

Q.6

Attempt any two:

- i. Calculate the rank coefficient of correlation from the following table:

X	Y
78	125
89	137
97	156
69	112
59	107
79	136
68	123
57	108

- ii. Calculate the mode from the following distribution:

Class	Frequency
10-20	2
20-30	17
30-40	7
40-50	18
50-60	6
60-70	18
70-80	4
80-90	8

- iii. The expenditure of 1000 families is given as under-

x	40-59	60-79	80-99	100-119	120-139
f	50	?	500	?	50

Where x denotes the expenditure in Rs, and f denotes number of families.

The median and mean for the distribution are both Rs. 87.50.

Calculate the missing frequencies.

Total No. of Questions: 6

Total No. of Printed Pages:4

Enrollment No.....



Faculty of Management Studies

End Sem Examination Dec 2024

MS5CO40 Business Statistics

Programme: MBA

Branch/Specialisation: Business

Analytics

Duration: 3 Hrs.

Maximum Marks: 60

Note: All questions are compulsory. Internal choices, if any, are indicated. Answers of Q.1 (MCQs) should be written in full instead of only a, b, c or d. Assume suitable data if necessary. Notations and symbols have their usual meaning.

- | | Marks | BL | PO | CO | PSO |
|---|-------|----|----|----|-----|
| Q.1 i. If the events A and B are mutually disjoint then $P(A \cap B) = \underline{\hspace{1cm}}$.
(a) 3 (b) 2 (c) 1 (d) 0 | 1 | 1 | 1 | 1 | |
| ii. How many possible two-digit numbers can be formed by using the digits 3,5 and 7 repetition of digits is allowed?
(a) 10 (b) 9 (c) 7 (d) 2 | 1 | 3 | 1 | 1 | |
| iii. What best describes a Z score?
(a) It is the average of all raw scores in a normal distribution.
(b) It is the measure of dispersion in a distribution of scores
(c) It is the position of a score relative to the mean.
(d) It is the frequency of a score in standardized units. | 1 | 1 | 1 | 2 | |
| iv. The purpose of pilot study is-
(a) To identify the problem before actual study
(b) To identify the problem after actual study
(c) To identify the problem at the time of actual study
(d) None of these | 1 | 1 | 1 | 2 | |
| v. The hypothesis which is accepted when the null hypothesis is rejected is called ____ hypothesis.
(a) Zero (b) Alternative
(c) One (d) None of these | 1 | 1 | 1 | 3 | |
| vi. When calculating ANOVA, F data will always fall within the range?
(a) (0, ∞) (b) (0, 1)
(c) (- ∞ , 0) (d) None of these | | 1 | 1 | 3 | |

P.T.O.

[2]

- vii. For Quartile deviation the coefficient of dispersion is defined by____, where the symbols have their usual meaning. **1** 1 1 4

(a) $(Q3 - Q1) / (Q3 + Q1)$ (b) $Q3 - Q1$
(c) $Q3 + Q1 / Q3 - Q1$ (d) None of these

- viii. Two data set of sizes 9 and 6 have standard deviation 3 and 4 respectively and arithmetic means 3 respectively. The standard deviation of combined data of size 15 is- **1** 3 1 4

(a) $\sqrt{\frac{177}{15}}$ (b) $\sqrt{\frac{176}{15}}$
(c) $\sqrt{\frac{178}{15}}$ (d) None of these

- ix. The Karl Pearson's coefficient of correlation lies between- **1** 1 1 5
(a) $-\infty$ to $+\infty$ (b) -1 to +1
(c) 0 to 1 (d) None of these

- x. If the mean of 6,4,7, p and 10 is 8, Find the value of p? **1** 3 1 5
(a) 15 (b) 13
(c) 10 (d) None of these

Q.2 Attempt any two:

