one city to another city on a bicycle. His speeds for each of the $\frac{1}{4}^{th}$ of the total distance are 30 km/hr, 60 km/hr, 40 km/hr, and 20 km/hr, respectively. What is the average speed for the cyclist?

(1) 20 km/hr (2) 24 km/hr (3) 30 km/hr (4) 32 km/hr (5) 34 km/hr

9. Consider the following statements.

A - The coefficient of variation of a data set is 100% when the mean is 5 and the standard deviation is 5.

B - When a constant value is subtracted from each observation in a group of observations, it directly affects the variability of the data set.

C - If the coefficient of variation of a data set is high, data are less consistent.

The true statement/s from the above is/are,

(1) only A. (2) only A and B. (3) only A and C.

(4) only B and C. (5) all A, B and C.

[see page three

10. What is the inter quartile range for the standard normal distribution?

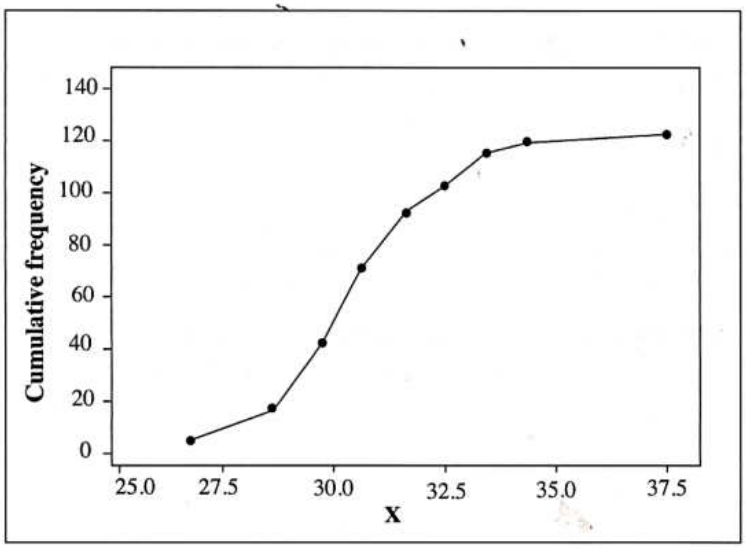
(1) −0.20 (2) 0.20 (3) 0.50 (4) 0.75 (5) 1.35

11. Consider the following statements with respect to the cumulative distribution function given below.

A - $(Q_3 - Q_2) = (Q_2 - Q_1)$

B - $(Q_3 - Q_2) > (Q_2 - Q_1)$

C - Box plot of the distribution has a longer right whisker than the left whisker.



The true statement/s from the above is/are

(1) only A. (2) only B. (3) only A and C.

(4) only B and C. (5) all A, B and C.

12. For a certain distribution, Pearson's coefficient of skewness is 0.24, the median is 24.50 and the mean is 25.14. What is the variance of the distribution?

(1) 32 (2) 48 (3) 62 (4) 64 (5) 72

13. The regression line of Y on X and X on Y are given by $Y = b_0+b_1X$ and $X=a_0+a_1Y$ respectively. The estimate of the Pearson's correlation coefficient between $X + k^2$ and Y, where k is a constant, is given by

(1) a_1b_1. (2) $k^2 + a_1b_1$. (3) $k \pm \sqrt{a_1b_1}$. (4) $\pm \sqrt{a_1b_1}$. (5) $k + a_1b_1$.

14. Which of the following statements is true with respect to scatter diagram for precise estimate of regression slope coefficient in simple linear regression?

(1) Small error variance and large variance of X

(2) Small error variance and small variance of X

(3) Large error variance and small variance of X

(4) Large error variance and large variance of X

(5) Small error variance and zero variance of X

15. Three individuals A, B and C were ranked by two judges. If B was ranked 2 by both judges then the possible value/s of rank correlation coefficient, r_k is/are

(1) $r_k = 0$. (2) $-1 < r_k < 1$. (3) $r_k = 0$ or $r_k = 1$.

(4) $r_k = -1$ or $r_k = 1$. (5) $r_k = -1$ or $r_k = 0$.

[see page four

16. Consider the following statements about rank (Spearman's) correlation coefficient.

A - It measures only the strength of linear relationship between two random variables.
B - It measures the consistency of direction in the relationship between two random variables.
C - It measures only the non-linearity of the relationship between two random variables.

The true statement/s from the above is/are,

(1) only A. (2) only B. (3) only A and C.
(4) only B and C. (5) all A, B and C.

17. After rolling a dice for 600 times under the same conditions, a student stated that the outcomes of the dice occur equally likely. Which probability approach/approaches has/have been used to make the above statement?

(1) Classical approach
(2) Subjective approach
(3) Relative frequency approach
(4) Classical approach and relative frequency approach
(5) Relative frequency approach and subjective approach

18. A, B and C are three collectively exhaustive and mutually exclusive events in the sample space S. It is given that $P(A)=7p^2-p$, $P(B)=1-4p+5p^2$ and $P(C)=1-2p$. If $P(A)<P(C)$ then what is the value of 'p'?

(1) $\frac{1}{12}$ (2) $\frac{1}{6}$ (3) $\frac{1}{4}$ (4) $\frac{1}{3}$ (5) $\frac{1}{2}$

19. If A and B are two events such that $P(A|B')= 3P(A|B)$ and $P(B')= 2P(B)$ then what is the value of $P(B'|A)$?

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

20. The probability of occurring event A is four times much as occurring an event B. If these two events are independent and $P(A\cap B)=p$, $(p>0)$ then what is the value of $P(A)$?

(1) $\frac{1}{4}$ (2) $\frac{p}{2}$ (3) $\frac{\sqrt{p}}{2}$ (4) $4p$ (5) $2\sqrt{p}$

21. The random variable X has the following probability distribution.

X:	1	2	3	4	5
$P(x)$:	a	0.1	b	0.3	0.1

If a and b are constant terms and it is given that $E[X-5]=-1.8$, then what is the value of the mode of the above distribution?

(1) 0.3 (2) 0.4 (3) 3 (4) 3.2 (5) 4

22. Which of the following statements is true?

(1) If X is a continuous random variable and b is a constant then $P(X=b)=p$ where $0 < p < 1$.
(2) The expectation of a discrete random variable is always a positive value.
(3) The mean and standard deviation of a set of negative values are also negative.
(4) The variance of a discrete random variable X with zero mean is equal to $E[X^2]$.
(5) If X is a discrete random variable, then $Var\left[\frac{X}{2}+7\right] = 4Var(X)$.

[see page five

23. If X follows a binomial distribution with $n = 5$ and $p = 0.4$ then what is $P(X>3 \mid X>2)$?

(1) $\frac{870}{6826}$ (2) $\frac{870}{3456}$ (3) $\frac{870}{3174}$ (4) $\frac{3174}{6630}$ (5) $\frac{2304}{3456}$

24. Customers arrive randomly at a certain bank according to a Poisson distribution at an average rate of 3 per minute period. What is the probability that at least two customers arrive in the 30 second interval ?

(1) $4e^{-3}$ (2) $1 - 4e^{-3}$ (3) $2.5e^{-1.5}$ (4) $1 - 2.5e^{-1.5}$ (5) $1 - 3e^{-1.5}$

25. A certain company sponsors a bicycle race. Suppose the time it takes to ride the race has a normal distribution with mean 62.5 minutes and standard deviation of 5 minutes. If the company decided to reward only the fastest 27% of the riders of the race, what is the cut off time to receive an award?

(1) 58.80 minutes (2) 59.45 minutes (3) 63.85 minutes (4) 65.55 minutes (5) 66.20 minutes

26. What is the most appropriate sampling technique for a population with linear trend?

(1) Stratified random sampling (2) Simple random sampling
(3) Systematic sampling (4) Cluster sampling
(5) Quota sampling

27. Consider the following statements about sampling.

A - Quota sampling is the combination of convenient and stratified sampling.
B - Non-probability sampling occurs when a researcher's goal is to study a specific sample of interest and be less concerned about generalizing to the total population.
C - Non-probability sampling estimates are always worse than that of probability sampling.

The true statement/s from the above is/are

(1) only A. (2) only A and B. (3) only A and C.
(4) only B and C. (5) all A, B and C.

28. Let a random variable Y has a N(μ, 597). What is the Var ($\bar{y}$) from a simple random sample of size n=20 without replacement if population size is 200?

(1) 5.20 (2) 27.00 (3) 27.30 (4) 29.85 (5) 33.00

29. Consider the following statements about the statistical properties of three estimators; $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ of μ, where

$$\hat{\theta}_1 = \frac{1}{n+1}\sum_{i=1}^{n} X_i, \qquad \hat{\theta}_2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i, \qquad \hat{\theta}_3 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

A - $\hat{\theta}_1$ is biased and more precise than $\hat{\theta}_3$
B - $\hat{\theta}_2$ is unbiased and more precise than $\hat{\theta}_1$
C - $\hat{\theta}_3$ is unbiased and less precise than $\hat{\theta}_1$

The true statement/s from the above is/are,

(1) only A. (2) only B. (3) only C.
(4) only A and C. (5) all A, B and C.

30. Consider the following statements about the standard error of an estimator.

A - It depends on the sample size.
B - It depends on the distribution of the estimator.
C - It depends on the values of at least one parameter of the distribution of the estimator.

The true statement/s from the above is/are,

(1) only A. (2) only B. (3) only C.
(4) only A and C. (5) all A, B and C.

[see page six

31. A simple random sample is to be taken to estimate the population proportion, P and it is intended to estimate P within ± 0.0196 of the true value at $\alpha = 0.05$. What is the required sample size to estimate P ignoring finite population correction if the population proportion is expected to lie between 0.1 and 0.2?

(1) 1000 (2) 1200 (3) 1300 (4) 1500 (5) 1600

32. A random sample of size 9 was selected from a population $N(\mu, \sigma^2)$. What is the point estimate of σ if 90% confidence interval for μ is (12.2, 30.8)?

(1) 12 (2) 14 (3) 15 (4) 16 (5) 18

33. Consider testing $H_0 : P = 0.4$ *vs* $H_1 : P > 0.4$, where P is the probability of observing a head in a single toss of a coin. What is the probability of type I error of the test if H_0 is rejected when there are 7 or more heads in 10 tosses?

(1) 0.032 (2) 0.041 (3) 0.050 (4) 0.055 (5) 0.172

34. Consider the following statements on the central limit theorem.

A - The distribution of sample mean has a normal distribution.