- i. A letter is known to have come either from TATANAGAR or from "CALCUTTA". On the envelope just two consecutive letters TA are visible. What is the probability that letter came from CALCUTTA? **5** 3 1 1
- ii. Out of 800 families with 4 children each, how many families would be expected to have- **5** 3 1 1
(a) 2 boys and 2 girls
(b) At least one boy.
Assume equal probability for boys and girls.
- iii. The lifetime of a certain kind of battery has a mean of 300 hours and a standard deviation of 35 hours. Assuming that the distribution of lifetimes, which are measured to the nearest hour is normal, find the percentage of batteries, which have lifetime of more than 370 hours. [P ($0 < z < 2$) = 0.4772]. **5** 3 1 1

Q.3 Attempt any two:

- i. Write uses of sampling in marketing research and campaigns. **5** 2 1 2

[3]

- ii. The record of weights of the male population follows the normal distribution its mean and standard deviation are 70 Kg and 15 Kg respectively. If a researcher considers the records of 50 males then using central limit theorem what would be the mean and standard deviation of the chosen sample? **5** 3 1 2
- iii. Explain in brief excel demo type of sampling method. **5** 2 1 2

Q.4 Attempt any two:

- i. The following data is given: **5** 3 1 3

Types of animals	Number of animals	Average domestic animals	Standard deviation
Dogs	5	12	2
Cats	5	16	1
Hamsters	5	20	4

Calculate the ANOVA coefficient.

- ii. To study the performance of three detergents and three different water temperatures, the following whiteness readings were obtained with Specially designed equipment. **5** 3 1 3

Water temperature	Detergent A	Detergent B	Detergent C
Cold water	57	55	67
Warm water	49	52	68
Hot water	59	46	58

Perform a two-way ANOVA using 5% level of significance.

- iii. How to find the critical region making the decision using P- value approach. **5** 3 1 3

Q.5 Attempt any two:

- i. Explain types of dispersion measure in detail. **5** 2 1 4
- ii. Find the standard deviation and coefficient of variation (C.V.) of the given series- **5** 3 1 4
13, 15, 18, 19, 20
- iii. Find the quartiles and quartile deviation of the following data- **5** 3 1 4
17,2,7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28

Marking Scheme

(MS5CO40) Business Statistics (T)

Programme:- MBA

Question 1

i. If the events A and B are mutually disjoint, then $P(A \cap B) =$

* Answer: (d) 0

ii. How many possible two-digit numbers can be formed by using the digits 3, 5, and 7 (repetition of digits is allowed)?

* Answer: (a) 9.

iii. What best describes a Z-score?

* Answer: (c) It is the position of a score relative to the mean..

iv. The purpose of a pilot study is-

* Answer: (a) To identify the problem before actual study.

v. The hypothesis which is accepted when the null hypothesis is rejected is called

* Answer: (b) Alternative hypothesis

vi. When calculating ANOVA, F data will always fall within the range?

* Answer: (a) $(0, \infty)$

vii. For Quartile deviation, the coefficient of dispersion is defined by:

The correct answer is (a) $(Q3 - Q1) / (Q3 + Q1)$.

viii. Two data sets of sizes 9 and 6 have standard deviations 3 and 4 respectively and arithmetic means 3 respectively. The standard deviation of the combined data of size 15 is:

The correct answer is (b) $\sqrt{177/15}$.

ix. The Pearson's coefficient of correlation lies between:

The correct answer is (b) -1 to +1.

x. If the mean of 6, 4, 7, p, and 10 is 8, find the value of p:

The correct answer is (b) 13.

Question 2:

i. A letter is known to have come either from TATANAGAR or from CALCUTTA. On the envelope just two consecutive letters TA are visible. What is the probability that the letter came from CALCUTTA?

Answer:- To solve this, we need to find the total number of possible arrangements of two consecutive letters from TATANAGAR and CALCUTTA and then calculate the probability based on the number of occurrences of "TA" in each case.

Let's analyze:

* TATANAGAR has three occurrences of "TA".

* CALCUTTA has no occurrences of "TA".

Therefore, the probability that the letter came from CALCUTTA is 0 (zero).

ii. Out of 800 families with 4 children each, how many families would be expected to have:

(a) 2 boys and 2 girls

Let's use the binomial distribution to calculate this. The probability of having a boy or a girl is 0.5.

The probability of having 2 boys and 2 girls is given by:

$$P(2 \text{ boys and } 2 \text{ girls}) = {}^4C_2 * (0.5)^2 * (0.5)^2 = 6 * (1/16) = 3/8$$

$$\text{Number of families expected to have 2 boys and 2 girls} = 800 * (3/8) = 300$$

(b) At least one boy.

The probability of having at least one boy is equal to 1 minus the probability of having no boys.

$$\text{Probability of having no boys} = (1/2)^4 = 1/16$$

$$\begin{aligned} \text{Number of families expected to have at least one boy} &= 800 * (1 - 1/16) = 800 * (15/16) \\ &= 750 \end{aligned}$$

iii. The lifetime of a certain kind of battery has a mean of 300 hours and 35 hours. Assuming that the distribution of lifetimes is normal, find the percentage of batteries that have a lifetime of more than 370 hours. Given $P(0 < z < 2) = 0.4772$.

Let's use the standard normal distribution to calculate this.

First, we need to standardize the value of 370 hours using the formula:

$$z = (x - \text{mean}) / \text{standard deviation}$$

where:

* x is the value we want to standardize (370 hours)

* mean is the mean of the distribution (300 hours)

the distribution (35 hours)

Plugging in the values:

$$z = (370 - 300) / 35 = 2$$

We know that $P(0 < z < 2) = 0.4772$. Since the standard normal distribution is symmetric, we can also say that $P(z > 2) = 0.5 - 0.4772 = 0.0228$.

Therefore, the percentage of batteries that have a lifetime of more than 370 hours is 2.28%.

Question 3.

i) Write uses of sampling in marketing research and campaign.

Answer:- Sampling is a cornerstone of effective marketing research and campaign strategy. By carefully selecting a representative subset of the target audience, marketers can gain valuable insights and make informed decisions without the need to survey the entire population. Here are some key uses of sampling in marketing:

1. Market Research:

* Understanding Consumer Preferences: Sampling helps identify consumer preferences, needs, and attitudes towards products, services, or brands. This information is crucial for product development, pricing strategies, and overall market positioning.

* Segmenting the Market: By analyzing sample data, marketers can identify distinct customer segments based on demographics, psychographics, and behavior. This segmentation allows for targeted marketing campaigns and personalized experiences.

* Testing Marketing Campaigns: Sampling is used to test the effectiveness of different marketing messages, channels, and creative concepts before launching full-scale campaigns. This helps optimize campaigns for better ROI.

* Measuring Brand Awareness and Perception: Sampling allows marketers to gauge brand awareness, perception, and loyalty among target audiences. This feedback helps refine brand positioning and messaging.

2. Campaign Development and Execution:

* Targeting the Right Audience: Sampling helps identify the most relevant audience segments for specific campaigns, ensuring that marketing efforts reach the most receptive consumers.

* Optimizing Campaign Messaging: By analyzing sample data, marketers can tailor campaign messages to resonate with specific audience segments, increasing the likelihood of engagement and conversion.

* Measuring Campaign Effectiveness: Sampling is used to track key metrics such as reach, engagement, and conversion rates. This data helps evaluate campaign performance and make necessary adjustments.

* Predicting Campaign Outcomes: Based on sample data, marketers can forecast potential campaign outcomes, such as sales revenue or customer acquisition costs. This information helps allocate resources effectively and set realistic goals.

Types of Sampling Methods:

* Probability Sampling: Each member of the population has a known probability of being selected. Common methods include simple random sampling, stratified sampling, and cluster sampling.

* Non-Probability Sampling: Selection is based on factors other than random chance. Common methods include convenience sampling, snowball sampling, and quota sampling.

The choice of sampling method depends on various factors, including research objectives, budget constraints, and the nature of the target population. By employing appropriate sampling techniques, marketers can ensure that their research and campaigns are data-driven, efficient, and effective.

ii. The record of weights of the male population follows the normal distribution. Its mean and standard deviation are 70 Kg and 15 Kg respectively. If a researcher considers the records of 50 males then using central limit theorem what would be the mean and standard deviation of the chosen sample?