B - The distribution of $\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ is a standard normal distribution as $n \longrightarrow \infty$

C - The sample mean $\overline{X}$ of a random sample from any distribution with finite variance, σ^2 and the mean, μ, is approximately normally distributed with μ and $\frac{\sigma^2}{n}$ when sample size, n, becomes large.

The true statement/s from the above is/are

(1) only A. (2) only A and B. (3) only A and C.
(4) only B and C. (5) all A, B and C.

35. Consider testing H_0: $\mu = 15$ vs H_1: $\mu > 15$ using a random sample size 36 from a $N(\mu, 9)$. What is the probability of type II error for a specific H_1: $\mu = 16$, if the probability of type I error is 0.05?

(1) 0.05 (2) 0.29 (3) 0.35 (4) 0.36 (5) 0.95

36. Consider testing H_0: $\mu = 110$ vs H_1: $\mu > 110$ using a random sample size 15 from a $N(\mu, 1500)$. What is the critical region if the power of the test is 0.8413 and H_1: $\mu = 125$?

(1) $\overline{X} > 115$ (2) $\overline{X} < 120$ (3) $\overline{X} > 135$ (4) $\overline{X} < 148$ (5) $\overline{X} > 148$

37. Consider testing H_0: $\mu = 200$ vs H_1: $\mu < 200$ using a random sample of size 9 from $N(\mu, 81)$. What is the sampling distribution of $\overline{X}$ for a specific alternative hypothesis, H_1: $\mu = 180$?

(1) $\overline{X} \sim N(-20, 9)$ (2) $\overline{X} \sim N(20, 9)$ (3) $\overline{X} \sim N(180, 3)$
(4) $\overline{X} \sim N(180, 9)$ (5) $\overline{X} \sim N(200, 3)$

38. Consider the following statements about the F-test in Analysis of Variance.

A - A required condition for F-test to be used is that the response variable is distributed normally.

B - If between group sum of squares is large, value of F-test statistic will be large.

C - F-test statistic is the sum of sample variances.

The true statement/s from the above is/are

(1) only A. (2) only B. (3) only C.
(4) only A and B. (5) all A, B and C.

[see page seven

39. Consider the following statements about Analysis of Variance model $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$.

A - Y_{ij} and ε_{ij} are distributed normally.

B - Y_{ij} and ε_{ij} both have same variance.

C - Y_{ij} and ε_{ij} both have same mean value.

The true statement/s from the above is/are,

(1) only A. (2) only B. (3) only C.
(4) only A and B. (5) all A, B and C.

40. A dice was tossed 60 times and the summary of the results is as follows.

Value	1	2	3	4	5	6
Frequency	15	13	9	5	11	7

The test statistic and the degrees to freedom for the test H_0 : fair dice vs H_1 : not a fair dice, respectively are,

(1) 5 and 7. (2) 7 and 5. (3) 7 and 6. (4) 9 and 5. (5) 9 and 6.

41. Consider the following statements.

A - The additive time series model assumes that the components of the model are independent of each other.

B - The multiplicative time series model assumes that the components of the model are of independent of each other.

C - All the components of a multiplicative time series model are expressed in percentages.

The true statement/s from the above is/are,

(1) only A. (2) only B. (3) only A and C.
(4) only B and C. (5) all A, B and C.

42. The estimated trend equation of sales (Y) of a commodity is given by $Y = 360 + 36X$ where X is year with the origin is 1988, and Y is number of units sold per year. What is the estimated sale for the month of October 1994?

(1) 47.80 (2) 48.03 (3) 48.88 (4) 55.00 (5) 58.87

43. Ratio to trend method for seasonal indices provides good results if,

(1) the periods are of long duration. (2) the periods are given six monthly.
(3) the periods are given monthly. (4) the periods are damped fluctuation.
(5) the periods are given in annually.

44. Which component/s of a time series is/are attached to the term recession?

(1) trend (2) seasonal (3) cyclical
(4) trend and seasonal (5) trend and irregular

45. Consider the following statements.

A - Lower control limit indicates the lowest limit for variation in the process due to assignable causes.

B - The vertical axis of the control chart represents magnitude of the quality statistic of each sample.

C - If all the sample points fall inside the two control limits, there is no presence of assignable causes of variation in the process of production.

The true statement/s from the above is/are,

(1) only A. (2) only B. (3) only A and B.
(4) only B and C. (5) all A, B and C.

[see page eight

46. Consider the following statements.

A - Sample points falling outside the three sigma limits from the mean ($\mu \pm 3\sigma$) is due to the presence of some assignable causes in the production process.

B - The assignable causes of variation are due to non-random causes.

C - Chance variation cannot be reduced with a modification in the process itself.

The true statement/s from the above is/are,

(1) only A. (2) only B. (3) only A and B

(4) only B and C. (5) all A, B and C.

47. Consider the following statements.

A - The minimum level of quality acceptable in a given lot of products is called Acceptable Quality Level (AQL).

B - The relationship between the probability of accepting the lot of and the percent defective lot is described by Operating Characteristic Curve.

C - Rejecting a good quality lot is called consumer's risk.

The true statement/s from the above is/are,

(1) only A. (2) only B. (3) only A and B.

(4) only B and C. (5) all A, B and C.

48. The wages per month for an employee were Rs. 20 000 in year 2000, and 25 000 in year 2005. The consumer price index for these years were 95 and 190, respectively. What is the increase in the real wage index for the period?

(1) −62.5 (2) −37.5 (3) 37.5 (4) 62.5 (5) 64

49. Which index/ indices satisfies/ satisfy both time reversal and factor reversal tests?

(1) Fisher's Index
(2) Laspeyre's Index
(3) Marshall-Edseworth's Index
(4) Fisher's Index and Laspeyre's Index
(5) Laspeyre's index and Marshall-Edseworth's Index

50. Suppose a worker of a factory earns daily salary of Rs. 500/= and his expenditure pattern is as given below.

Group	Index	Expenditure (Rs.)
Food	125	x
Clothing	150	200
House rent	160	y
Fuel	200	50

If the cost of living index for a day is 150, and he spends all his income during a day, the expenses on food and house rent in rupees, respectively are,

(1) 142 and 108. (2) 143 and 107. (3) 140 and 110.

(4) 145 and 105. (5) 150 and 100.

* * *

G.C.E. (A/L) Examination - 2021 (2022)

Subject code **31** **Subject** **Business Statistics**

Marking scheme

Paper I

Question No	Answer No	Question No	Answer No	Question No	Answer No	Question No	Answer No	Question No	Answer No
01.	3	11.	4	21.	3	31.	5	41.	1
02.	3	12.	4	22.	4	32.	3	42.	3
03.	1	13.	4	23.	3	33.	4	43.	3
04.	3	14.	1	24.	4	34.	4	44.	3
05.	5	15.	4	25.	2	35.	4	45.	2
06.	3	16.	2	26.	1	36.	1	46.	5
07.	2	17.	3	27.	2	37.	4	47.	3
08.	4	18.	3	28.	2	38.	1	48.	2
09.	3	19.	4	29.	4	39.	4	49.	1

✪ **Special Instruction :**

One Mark for One Answer
Total Marks 1 × 50 = 50

Paper II

Part I

1. (a) Briefly explain the importance of Business Statistics to a business institute.

1. In order to evaluate the progress of the daily business affairs
2. In order to maintain a high quality in the products, statistical quality control techniques can be applied
3. In order to maintain a stable business quota through market surveys and marketing research.
4. Related to Human Resource Management statistical techniques such as weighted mean, Rank correlation etc.. can be applied in enrolment of employees, training promoting and motivating the employees
5. In order to make important business decisions examining the relationship between various business variables using regression and correlation analysis
6. In order to determine the real values of money values.
7. To Provide in the basis for inferential analis
8. In order to make optimal decisions through uncertainties.

(3 marks)

(b) Giving reasons, state whether each of the following statements is true or false.

(i) Statistics are used to study individual phenomena.

False - Statistics is about studying aggregate of Data

(1 mark)

(ii) In the case of a personal interview method, the interviewer may not be able to bias the responses.

False -The interviewer may be able to bias the responses, if he does not work honestly or strategically.

(1 mark)

(iii) Cumulative frequency percentages of class intervals can be determined directly using Ogives.

False - Ogive is given on the basis of cumulative frequency, not on the basis of cumulate frequency percentage

(1 mark)

(iv) It is not always necessary to examine the data before analysis.

False- it is always good to examine the data before analysis since we cannot guarantee the reliability of the data.

Examining the data before analysis is necessary since reliability of data cannot be guarantee

(1 mark)

(c) Frequency distribution of marks obtained by 100 students for a certain subject is given below.

Mark interval	0-19	20-39	40-59	60-79	80-99
Frequency	10	20	50	15	05

(i) Draw greater than and less than cumulative frequency curves on the same plot. (03 *marks*)

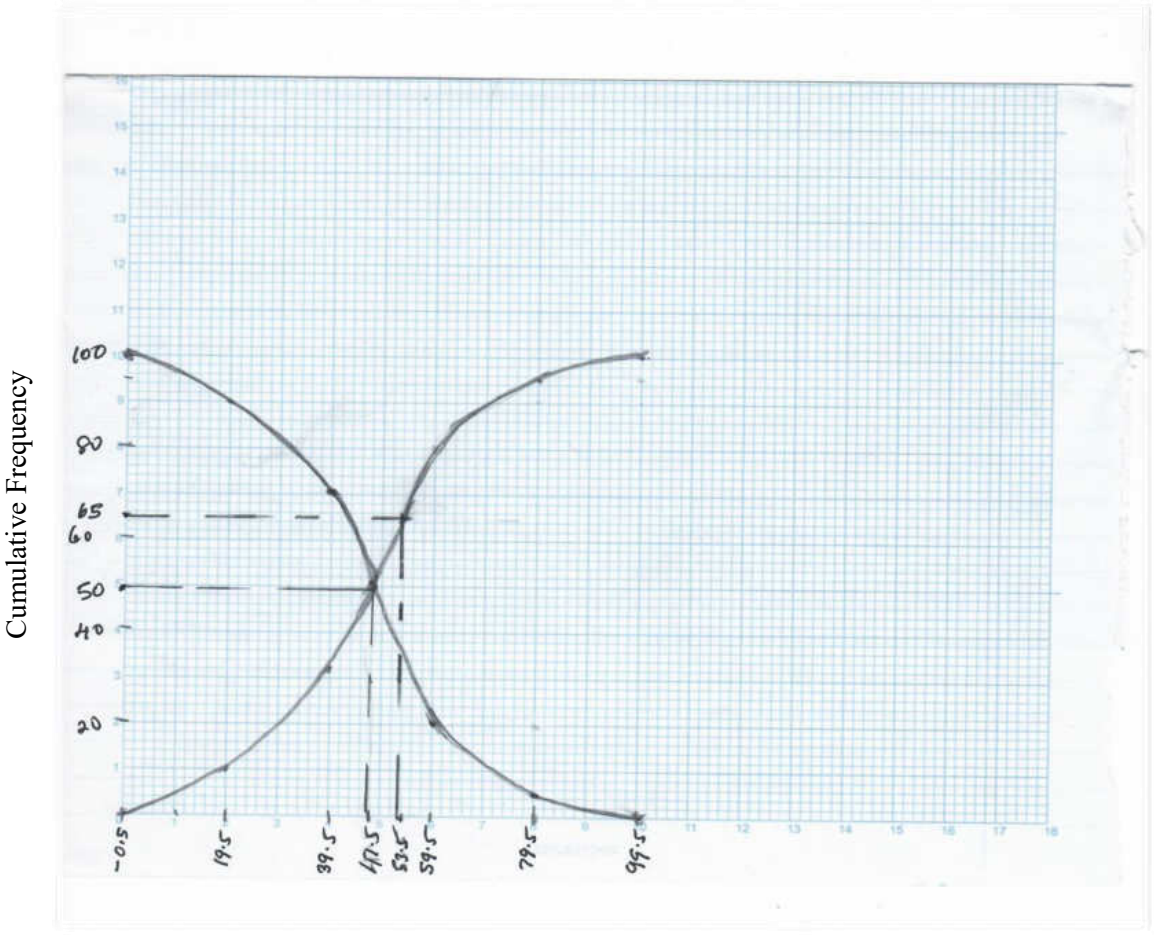
"less than" cumulative frequency distribution

Cl : Int.	**Frequency**	**Less than UCB**	**Cumulative Frequency**
		Less than 0.5	0
0 - 19	10	Less than 19.5	10
20 - 39	20	Less than 39.5	30
40 - 59	50	Less than 59.5	80
60 - 79	15	Less than 79.5	95
80 - 99	05	Less than 99.5	100

" greater than" cumulative frequency distribution

Cl : Int.	**Frequency**	**LCB**	**Cumulative Frequency**
20 - 39	20	19.5 or more	90
40 - 59	50	39.5 or more	70
60 - 79	15	59.5 or more	20
80 - 99	05	79.5 or more	05
		99.5 or more	0

(i)



(03 marks)

(ii) Find the median of the distribution using the plot. (01 *mark*)

From sraph,

Median= 47.5

or

= 46.5 - 48.5

(01 mark)

(iii) Find the 65th percentile of the marks using the plot. (02 *marks*)

$$P_{65} = 53.5$$

or = 52.5 , - 54.5

(02 marks)

(d) The following data are the heights (cm) of chilly plants recorded from two nurseries A and B.

Nursery A

8	7	6	5	4	3	9	10	11	12
11	10	9	7	7	8	9	10	7	8
8	9	10	10	9	8	8	9	9	8

Nursery B

4	5	7	6	6	8	9	10	11	11
9	10	9	8	7	10	9	8	7	7
8	9	8	9	9	9	9	9	9	9

(i) Construct stem-and-leaf plot for each data set separately. (02 *marks*)

Nursery A

stem	Leaf
0	3 4 5 6 7 7 7 7 8 8 8 8 8 8 8 9 9 9 9 9 9 9
1	0 0 0 0 0 1 1 2

Nursery B

stem	Leaf
0	4 5 6 6 7 7 7 7 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9 9 9
1	0 0 0 1 1

(02 marks)

optional

1 3 0
2 4 0
3 5 0
4 6 0
8 7 0000
15 8 0000000
15 9 0000000
8 10 00000
3 11 00
1 12 0

Stem-and-leaf of B N = 30
Leaf Unit = 0.10

1 4 0
2 5 0
4 6 00
8 7 0000
13 8 00000
(12) 9 000000000000
5 10 000
2 11 00

Ordered series

A

3	4	5	6	7	7	**7**	**7**	8	8	8	8	8	8	**8**	**9**	9	9	9	9	9	9	10	10	10	10	10	11	11	12

B

4	5	6	6	7	7	**7**	**7**	8	8	8	8	8	9	**9**	**9**	9	9	9	9	9	9	9	9	9	10	10	10	11	11

(ii) Construct box and whicker plot for each data set separately on the same graph. (03 *marks*)

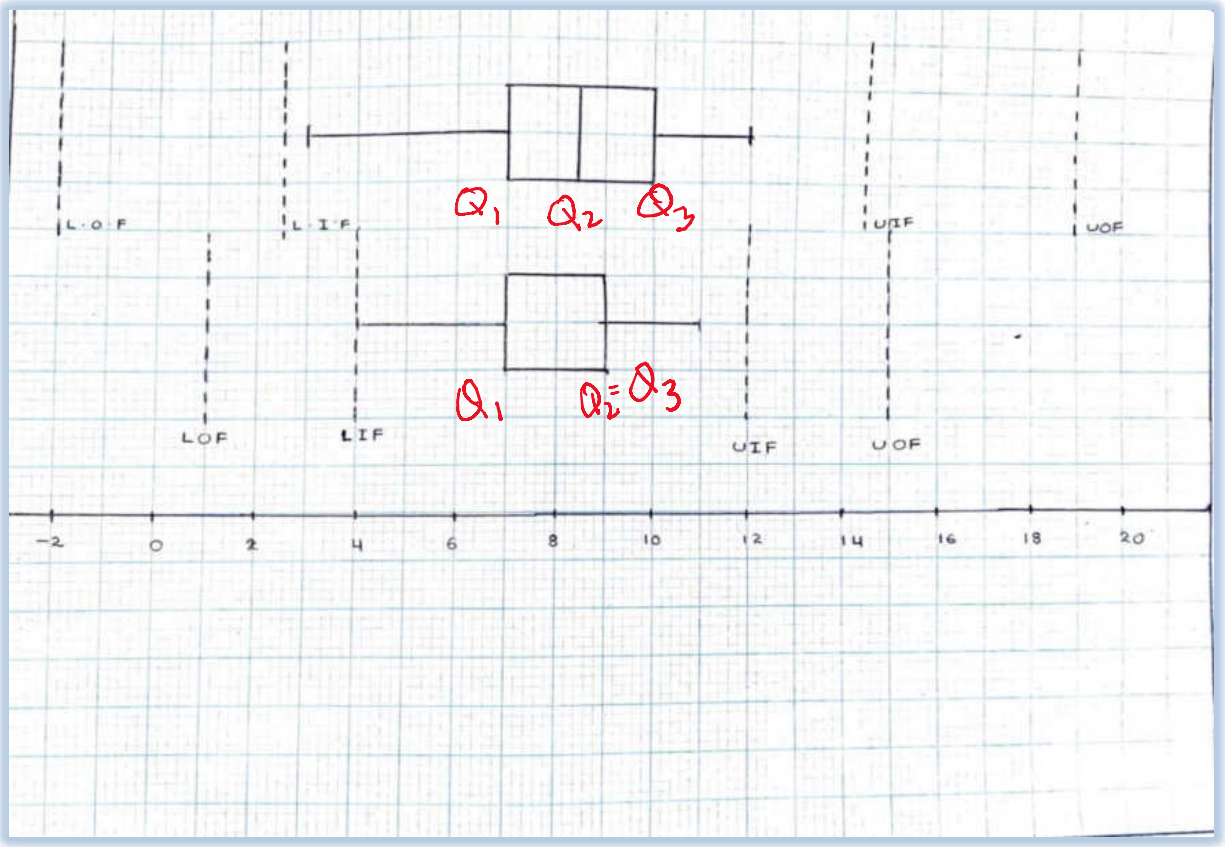
	A	B
Q_1 =	**7**	**7**
Q_2 =	**8.5**	**9**
Q_3 =	**10**	**9**
Lower Inner Fence $Q1_{.}$ -1.5(IQR)	7-4.5 =**2.5**	=7-3 =**4**

Upper Inner Fence Q_3+1.5(IQR)	10+4.5 **=14.5**	=9+3 =**12**
Lower Outer Fence Q_1-3(IQR)	7-9 **= -2**	=7-6 =**1**
Upper Outer Fence Q_3 +3(IQR)	10+9 =**19**	=9+6 =**15**
Lower adjacent value	**3**	**4**
Highest adjacent value	**12**	**11**

(03 marks)

A

B



A

B

(iii) Compare the shape of the distributions of the two data sets based on the plots. (02 *marks*)

A	B
$(Q_2 - Q_1)_A < (Q_2 - Q_1)_B$	
$(Q_3 - Q_2)_A > (Q_3 - Q_2)_B$	
L.W.L of A >L.W.L B	

2 marks)

2. (a) What are the properties of a good central tendency measure? (02 *marks*)

Properties of a good measure of central tendency

i. It should be a unique measure
ii. It should be rigidly defined
iii. It should be easy to understand
iv. It should be easy to interpret
v. It should represent all observations of the data
vi. It should not be unduly affected by the extreme values

(02 marks)

(b) What are the basic characteristics of a distribution of data? Briefly explain the importance of each of them. (05 *marks*)

Basic characteristics of a distribution of data

i. Central tendency
ii. Variability (Dispersion)
iii. Skewness (asymmetry)
iv. Kurtosis (Flatness/peakedness)

Importance

- Understand the center of the population/ Obtain a representative single measure
- Understand the individual difference between observations of the population .
- Understand the symmetricity of the distribution
- Understand the curvature of the distribution

(05 marks)

(c) The following is a frequency distribution of diseases identified in a health camp attended by 300 individuals. The mean for the data is 2.15. Find the frequencies that correspond to the blanks in the frequency distribution.

No. of diseases	0	1	2	3	4	5	6
Frequency	55	--	80	--	30	20	10

(04 *marks*)

X	f		fx
0	55		0
1	x		x
2	80		160
3	y		3y
4	30		120
5	20		100
6	10		60
	300		

$\sum f = 300$ ----------1

$\bar{X} = \frac{\sum fX}{\sum f}$=2.15 ----------2

x + y + 195 = 300

x + y = 105 .....(1)

$\frac{(x + 160 + 3y + 120 + 100 + 60)}{300} = 2.15$

x + 3y + 440 = 645 ---------- (2)

x + y = 105 .....①

x + 3y = 205 .....②

② - ①

2y = 100

y = 50

x + 50 = 105

x - 50 = 105 - 50

x = 55

(04 marks)

(d) What are the limitations of standard deviation in comparing dispersion of different data sets? (03 *marks*)

- Standard deviation is inappropriate to compare the variations in data sets with different measuring units, because it is not independent from scale units.
- Even though the standard deviation evaluates the dispersion between the observation, it is not a relative comparison with the central tendency measure.