Solution:

According to the Central Limit Theorem, for a sufficiently large sample size (usually $n \geq 30$), the distribution of sample means will be approximately normal, regardless of the underlying population distribution.

Mean of the sample:

The mean of the sample means is equal to the mean of the population.

Therefore, the mean of the sample = 70 Kg

Standard deviation of the sample (standard error):

The standard deviation of the sample means (standard error) is calculated as:

Standard error = (Population standard deviation) / $\sqrt{\text{Sample size}}$

Standard error = $15 / \sqrt{50} \approx 2.12$ Kg

Therefore, the mean of the chosen sample is 70 Kg, and the standard deviation of the chosen sample (standard error) is approximately 2.12 Kg.

iii. Explain in brief the Excel demo type of sampling method.

Solution:

Excel provides a convenient way to simulate various sampling methods. Here's a brief explanation of how it works:

- * Generate a population: Create a column in Excel with the values representing the population you want to sample from.

- * Use the RAND() function: Generate a random number between 0 and 1 for each individual in the population.

- * Sort the data: Sort the population data based on the randomly generated numbers.

- * Select a sample: Choose the first n rows (where n is your desired sample size) from the sorted data. This gives you a random sample from the population.

Example:

Let's say you have a population of 1000 individuals with their ages listed in column A.

- * In column B, enter =RAND() in each cell corresponding to the population.

- * Sort the entire dataset based on column B (the random numbers).

- * To select a sample of 50 individuals, choose the first 50 rows from the sorted data.

Key points:

- * Excel's RAND() function generates truly random numbers, ensuring a fair and unbiased sample.

- * This method is easy to implement and can be used to simulate various sampling methods like simple random sampling.

Question 4

i. Calculate the ANOVA coefficient.

Solution:

The ANOVA (Analysis of Variance) coefficient is used to compare the means of multiple groups. In this case, we are comparing the average domestic animals of Dogs,

Cats, and Hamsters.

Formula:

ANOVA Coefficient = (Sum of Squares Between Groups) / (Sum of Squares Within Groups)

Calculations:

- * Calculate the overall mean:

- * Total number of animals = 5 (Dogs) + 5 (Cats) + 5 (Hamsters) = 15

- * Total average domestic animals = 12 (Dogs) + 16 (Cats) + 20 (Hamsters) = 48

- * Overall mean = $48 / 15 = 3.2$

- * Calculate the Sum of Squares Between Groups (SSB):

- * $SSB = \sum(n_i * (\text{mean}_i - \text{overall mean})^2)$

- * For Dogs: $5 * (12 - 3.2)^2 = 409.6$

- * For Cats: $5 * (16 - 3.2)^2 = 846.4$

- * For Hamsters: $5 * (20 - 3.2)^2 = 1422.4$

- * $SSB = 409.6 + 846.4 + 1422.4 = 2678.4$

- * Calculate the Sum of Squares Within Groups (SSW):

- * $SSW = \sum(n_i * \text{standard deviation}_i^2)$

- * For Dogs: $5 * (2^2) = 20$

- * For Cats: $5 * (1^2) = 5$

- * For Hamsters: $5 * (4^2) = 80$

- * $SSW = 20 + 5 + 80 = 105$

- * Calculate the ANOVA Coefficient:

- * $ANOVA \text{ Coefficient} = SSB / SSW = 2678.4 / 105 \approx 25.46$

ii. Perform a two-way ANOVA using 5% level of significance.

Solution:

A two-way ANOVA is used to analyze the effects of two factors (water temperature and detergent) on the dependent variable (whiteness readings).

To perform a two-way ANOVA, you would typically use statistical software like SPSS, R, or Excel. The software will calculate the F-statistic and p-value for each factor and their interaction.

Interpretation:

- * If the p-value for a factor is less than the significance level (0.05), it indicates that the factor has a statistically significant effect on the whiteness readings.

iii. How to find the critical region making the decision using the P-value approach.