(03 marks)

(e) The following table shows the data recorded on lifetime of two new electronic products from a market survey.

Lifetime (No. of years)	Product A	Product B
0 - 2	8	6
2 - 4	14	8
4 - 6	12	16
6 - 8	8	11
8 -10	6	8
10 - 12	2	1

(i) Which product has higher mean lifetime? (02 *marks*)

Product A						Product B				
x	f	xi	Ui	fiUi	Fu^2	f	xi	Ui	fiUi	Fu^2
0 - 2	8	1	-3	-24	72	6	1	-2	-12	24
2 - 4	14	3	-2	-28	56	8	3	-1	-8	8
4 - 6	12	5	-1	-12	12	16	5	0	0	0
6 - 8	8	7	0	0	0	11	7	1	11	11
8 - 10	6	9	1	6	6	8	9	2	16	32
10 - 12	2	11	2	4	8	1	11	3	3	9
	Σf=50			**(-54)**	**154**	**Σf=50**			**10**	**84**

Assumed mean=A=7, C=2, $\boldsymbol{u = \frac{x-A}{c}}$

$$\bar{X} = A + \left(\frac{\Sigma fu}{\Sigma f}\right) \times C$$

$$= 7 + \left(\frac{-54}{50}\right) \times 2$$

$$= 7 - 2.16$$

$$\underline{\bar{X} = 4.84}$$

$$\bar{X} = A + \left(\frac{\Sigma fu}{\Sigma f}\right) \times C$$

$$= 5 + \left(\frac{10}{50}\right) \times 2$$

$$= 5 + 0.4$$

$$\underline{\bar{X} = 5.4}$$

The product B has a greater average lifetime

(02 marks)

(ii) Which product has less variation of lifetime? (04 *marks*)

Product A	Product B
$S^2 = \left[\frac{\sum fu^2}{\sum f} - \left(\frac{\sum fu}{\sum f}\right)^2\right] \times c^2$ $= \left[\frac{154}{50} - \left(\frac{54}{50}\right)^2\right] \times 4$ $= [3.08 - 1.1664] \times 4$ $= 1.9136 \text{ X } 4$ $= \sqrt{7.6544}$ $S = 2.77$ C.V $= \frac{S}{X}$ x 100 $= \frac{2.77}{4.84}$ x 100 $= 57.23\%$	$S^2 = \left[\frac{\sum fu^2}{\sum f} - \left(\frac{\sum fu}{\sum f}\right)^2\right] \times c^2$ $= \left[\frac{84}{50} - \left(\frac{10}{50}\right)^2\right] \times 4$ $= [1.68 - 0.04] \times 4$ $= 1.64 \text{ X } 4$ $= \sqrt{6.56}$ $S = 2.56$ C.V $= \frac{S}{X}$ x 100 $= \frac{2.56}{5.4}$ x 100 $= 47.41\%$

Product B less variability of lifetime

(04 marks)

3. (a) (i) Why is the Consumer Price Index (CPI) important in the economy of a country? (02 *marks*)

(1) To measure changes in general price level of consumer goods (inflation)

(2) To convert monetary income into real income (deflate)

(3) To compare the inflation with other countries

(4) To measure purchasing power of money

(02 marks)

(ii) List the indices used to measure the inflation in Sri Lanka. (01 *mark*)

Indices used to measure the inflation in Sri Lanka

1) National consumer price Index (NCPI)
2) Colombo Consumer Price Index (CCPI)
3) GDP deflator
4) Producer's Price Index
5) Whole sale price index

(01 mark)

(iii) Compare Colombo Consumer Price Index (CCPI) and National Consumer Price Index (NCPI). (04 *marks*)

Colombo Consumer Price Index - CCPI	National Consumer Price Index - NCPI
• Base Year 1952, 2002, 2006/07, 2013	Base Year 2013
• 393 consumer items under 12 major groups are considered.	407 consumer items under 12 major groups are considered
• Prices are collected from 14 price collection centers from urban areas of Colombo district	Prices are collected from 3 price collecting centers in each district in Sri Lanka
• Food items and non-food items have been weighed as 24% and 76% respectively	Food items and non-food items have been weighed as 44% and 56% respectively
• Laspeyres' formula is used	Modified Laspeyres' formula (Two stage) is used

(04 marks)

(b) Calculate Fisher's Ideal Index for price using the data given in the following table.

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	10	40	12	50
B	12	25	15	20
C	15	10	20	12
D	20	5	30	2

(03 *marks*)

Item	p_o	q_o	p_n	q_n	$p_o q_o$	$p_n q_o$	$p_n q_n$	$p_o q_n$
A	10	40	12	50	400	480	600	500
B	12	25	15	20	300	375	300	240
C	15	10	20	12	150	200	240	180
D	20	5	30	2	100	150	60	40
					950	1205	1200	960

$$FP_{n/o} = \sqrt{\frac{\sum P_n q_o}{\sum P_o q_o} \times \frac{\sum P_n q_n}{\sum P_o q_n}} \times 100$$

$$= \sqrt{\frac{1205}{950} \times \frac{1200}{960}} \times 100$$

$$= \underline{125.9}$$

(03 marks)

(c) (i) State the importance of studying the trend component of a time series. (02 *marks*)

Importance of studying the trend component of a time series.

1. to forecast past patterns/ trend .
2. to make production plans and sales plans.
3. to assess the need and feasibility of expanding the business.

and with similar variables in other firms.

5. To describe (historical) the long-term behavior of a variable

(02 marks)

(ii) Mention **two** methods to measure the trend and list **two** advantages of each method. (02 *marks*)

1. Free Hand Method

Advantages

1. Ability to obtain a rough idea regarding the trend of the variable at a glance.
2. Being Free from complicated calculations.
3. simplest and flexible method

2. Semi Average Method Advantages

1. A unique (identical) trend can be identified, because a simple mathematical technique is applied
2. This method is not subjective.

3. Least square Method

1. trend can be derived minimizing the errors
2. Highly appropriate for forecasting.
3. Unique trend can be identified

4 Moving Average Method

1. Useful in reducing random variation
2. Deriving a smooth trend Line which is more practical.
3. reduces the effect from extreme vales
4. A simple mathematical technique.

(02 marks)

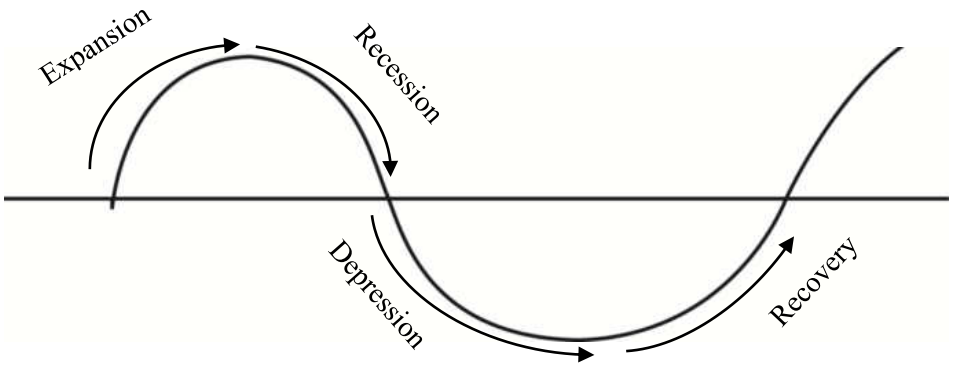
(iii) Explain cyclic variation considering the four main stages of a business cycle. (02 *marks*)

Cyclical variations refer to the long term oscillations about a trend line

A business cycle is a cycle of a fluctuation in a macro variable (i - e. GDP) around its long-term trend. It explains **expansion** and **contraction i**n economic activities that an economy experiences over time.

The four main stages of a business cycle is shown in the diagram below

Phases: Prosperity, recession, depression, and recovery



(02 marks)

(d) Quarterly sales (in thousand rupees) of a commodity are given in the following table.

Year	Quarterly Sales			
	I	II	III	IV
2007	46	40	39	38
2008	38	34	32	25
2009	51	28	28	33

Calculate seasonal index for each quarter using the simple average method and comment on the values. (04 *marks*)

	Quarterly Sales				
Year	I	II	III	IV	
2007	46	40	39	38	
2008	38	34	32	25	
2009	51	28	28	33	
total	135	102	99	96	
Quarterly average	45	34	33	32	36
Grand average	36	36	36	36	
Seasonal Index	45/36*100 .= **125**	34/368*100 **94.44444**	33/36*100 **91.66667**	32/36*100 **88.88889**	400

Comment:

Comment: quarterly sales decline over the seasons (Quarters).

In the first quarter, Seasonal index of 125 indicates that quarter I had 25% more sales than the seasonal average sales. However, the quarter II , III and IV had 6%, 8% 12% less than seasonal average sales, respectively.

(04 marks)

4. (a) Why do we need visual inspections of data as the first step in regression analysis? (01 *mark*)

A visual inspection of data is needed as the first step in regression analysis,

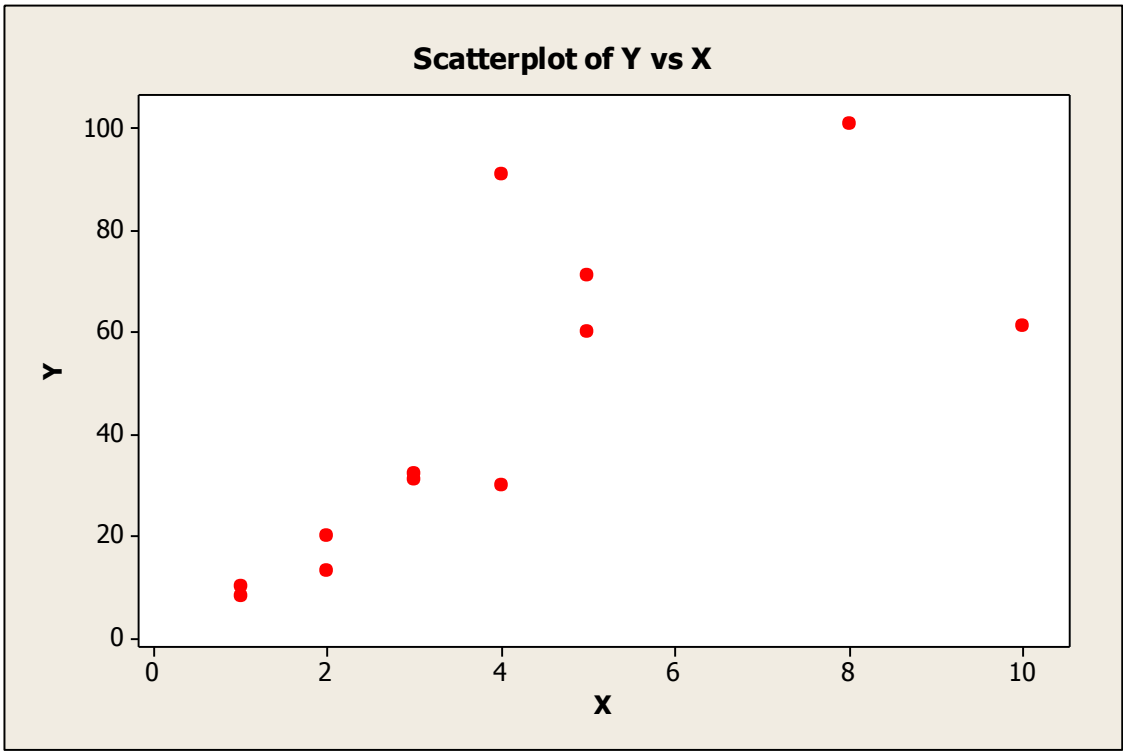
1. To identify whether there is a relationship
2. If there is a relationship to identify the relationship whether the relationship is deterministic or not deterministic
3. If it is not deterministic , to identify the nature of the relationship

(01 mark)

(b) In order to predict the performance of sales employees based on their sales experience, the manager of a large furniture store selected a random sample of 12 salesmen. Their recorded annual sales values (in ten thousand) and sales experience (in years) as follows.

Sales person	1	2	3	4	5	6	7	8	9	10	11	12
Sales experience (X)	2	2	1	1	5	5	3	4	4	3	8	10
Annual Sale (Y)	20	13	8	10	71	60	32	91	30	31	101	61

(i) Draw a scatter plot for the above data and comment on the possible relationship between two variables. (01 *mark*)



Comment:

- Scatter diagram shows that the relationship between X and Y is non - deterministic, it is stochastic
- There are various possible relationships that can be thought from the scatter diagram.

(01 mark)

(ii) The estimated equation for the regression line using least square method for the above data is as follows.

$$\hat{Y} = 8.98 + 2.46\,X$$

Interpret of the estimated slope coefficient and intercept. (02 *marks*)

Slope coefficient = 2.46

$\frac{\partial \hat{E}(Y|X)}{\partial X}$=2.46

When X increase by one unit (sales person gets one year experience) annual sales value on average would increase by 24600 rupees.

.

Intercept

$$\hat{\beta}_0 = 8.98$$

$$\hat{E}(Y|X = 0) = 8.98$$

when sales employees have no experience, the annual sales on average would be Rs. 89800.

(02 marks)

(iii) If R^2 for the above estimated regression line is 0.56, interpret the value. (01 *mark*)

56% of the total variability of annual sales is explained by sales employees' experience

(01 mark)

(iv) Comment on predicting the sales of a person with 15 years of experience by using estimated regression line. (01 *mark*)

It is not appropriate to do extrapolation based on the estimated sample regression line, since 15 years is out of the sample domain.

(01 mark)

(c) (i) Give the situations for which each of the methods Pearson's correlation coefficient and Spearman's rank correlation coefficient is appropriate. (01 *mark*)

	Pearson's correlation	Spearman's correlation
relationship	Linear relationship	Consistent/directional relationship
Data type	interval /ratio	Ordinal/ interval

(01 mark)

(ii) Following marks have been given by two judges **A** and **B** for 10 handcrafts in a competition.

Handcraft	1	2	3	4	5	6	7	8	9	10
Judge A	5	8	4	6	8	9	6	7	2	3
Judge B	7	6	5	9	5	7	8	4	2	1

Calculate Spearman's rank correlation coefficient and comment on the value. (03 *marks*)

Handcraft	**Judge A**	r_1	**Judge B**	r_2	**d**	d^2
1	5	7	7	3.5	3.5	12.25
2	8	2.5	6	5	2.5	6.25
3	4	8	5	6.5	1.5	2.25
4	6	5.5	9	1	4.5	20.25
5	8	2.5	5	6.5	4	16
6	9	1	7	3.5	2.5	6.25
7	6	5.5	8	2	3.5	12.25
8	7	4	4	8	4	16
9	2	10	2	9	1	1
10	3	9	1	10	1	1
						93.5

$$\gamma_k = 1 - \frac{6\,\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 93.5}{990}$$

$$= 1 - \frac{561}{990}$$

$$= 1 - 0.567$$

$$\boldsymbol{\gamma_k = 0.433}$$

The estimated correlation coefficient indicates that two judges' decisions are moderately/weakly associated.

(03 marks)

(d) (i) State the uses of R-chart in statistical quality control? (01 *mark*)

1. R chart is used for controlling quality dispersion (variability) of the product in a process.

2. To examine the validity of $\bar{X}$ chart

(03 marks)

(ii) List **four** types of pattern that indicate a process which is out of control. (02 *marks*)

- when one or more points falling outside the control limits
- when 3 or more successive points follow a certain pattern
- When majority of sample points fall close to any of the two control limits (upper control limit or lower control limit)
- A series of several sample points showing a trend
- A series of long run of consecutive sample points falling above or below the central line

(02 marks)

(e) The number of defective items in 12 samples of 50 items each are given below.

Sample number	1	2	3	4	5	6	7	8	9	10	11	12
Number of defective items	3	4	3	7	8	5	4	4	8	2	1	5

(i) Based on the P-chart, explain whether the process is statistically in control. (03 *marks*)

Sample number	**Number of defective items**	**Proportion of defectives**
1	3	0.06
2	4	0.08
3	3	0.06
4	7	0.14
5	8	0.16
6	5	0.1
7	4	0.08
8	4	0.08
9	8	0.16
10	2	0.04
11	1	0.02
12	5	0.01

$$\bar{P} = \frac{54}{12\ X\ 50} = 0.09 = CL$$

$$UCL = \bar{P} + 3\sqrt{\frac{\bar{P}\ (1-\bar{P}\ )}{n}}$$

$$= 0.09 + 3\sqrt{\frac{0.09\ X\ 0.91}{50}}$$

$$= 0.09 + 3\ \times 0.040$$

$$= 0.09 + 0.12$$

$$\underline{\underline{= 0.21}}$$

$$LCL = \bar{P} - 3\sqrt{\frac{\bar{P}(1-\bar{P})}{n}}$$

$$= 0.09 - 3\sqrt{\frac{0.09\ X\ 0.91}{50}}$$

$$= 0.09 - 0.12$$

$$LCL = -0.03 \equiv \simeq 0$$

P chart

The process is under control, since all the points fall within the control limits

(03 marks)

(ii) Explain the importance of Operating Characteristic (OC) curve. (02 *marks*)

1. ability to discriminate good lot and bad lot
2. ability to decide how successful of sampling plan
3. ability to know consumer's risk and producers.

(02 marks)

(iii) Suppose a company has received a large lot of items. It has been decided to use an acceptance sampling plan to examine whether the lot is acceptable. If the acceptance number is 3 for a random sample of 150 items, calculate the acceptance probability at each of the defective percentages 1% and 4%. (02 *marks*)

n = 150 C = 3

Defective proportion	**λ = np**	**Probability of accepting**
		0.22 31 + 0.3347 + = <u>0.9343</u>
0.04	6	0.0025 + 0.0149 + 0.0446 + 0.0892 = <u>0.1512</u>

(02 marks)

Part II

5. (a) (i) Distinguish probability of an event from conditional probability of an event. (02 *marks*)

Probability :

Probability is a measure of the likelihood of occurrence of an uncertain event

Conditional probability

Conditional probability is a measure of the probability of an occurring event given that another event has already occurred.

(02 marks)

(ii) Determine whether the two events A and B are exhaustive, if

$P(A) = \frac{4}{7}$, $P(B) = \frac{4}{5}$ and $P(A' \cap B) = \frac{3}{7}$. (02 *marks*)

$P(A) = \frac{4}{7}$, $P(B) = \frac{4}{5}$, $P(A' \cap B) = \frac{3}{7}$

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

$\frac{3}{7} = \frac{4}{5} - P(A \cap B)$

$P(A \cap B) = \frac{4}{5} - \frac{3}{7} = \frac{28-15}{35}$

$= \frac{13}{35}$

If A, B are exhaustive events, $P(A \cup B) = 1$

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

$P(A \cup B) = \frac{4}{7} + \frac{4}{5} - \frac{13}{35} = \frac{20+28-13}{35} = \frac{35}{35} = 1$

Therefore A, B are exhaustive events

(02 marks)

(iii) Let A and B are two events of a sample spaces S.

If $P(A \cap B') = \frac{21}{100}$, $P(A' \cap B') = \frac{13}{50}$ and $P(A \cap B)' = \frac{16}{25}$,

find $P(B)$, $P(B|A)$ and $P(A'|B)$. (03 *marks*)

$$P(A' \cap B') = P(A \cup B)' = 1 - P(A \cup B) = \frac{13}{50}$$

$$1 - P(A \cup B) = \frac{13}{50}$$

$$P(A \cup B) = 1 - \frac{13}{50} = \frac{37}{50}$$

$$P(A \cap B)' = \frac{16}{25}$$

$$P(A \cap B)' = 1 - P(A \cap B) = \frac{16}{25}$$

$$P(A \cap B) = 1 - \frac{16}{25} = \frac{9}{25}$$

$$P(A \cap B') = P(A) - P(A \cap B) = \frac{21}{100}$$

$$P(A) = \frac{21}{100} + \frac{9}{25} = \frac{21 + 36}{100} = \frac{57}{100}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{37}{50}$$

$$\frac{57}{100} + P(B) - \frac{9}{25} = \frac{37}{50}$$

$$* \quad P(B) = \frac{37}{50} + \frac{9}{25} - \frac{57}{100} = \frac{74+36-57}{100} = \frac{53}{100}$$

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

$$P(B|A) = \frac{9/25}{57/100} = \frac{36}{57}$$

$$\mathbf{P(A'|B) = \frac{P(A' \cap B)}{P(B)} = \frac{P(B) - P(A \cap B)}{P(B)} = \frac{\frac{53}{100} - \frac{9}{25}}{\frac{53}{100}} = \frac{\frac{17}{100}}{\frac{53}{100}} = \frac{17}{53}}$$

(03 marks)

(b) A Statistics Association in a certain country studied whether there is an association between students' status of interest in Statistics and their A/L subject stream using a random sample of 300 students. Summary of the data is given below.