Solution:

- * Set the significance level: In this case, it's 5% ($\alpha = 0.05$).
- * Calculate the p-value: Using statistical software, calculate the p-value for each factor (water temperature and detergent) and their interaction.
- * Compare the p-value to the significance level:
 - * If $p\text{-value} \leq \alpha$, reject the null hypothesis. This means there is a statistically significant effect of the factor.
 - * If $p\text{-value} > \alpha$, fail to reject the null hypothesis. This means there is no statistically significant effect of the factor.

Question.5

i. Explain types of dispersion measure in detail.

Solution:

Dispersion measures quantify how spread out the data points are in a dataset. Here are some common types:

- * Range: The difference between the maximum and minimum values in the data.
- * Variance: The average squared deviation of each data point from the mean.
- * Standard Deviation: The square root of the variance, providing a measure in the same units as the data.
- * Coefficient of Variation (CV): The ratio of the standard deviation to the mean, expressed as a percentage. It allows for comparison of variability between datasets with different scales.
- * Quartile Deviation: Half the difference between the third quartile (Q3) and the first quartile (Q1). It measures the spread of the middle 50% of the data.

ii. Find the standard deviation and coefficient of variation (C.V.) of the given series:

Solution:

Data: 13, 15, 18, 19, 20

- * Calculate the mean: $\text{Mean} = (13 + 15 + 18 + 19 + 20) / 5 = 17$
- * Calculate the variance:
 - * $\text{Variance} = [(13 - 17)^2 + (15 - 17)^2 + (18 - 17)^2 + (19 - 17)^2 + (20 - 17)^2] / 5$
- $= 8$
- * Calculate the standard deviation:

- * $\text{Standard deviation} = \sqrt{\text{variance}} = \sqrt{8} \approx 2.83$

- * Calculate the coefficient of variation (CV):

- * $\text{CV} = (\text{standard deviation} / \text{mean}) * 100 = (2.83 / 17) * 100 \approx 16.6\%$

iii. Find the quartiles and quartile deviation of the following data:

Solution:

Data: 17, 2, 7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28

- * Arrange the data in ascending order: 2, 5, 7, 7, 8, 8, 10, 10, 14, 15, 17, 18, 24, 27, 28, 48
- * Calculate the first quartile (Q1):
 - * Q1 is the median of the lower half of the data.
 - * Lower half: 2, 5, 7, 7, 8, 8, 10, 10
 - * Median of lower half: $(7 + 8) / 2 = 7.5$
 - * $Q1 = 7.5$
- * Calculate the third quartile (Q3):
 - * Q3 is the median of the upper half of the data.
 - * Upper half: 14, 15, 17, 18, 24, 27, 28, 48
 - * Median of upper half: $(18 + 24) / 2 = 21$
 - * $Q3 = 21$
- * Calculate the quartile deviation:
 - * $\text{Quartile deviation} = (Q3 - Q1) / 2 = (21 - 7.5) / 2 = 6.75$

Question 6.

i. Calculate the rank coefficient of correlation from the following table:

Solution:

We will use Spearman's Rank Correlation Coefficient to calculate the correlation between the given X and Y values.

Steps:

- * Rank the X and Y values separately. Assign 1 to the highest value, 2 to the next highest, and so on. If there are ties, assign the average rank to the tied values.

X	Y	Rank of X	Rank of Y
178	125	15	5
189	137	13	3
197	156	11	1
169	112	16	6

59	107	8	7
79	136	4	4
68	123	7	8
57	108	8	7

* Calculate the difference in ranks (D) for each pair of values.

X	Y	Rank of X	Rank of Y	D = Rank X - Rank Y	D^2
---	---	---	---	---	---
78	125	5	5	0	0
89	137	3	3	0	0
97	156	1	1	0	0
69	112	6	6	0	0
59	107	8	7	1	1
79	136	4	4	0	0
68	123	7	8	-1	1
57	108	8	7	1	1

* Calculate the sum of the squared differences ($\sum D^2$).

$$\sum D^2 = 0 + 0 + 0 + 0 + 1 + 0 + 1 + 1 = 3$$

* Calculate Spearman's Rank Correlation Coefficient (rs):

$$rs = 1 - (6 * \sum D^2) / (n * (n^2 - 1))$$

where n is the number of data points (in this case, n = 8)

$$rs = 1 - (6 * 3) / (8 * (8^2 - 1)) = 1 - (18) / (8 * 63) = 1 - 0.0357 = 0.9643$$

Therefore, the rank coefficient of correlation is approximately 0.9643. This indicates a very strong positive correlation between X and Y.

ii. Calculate the mode from the following distribution:

Solution:

The mode is the value that appears most frequently in the data. In a grouped frequency distribution, the modal class is the class with the highest frequency.