Status of Interest in statistics	A/L subject stream		
	Arts	Commerce	Science
Interested	50	66	32
Not interested	74	37	41

	Arts	Com	Sc	
I	50	66	32	148
NI	74	37	41	152
	124	103	73	300

n = 300

(i) Calculate the probability that a randomly selected student is from commerce subject stream. (02 *marks*)

$\frac{103}{300} = 0.343$

(02 Marks)

(ii) Calculate the probability that a randomly selected student is interested in Statistics, given the student is from Arts subject stream. (02 *marks*)

$$P(I|A) = \frac{P(I \cap A)}{P(A)} = \frac{50/300}{124/300} = \frac{50}{124} = 0.403$$

(02 Marks)

(iii) Determine whether the A/L subject stream is independent from the status of interest in Statistics. (02 *marks*)

First Method

$$P(Arts) = P(A) = \frac{124}{300}$$

$$P(Interest\ to\ Statistics) = P(I) = \frac{148}{300}$$

$$P(A \cap I) = \frac{50}{300} \neq P(A)P(I) = \left(\frac{124}{300}\right)\left(\frac{148}{300}\right)$$

$\therefore$ *Not independence*

Second Method

If P(I|A)=P(I) A and I are independent event.

From (ii) P(I|A)=50/124. From the data P(I) = 148/300. Since P(I|A)≠P(I), A and I are not independent.

Subject stream is not independent from the status of Interest in statistics

(02 Marks)

(c) A multinational company is planning to start a subsidiary in a certain country. The management of the company realizes that the success of the new subsidiary depends on the political situation in the target country. The management estimates that the probability of success of the subsidiary is 0.64 if the prevailing political situation is favourable, 0.28 if the political situation is neutral and 0.12 if the political situation is unfavourable during the year. The management further believes that the probabilities of favourable and unfavourable political situations are 0.54 and 0.21 respectively.

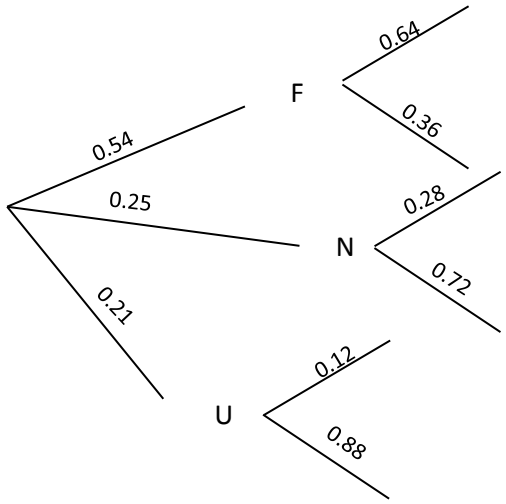
S= Success,

NS = not success

F - Favourable

N - Neutral

U - Unfavorable -



$P(S|F) = 0.64$

$P(S|U) = 0.28$

$P(S|F) = 0.12$

$P(F) = 0.54$

$P(N) = 0.25$

$P(U) = 0.21$

(i) What would be the probability of success of the subsidiary? (02 *marks*)

$$P(S) = P(S|F)P(F) + P(S|N)P(N) + P(S\}U)P(U)$$

$$P(S) = (0.64)(0.54) + (0.28)(0.25) + (0.12)(0.21) = 0.4408$$

(2 Marks)

(ii) What would be the probability of prevailing political situation is unfavourable given that the subsidiary is successful? (02 *marks*)

$$P(U|S) = \frac{P(S \cap U)}{P(S)} = \frac{0.0252}{0.4408} = 0.0572$$

(2 Marks)

(d) A device consists of three components A, B and C, and it is necessary to have at least two components non-defective in order for it to work. If the probabilities of these components A, B and C being defectives are 0.1, 0.15 and 0.2 respectively, find the probability that this device will operate when necessary assuming the probability of each component becoming defective is independent from that of others. (03 *marks*)

A - A being non-defective

B - B being non-defective

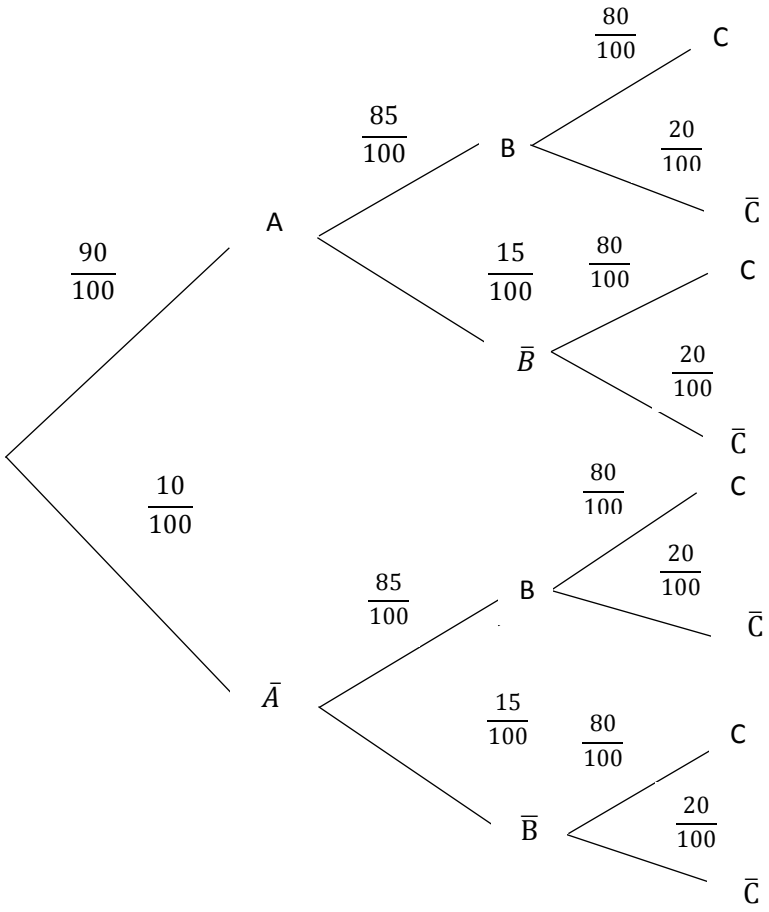
C - C being non-defective

$P(A') = 0.1, \quad P(A) = 0.9$

$P(B') = 0.15, \quad P(B) = 0.85$

$P(C') = 0.2, \quad P(C) = 0.8$

A – Non - defective A' - defective



P (Device being Operative)

= P(A ∩ B ∩ C') + P (A ∩ B' ∩ C) + P (A' ∩ B ∩ C) + P (A ∩ B ∩ C)

= P (A) P(B) P (C') + P(A) P (B') P(C) + P(A') P(B) P(C) + P(A) P(B) P(C)

= 0.9 x 0.85 x 0.2 + 0.9 x 0.15 x 0.8 + 0.1 x 0.85 x 0.8 + 0.9 x 0.85 x 0.8

= 0.153 + 0.108 + 0.068 + 0.612

= 0.941

3 Marks)

6. (a) (i) Distinguish a binomial experiment from a binomial random variable. (02 *marks*)

Binomial Experiment

1. It is a random experiment
2. It consists of n identical trials (Bernoulli trials)
3. Each trial results in one of the two mutually disjoint outcomes - called success or failure.
4. Number of trials is finite/fixed)
5. The probability of a success (p) or failure (1-p) remains same (constant) in each trial.
6. The trials are independent , hence the outcomes of the trials are independent

Binomial Random variable

A random variable (X)which follows binomial probability distribution is called a binomial random variable.

(02 Marks)

(ii) State characteristics of the binomial probability distribution. (02 *marks*)

Characteristics of the Binomial probability distribution

Let X be number of success out of n trials

i. Binomial distribution has two parameters (n, p)
ii. The mean of the binomial distribution is $E(x) = np$
iii. The variance of the distribution, $V(x) = npq$ where $q = 1 - p$.
iv. If $P(x) = 0.5$, X has symmetric distribution
v. Binomial distribution tends to normal distribution as n increases
vi. If $n > 50$ and $p < 0.1$, Binomial RV $X \sim Po(np)$
vii. If $n > 10$ and p close to 0.5 the Bin RV $X \sim N(np, npq)$
viii. If $n > 30$ the Bin RV $X \sim N(np, npq)$

(02 Marks)

(b) Suppose the probability of success of a mobile salesman at a household is 0.2.

X = Number of successes in visiting n households

$n = 5$ = Number of households

Probability of success = 0.2 , q=0.8,

(i) What is the probability that the salesman will be successful at two households out of five households? (01 *mark*)

$$P(X=2) = {}^{5}c_{2}(0.2)^{2}(0.8)^{5}$$

$$= \underline{0.2048}$$

(01 mark)

(ii) What is the probability that the salesman will be successful in at least two households out of five households? (02 *marks*)

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

$$= 1 - [P(X=0) + P(X=01)]$$

$$= 1 - [0.3277 + 0.4096]$$

$$= 1 - 0.7373$$

$$= \underline{0.2627}$$

(02 marks)

(iii) What are the assumptions that you take to calculate the above probabilities? (01 *mark*)

Assumptions

2) Probability of success each household is same

3) Tested households are independent from each other

4) Outcome at each household is success or failure

(01 mark)

(c) (i) State **three** examples for a Poisson random variable. (02 *marks*)

Examples for a Poisson Ramona Variable

(1) # of accidents in particular junction in an hour.
(2) # of calls come to a telephone exchange 10 miniature
(3) # of customers arrive to a bank within a period of 1 hour
(4) # of errors in a page
(5) # of defects within 1 meter length of cloth

(02 marks)

(ii) What conditions are required for Binomial distribution to be approximated by the Poission distribution? (01 *mark*)

Condition **Bin (n, p) ⟶ Poisson(np)**

i) **N > 50**

ii) **P small** (P < 0.1)

X ~ Po (np)

(01 mark)

(iii) Let X~Bin(100, 0.05). Find $P(X=3)$ using the normal and Poisson approximation.
[Assume $\sqrt{4.75} \approx 2$] (03 *marks*)