From the given table, the modal class is 40-50 with a frequency of 18.

To find the exact mode, we can use the following formula for grouped data:

$$\text{Mode} = L + [(f_1 - f_0) / (2f_1 - f_0 - f_2)] * h$$

where:

* L is the lower limit of the modal class (40)

* f1 is the frequency of the modal class (18)

* f0 is the frequency of the class preceding the modal class (7)

* f2 is the frequency of the class succeeding the modal class (6)

* h is the class width (10)

$$\text{Mode} = 40 + [(18 - 7) / (2 * 18 - 7 - 6)] * 10 = 40 + (11 / 13) * 10 = 40 + 8.46 \approx 48.46$$

Therefore, the mode of the given distribution is approximately 48.46.

iii)

Solutions:-

1. Understanding the Data

We are given the following information:

* Expenditure of 1000 families is grouped into different expenditure ranges.

* The median and mean for the distribution are both Rs. 87.50.

* We need to find the missing frequencies for the 60-79 and 100-119 expenditure ranges.

2. Approach

We can use the given information about the median and mean to set up equations and solve for the missing frequencies.

3. Calculation

a. Median

* Since the median is Rs. 87.50, it lies in the 80-99 expenditure range.

* We can assume that the median class is evenly distributed.

b. Mean

* We can use the formula for the mean of a grouped frequency distribution:

$$\text{Mean} = \sum (m_i * f_i) / \sum f_i$$

where:

* m_i is the midpoint of the i-th class

* f_i is the frequency of the i-th class

* Let's denote the missing frequencies as f1 (for 60-79) and f2 (for 100-119).

* We can calculate the midpoints of each class:

| Class | Midpoint (m_i) |

| --- | --- |

| 40-59 | 49.5 |

| 60-79 | 69.5 |

| 80-99 | 89.5 |

| 100-119 | 109.5 |

| 120-139 | 129.5 |

* Now we can set up the equation for the mean:

$$87.50 = [(49.5 * 50) + (69.5 * f1) + (89.5 * 500) + (109.5 * f2) + (129.5 * 50)] / (50 + f1 + 500 + f2 + 50)$$

* Simplifying the equation:

$$87.50 = (2475 + 69.5f1 + 44750 + 109.5f2 + 6475) / (600 + f1 + f2)$$

$$87.50 * (600 + f1 + f2) = 54700 + 69.5f1 + 109.5f2$$

$$52500 + 87.50f1 + 87.50f2 = 54700 + 69.5f1 + 109.5f2$$

$$18f1 - 22f2 = 2200$$

c. Solving for f1 and f2

We have one equation and two unknowns. We need another equation to solve for f1 and f2.

* We know that the total number of families is 1000:

$$50 + f1 + 500 + f2 + 50 = 1000$$

$$f1 + f2 = 400$$

* Now we have two equations:

$$18f1 - 22f2 = 2200$$

$$f1 + f2 = 400$$

* We can solve these equations simultaneously using any method (substitution, elimination, etc.).

* Let's use the substitution method:

$$f1 = 400 - f2$$

* Substitute this value of f1 into the first equation:

$$18(400 - f2) - 22f2 = 2200$$

$$7200 - 18f2 - 22f2 = 2200$$

$$-40f2 = -5000$$

$$f2 = 125$$

* Now substitute the value of f2 into the equation for f1:

$$f1 = 400 - 125$$

$$f1 = 275$$

Therefore, the missing frequencies are:

* f1 (for 60-79 expenditure range) = 275

* f2 (for 100-119 expenditure range) = 125