Using Normal approximation

$X \sim Bin(100,\ 0.05)$

Condition

(1) n > 10 and P≈ 0.5
or
(2) n > 30
or
(3) $np \geq 5$ *or* $nq \geq 5$

np = 100(0.05)=5
npq = 100 (0.05)(0.95)=0.475

$\sqrt{npq} = \sqrt{4.75} = 2.179 \cong 2$

$X \sim N(np, npq)$

$X \sim (5, 4.75)$

P (X =3) = P (2.5< X <3.5)

$= P[\frac{2.5-5}{\sqrt{4.75}} < Z < \frac{3.5-5}{\sqrt{4.75}}]$

$= P[\frac{-2.5}{\sqrt{4.75}} < Z < \frac{-1.5}{\sqrt{4.75}}]$

$= P[\frac{-2.5}{2} < Z < \frac{-1.5}{2}]$

$= P[-1.25 < Z < -0.75]$

= 0.227 – 0.106

= 0.121

Using Poisson approximation

When n > 50, P <0.1, we can use Poisson approximation

$\lambda = np$ = 100(0.05)= 5

X~$Po(5)$

P(X =*x*) = $\frac{e^{-5}\,5^{x}}{x\,!}$

P(X =*3*) = $\frac{e^{-5}\,5^{3}}{3!}$

= 0.1404

Since n ≈ large with P ≈ Small
n> $50, P < 0.1, This\ is\ good\ \ approximation$

(03 marks)

(d) (i) State the descriptive methods for examining normality of a data set. (02 *marks*)

Descriptive methods of testing normality

1. Empirical rule

   $\bar{X} \pm 1SD$ 68 %

   $\bar{X} \pm 2SD$ 95 %

   $\bar{X} \pm 3SD$ 99 %

2. Histogram

3. Stem and leaf plot

4. Box Plot

5. normal probability plot

6. $\frac{IQR}{SD} \cong 1.35$

7. kurtosis and skewness

(02 marks)

(ii) The mean and standard deviation of marks of an examination are 72 and 9, respectively. If the top 10 % of students are to receive A grades, determine the cutoff mark for the grade assuming marks are distributed normally. (02 *marks*)

$X \sim N\ (72, 9^2)$

$P\ (z \geq a) \quad = 0.1$

$a = 1.28$

$\frac{X-72}{9} \quad = 1.28$

$X = 72 + 1.28\ (9)$

$= \quad 83.5$ cutoff marks

(02 marks)

(iii) Using the above distribution of marks, find k if $P\ [|X-72|>k] = 0.05$. (02 *marks*)

$P(|X - 72| > k) = 0.05$

$P\left(\left|\frac{X-72}{9}\right| > \frac{k}{9}\right) = 0.05$

$P(|Z| > a) = 0.05$

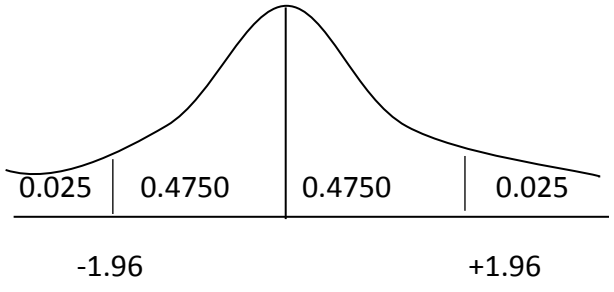
$P(|Z| > 1.96) = 0.05$

$P(Z > 1.96) + P(Z < -1.96) = 0.05$

Where $\frac{k}{9}$=a=1.96

k=1.96(9)

K=17.64



(02 marks)

7. (a) A random sample consisting units X_1, X_2, X_3, X_4 and X_5 was drawn from a normally distributed population with unknown mean, μ and variance, σ^2. Suppose following estimators are suggested to estimate the population mean μ.

$$T_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}, \quad T_2 = \frac{X_1 + X_2}{2} + X_3, \quad T_3 = \frac{2X_1 + X_2 + kX_3}{3}$$

$$T_1 = \frac{X_1+X_2+X_3+X_4+X_5}{5}$$

$$T_2 = \frac{X_1+X_2}{2} + X_3$$

$$T_3 = \frac{2X_1+X_2+kX_3}{3}$$

(i) Find whether T_1 and T_2 are unbiased estimators of μ. (02 *marks*)

$$T_1 = \frac{X_1+X_2+X_3+X_4+X_5}{5}$$

$$E(T_1) = \frac{E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5)}{5}$$

$$E(T_1) = \frac{\mu + \mu + \mu + \mu + \mu}{5}$$

$$E(T_1) = \frac{5\mu}{5} = \mu$$

T_1 is unbiased

$$T_2 = \frac{X_1 + X_2}{2} + X_3$$

$$E(T_2) = \frac{E(X_1) + E(X_2)}{2} + E(X_3)$$

$$E(T_2) = \frac{\mu + \mu}{2} + \mu$$

$$E(T_2) = \frac{4\mu}{2} = 2\mu$$ ----Biased

02 marks)

(ii) Find the value of k such that T_3 is an unbiased estimator of μ. (01 *mark*)

$$T_3 = \frac{2X_1+X_2+kX_3}{3}$$

$$E(T_3) = \frac{2E(X_1) + E(X_{2)} + kE(X_{3)}}{3}$$

$$E(T_3) = \frac{2\mu + \mu + k\mu}{3} = \mu$$

$$2\mu + \mu + k\mu = 3\mu$$

$$k\mu = 0 \quad \text{since} \quad \mu \neq 0$$

(01 mark)

(iii) Consider an estimator T given by $T = l_1 T_4 + l_2 T_5$, where l_1, l_2 are constants and T_4 and T_5 are unbiased estimators of μ. Show that $l_1 + l_2 = 1$ when T is an unbiased estimator of μ. (02 *marks*)

$T = l_1T_4 + l_2T_5$, T_4, T_5 are unbiased

$$E(T) = l_1E(T_4) + l_2E(T_5)$$

$$E(T) = l_1\mu + l_2\mu = \mu$$

$$E(T) = (l_1 + l_2)\mu = \mu$$

$$l_1 + l_2 = 1$$

(02 marks)

(b) In an examination a student received 62 and 48 marks for two subjects Statistics and Economics respectively. The mean and variance for the marks of Statistics are 50 and 36 respectively, whereas mean and variance for the marks of Economics are 42 and 4 respectively. Using Z score, compare performance of the student for the two subjects. (02 *marks*)

Statistics $\mu = 50 \quad \sigma^2 = 36$

Economics $\mu = 42 \quad \sigma^2 = 4$

subject	marks	z score	z
stat	62	$\frac{62-50}{6}$	$\frac{12}{6} = 2$
Econ	48	$\frac{48-42}{2}$	$\frac{6}{2} = 3$

Since Z score for Economics is higher than that for Statistics,

the student has performed better in economics (z = 3), than in Statistics (z = 2).

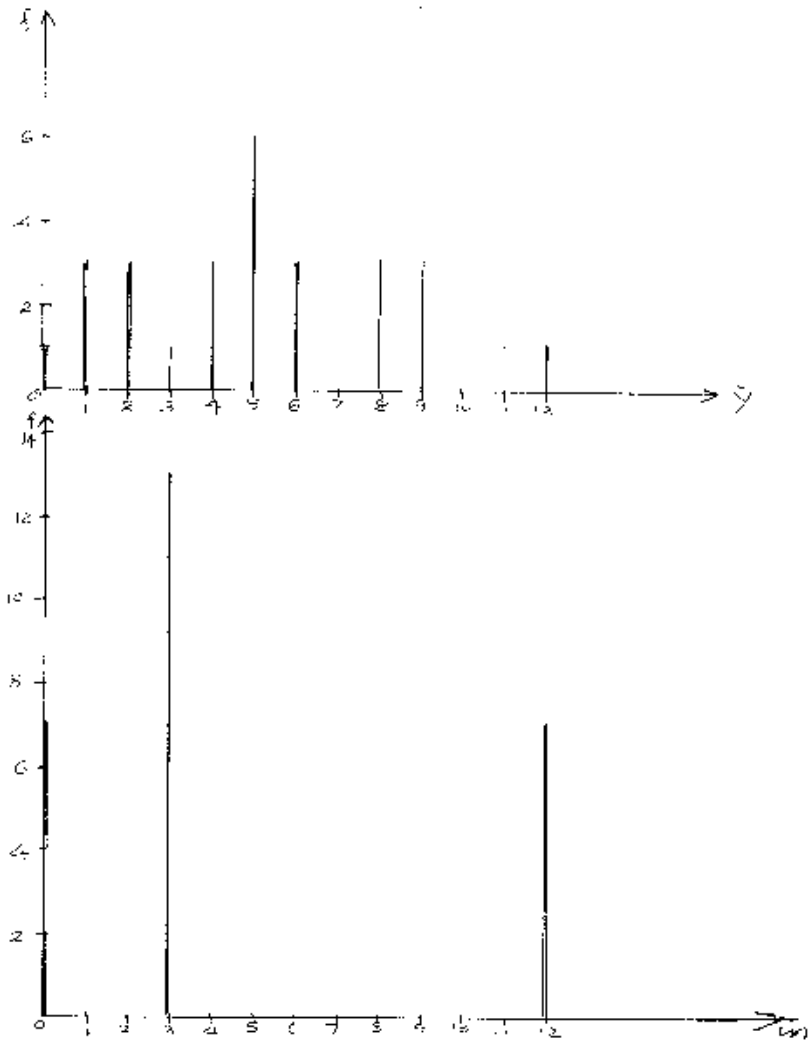
(02 marks)

(c) For a population consisting the measurements $y_1 = 0$, $y_2 = 3$ and $y_3 = 12$ with equal probability of occurrence; the frequency (f) distributions of sample mean, $\bar{y}$ and sample median, m for the sampling distribution of size 3 with replacement are given below.

$\bar{y}$	0	1	2	3	4	5	6	8	9	12
f	1	3	3	1	3	6	3	3	3	1

m	0	3	12
f	7	13	7

(i) Plot the distributions of $\bar{y}$ and m and compare them. (02 *marks*)



Both distributions are positively skewed. But median distribution has a higher skewness

(02 marks)

(ii) State the uses of these distributions in inferential analysis. (04 *marks*)

. Sampling distributions of sample mean ($\bar{y}$) and sample median (m) provides information about the corresponding parameters in the inferential analysis (Confidence interval, hypothesis testing),

1. Probability distributions of sample mean is used for the inferential analysis of population mean.
2. Probability distribution of sample median is used for the inferential analysis of population median

(04 marks)

(iii) The expected values and the variances of the above estimators are given below.

Estimators	Expected value	Variance
$\bar{y}$	5.00	8.66
m	4.56	20.91

Comment on the two estimators. (02 *marks*)

Population values

Population mean, μ= (0+3+12)/3=5

E(y)=5
Var(y)=128

Sample statistics

Given E($\bar{y}$) = 5

Since μ= E($\bar{y}$), the ybar is an unbiased estimator.

E(med) = 4.56

V($\bar{y}$) = 8.67 SD = 2.94
V(m) = 20.91 SD = 4.57

ii. $\bar{y}$ is unbiased estimator of mean

iii. $\bar{y}$ has minimum variance

$Var(\bar{y}) < Var(m)$

8.67 < 20.91

(02 marks)

(d) Show that the $\bar{y}$ from simple random sampling without replacement is more precise than $\bar{y}$ from simple random sampling with replacement. (02 *marks*)

with replacement	without replacement
$E(x) = \mu$	$E(x) = \mu$
$Var(x) = \frac{\sigma 2}{n}$	$Var(x) = \frac{\sigma 2}{n}\left(\frac{N-n}{N-1}\right)$

$\frac{\sigma 2}{n} > \frac{\sigma 2}{n}\frac{N-n}{N-1}$

Since $n > 1$, $N - 1 > N-n$

$\frac{N-n}{N-1} < 1 \Rightarrow \frac{\sigma 2}{n}\frac{N-n}{N-1}$

Therefore sampling without replacement is more precise than sampling with replacement

(02 marks)

(e) Sales profit from a specific product was studied using a simple random sample of 35 sales centers. From this sample, a sub sample of size 20 was also studied. Let the means from the sub-sample of size 20 and the balance 15 are $\bar{y}_1$ and $\bar{y}_2$ respectively. Moreover, let the mean from the total sample is $\bar{y}$, assume $\sigma^2 = 200$.

(i) Find $Var(\bar{y}_1 - \bar{y}_2)$. (02 *marks*)

From the sampling theory

$$V(\bar{y}) = \frac{\sigma 2}{n}$$

Total sample size $= n = 35$

Sample mean $= \bar{y}$

Sub sample 1 $= n_1 = 20$

Sub sample mean $= \bar{y}_1$

Sub sample 2 size $= n_2 = 15$

Sub sample mean $= \bar{y}_2$

(i) $Var(\bar{y}_1 - \bar{y}_2) = Var(\bar{y}_1) + Var(\bar{y}_2)$

because sub samples are independent

$$Var(\bar{y}_1 - \bar{y}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$$

$$= \sigma^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$$

$$200\left[\frac{1}{20} + \frac{1}{15}\right]$$

$$= 23.33$$

(02 marks)

(ii) What extra information would you need to consider in the calculation of $Var(\bar{y}_1 - \bar{y})$, compared to that of part (i)? (01 *mark*)

Sub sample mean $\bar{y}_1$ is not independent from the total sample mean $\bar{y}$

$$Var(\bar{y}_1 - \bar{y}) = Var(\bar{y}_1) + Var(\bar{y}) - 2cov(\bar{y}_1, \bar{y})$$

To find $Var(\bar{y}_1 - \bar{y})$

Non independence of $\bar{y}_1$ and $\bar{y}$

(01 mark)

8. (a) Giving reasons, state whether each of the following statements is true or false.

(i) Probability of type I error is calculated using the distribution of test statistic under the null hypothesis. (02 *marks*)

1. True

$P(Type\ I\ error) = P(Reject\ H_0 | H_0 is\ true)$, thus Probability of Type I error is calculated using H0 distribution.

(02 marks)

(ii) Precision of sample mean depends on the sample size. (02 *marks*)

2. True

Precision of sample mean is measured by variance of the sample mean $Var(\bar{X}) = \frac{\sigma^2}{n}$, therefore the precision depends on sample size.

(02 marks)

(iii) By increasing the sample size, power of the test can be increased without changing the probability of type I error. (02 *marks*)

3. True

V) in $SE(\bar{X})$

terms of absolute value will be smaller . thus P(Type II) become smaller. So that power is increasing.

(02 marks)

(b) A cross classification of a random sample of 200 people by status of vaccination (X) and status of occurrence of Covid-19 (Y) is given below.

X \ Y	Occured	Not occured
Vaccinated	10	150
Not Vaccinated	10	30

Suppose it is intended to study the impact of vaccination on occurrence of Covid-19 using the above data.

(i) State the hypotheses H_0 and H_1. (01 *mark*)

i. H_o: Vaccination has no impact on occurrence the disease

H_1: vaccination has impact on occurrence the disease

(01 mark)

(ii) State the test statistic and its distribution. (01 *mark*)

$$TS = \chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad \text{or} \quad Z = \frac{\hat{P}_1 - \hat{P}_2}{SE\ (\hat{P}_1 - \hat{P}_2)} = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$TS \sim \chi^2_{(r-1)\,(c-1)} \quad \text{or} \quad TS \sim Z \sim n\,(0,1)$$

$$TS \sim \chi^2_1$$

(01 mark)

(iii) Test whether the vaccination has an impact on the status of occurrence of Covid-19 at 5% significant level. (03 *marks*)

	Yes	No	
Vacc	10	150	160
No Vaccin	10	30	40
	20	180	200

Y / X	**Occurred**	**Not Occurred**	**Total**
Vaccinated	10 $\frac{160x20}{200}$ = 16 (16)	150 $\frac{160x180}{200}$ = 144 (144)	160
Not vaccinated	10 $\frac{40x20}{200}$ = 4 (4)	30 $\frac{40x180}{200}$ = 36 (36)	40
Total	20	180	

O	E	(O-E)	$(O-E)^2$	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
10	16	-6	36	2.25
150	144	6	36	0.25
10	4	6	36	9
30	36	6	36	1
				12.5

$Test\ Statistic = \chi^2 = 12.5$

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

= (2 - 1) (2 - 1)

= 1

CV x^2 = $x^2_{1,0.05}$ = 3.84

Decision rule:

As TS>CV, we reject Ho : No association at 5% significant level.

Sample evidence shows that vaccination has an impact on the status of occurrence of Covid-19

at $\alpha = 0.05$

Alternative answer

H_o: $P_1 = P_2$

H_1: $P_1 < P_2$

Test statistic $\not{Z} = \frac{\hat{P}_1 - \hat{P}_2}{SE\ (\hat{P}_1 - \hat{P}_2)} = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

$\hat{P}_1 = \frac{10}{160}$, $\hat{P}_2 = \frac{10}{40}$, $\hat{P} = \frac{10+10}{160+40} = \frac{20}{200} = 0.1$, $\hat{q} = 1 - \hat{p} = 0.9$

$n_1 = 160$

$n_2 = 40$

$\hat{P}_1 - \hat{P}_2 = \frac{10}{160} - \frac{10}{40} = \frac{10-40}{160} = \frac{-3}{16} = 0.1875$

$\not{Z} = \frac{-0.1875}{\sqrt{(0.1)(0.9)\left(\frac{1}{160}+\frac{1}{40}\right)}}$

$= \frac{-0.1875}{\sqrt{(0.1)(0.9)(0.03125)}} = \frac{-1.1875}{0.05303} = -3.5357$

$\not{Z}_{0.05} = 1.65$

Since $|\not{Z}| > \not{Z}_\alpha$, $H_0\ is\ rejected\ at\ 5\%\ level. Thus\ P_1 < P_2$

∴ There is an impact of vaccination

(03 marks)

(c) The Managing Director of a firm wanted to determine whether three different training programs have different effects on employees' proficiency level. Accordingly, he selected 14 employees and assigned 4 employees to the program I, and 5 each for the other two programs II and III randomly. Upon completion of the training, each employee was given a test to assess his/her proficiency. The test scores were analysed and the output from the analysis is given below.

Source	df	Sum of Square	Mean Square
Between samples	2	65.71	32.85
Within samples	.....	.....	.....
Total	.....	251.71	

(i) Fill the blanks in the above table. (02 *marks*)

Source	**DF**	**SS**	**MS**
Between sample	2	65.71	32.85
within sample	11	186	16.909
Total	13	251.71	

(02 marks)

(ii) State the hypotheses H_0 and H_1 for the test. (01 *mark*)

$H_O : \mu_1 = \mu_2 = \mu_3$

H_1 : at least two of the means are different

(01 mark)

(iii) State the assumptions that are required to perform this analysis. (02 *marks*)

Assumptions

1) Population is normally distributed, Population Variances are equal $(Y_i \sim N(\mu_i, \sigma^2))$

2) Error is normally distributed, error variances are equal $(U_i \sim N(0, \sigma^2))$

3) Each sample in drown ***independently***, ***randomly*** from a ***normal population***

(02 marks)

(iv) Compute the test statistic and state its probability distribution. (02 *marks*)

TS = $\frac{32.85}{16.91} = 1.94$

Its probability distribution is $TS \sim F_{(k-1)(N-k)} = F_{2,11}$

F distribution with df 2, 11

(02 marks)

(v) Perform the test at 5% significant level to determine whether mean proficiency levels of three training programs are the same. (02 *marks*)

V. TS=1.94,

CV = $F_{2,11,0.05}$ = 3.98

As TS<CV , we do not reject H0 at 5% significance level

Not enough evidence to say that the training programs affect the employee's means proficiency levels.

(02 marks)

****

Appendix

(3) d

Year	I	II	III	IV	yearly average
2017	46	40	39	38	40 . 75
2018	38	34	32	25	32 . 25
2019	51	28	28	33	35 . 00

2017	$\frac{46}{40.75} \times 100$	$\frac{40}{40.75} \times 100$	$\frac{39}{40.75} \times 100$	$\frac{38}{40.75} \times 100$
	112.9	98.2	95.7	93.3
2018	$\frac{38}{32.25} \times 100$	$\frac{34}{32.25} \times 100$	$\frac{32}{32.25} \times 100$	$\frac{25}{32.25} \times 100$
	116.0	105.4	99.2	77.5
2019	$\frac{51}{35} \times 100$	$\frac{28}{35} \times 100$	$\frac{28}{35} \times 100$	$\frac{33}{35} \times 100$
	145.7	80.0	80.0	94.3
Total	374.6	283.6	274.9	265.1
average	124.87	94.53	91.63	88.37
Seasonal Index	125	95	92	88