CS40003
Data Analytics

Mid-Autumn Semester Test
(Session 2019-2020)

Full Marks: 50 **Time: 120 minutes**

Instructions

- This is a question-cum-answer booklet. **No separate sheet is required** for solving any problem and answering.
- There are two parts in the question paper. Answer to both the parts.
- To give your answers, **use the space provided at the end of each question**. Do not give answer anywhere else.

Part A

All questions in this part are of small answer type questions.
For a question, there may be one or more option(s) is(are) correct.
For question with more than one correct options, credit will be given on pro-rata basis.
No negative marking.
Give your answers ONLY in the space provided at the end of each question.

1.

i. In Table QI, there are attributes of different types. Out of mean, median, and mode, which is (are) the descriptive measure(s) that can be calculated for the "Income Group"?

Table QI

Patient ID	Gender	Age	Income Group

Answer to (i)
The descriptive measure(s) which can be obtained for Income Group are median and mode.

Explanation of the answer:
"Income Group" is of ordinal type. For the mode and median calculation, ordering is required and hence median and mode can be calculated.

ii. A frequency distribution of a set of 10 data is given below (see Table QII). Calculate the coefficient of variance of the data.

Table QII

x	1	2	3	4	5	6	7	8	9	10
$f(x)$	1	3	5	7	9	2	4	6	1	0

Answer to (ii)
The expression for coefficient of variance is $CV = \frac{\sigma}{\mu} \times 100$

Here, $\mu = \, = \frac{1+6+15+28+45+12+28+48+9+0}{1+3+5+7+9+2+4+6+1+0} = 5.18$

and $\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i-} \, \overline{x})^2 = \; 17.60$, that is $\sigma = 4.19$

Hence, for the given data, CV = $\frac{\sigma}{\mu}$ *100 = 80.88%

iii. Which of the following can be considered to remove outliers in data? Mark a circle on the correct answer(s).

(a) Box plot
(b) Mid range
(c) IQR (Inter Quartile Range)
(d) 0-1 normalization

iv. From which of the following measurements, the "coefficient of determination" can be calculated? Mark a circle on the correct answer(s).

(a) Degree of correlation
(b) Geometric mean
(c) Harmonic mean
(d) Number of "Type I errors"

v. If X is a random variable and $P(X = x)$ denotes the probability that X=x over a discrete domain of values of x, then which of the following is NOT true?

(a) $\sum_{\forall x} P(X = x) = 1$
(b) $\sum_{\forall x} P(X = x) = \infty$
(c) $0 \leq \sum_{\forall x} P(X = x) \leq \infty$
(d) Cannot be determined until the set of all values of x is given

vi. Which of the following is true about the sampling distribution from a normally distribution population? (All symbols in this question bear their usual meanings).

(a) $\bar{X} = \mu$ (Distribution of samples' mean is approximately normal with mean μ).
(b) S = $\sigma/\sqrt{n}$ (It is not true that sample's STD; it is distribution of samples' mean is approximately normal with STD $\sigma/\sqrt{n}$))
(c) Variance of the mean of samples' mean is σ^2/n
(d) S = $\sigma/\sqrt{n}$ is true for a large value of n

vii. If the value of α, the significant level is increased, then

(a) Type I error increases while Type-II error decreases
(b) Type I error decreases while Type-II error increases
(c) Both Type I and Type II errors increase
(d) Both Type I and Type II errors decrease

viii. Which of the following statement(s) is(are) NOT true?

(a) Pearson's correlation analysis is applicable to only numeric data.

(b) Spearman's correlation analysis is applicable to only ordinal data.

(c) χ^2 correlation analysis is applicable to only categorical data.

(d) Any non-parametric statistical learning approach is applicable when the entire population is known.

ix. Let $Y_T = \{ Y_1, Y_2, \ldots , Y_n\}$ denotes a time series data, where Y_i $(i = 1, 2, \ldots, n)$ denotes the data for any i-th period. Mark the following statements as True and False.

(a) $Y_r = \beta_0 + \sum_{i=1}^{n} \beta_i {\rho_i}^r$ denotes a auto-regression model to predict a data in r-th period where r > n and ρ^j denotes the j-th auto correlation coefficient **[False]**

(b) Auto regression analysis is possible if each Y_i $(i=1, 2, \ldots,n)$ satisfies the stationary property. **[True]**

(c) All periodic data if available in uniform and continuous manner, then only ρ^i , the i-th auto-correlation coefficient is possible. **[True]**

(d) If Y_j is predicted accurately from Y_T , then Y_k will be predicted for $k>j$ from $Y_T \cup \{Y_j\}$. **[True]**

x. Which of the following statement(s) is(are) NOT true? Mark the correct option with a circle.

(a) If confidence level is high, then probability that the null hypothesis will be **rejected** is high.

(b) The null hypothesis in Chi-Square test is that the there is no association between the attributes under test.

(c) Ogive polygon can be used to calculate the mean of a sample.

(d) From the box plot for a sample, median value can be obtained.

Part B

This part includes 5 concept level questions.
You should solve each question and give your answer in the space provided in the booklet.
Don't use any extra sheet for problem solving.

Do not give answer elsewhere.

2. The marks for 15 students on mid-term and end-term examinations in Data Analytics course are given in Table Q2.

Table Q2

Mid-term	82	73	95	66	84	89	51	82	75	90	60	81	34	49	87
End-term	76	83	89	76	79	73	62	89	77	85	48	69	51	25	74

a) Obtain the simple linear regression analysis to predict the score on the end-term examination from the mid-term examination score.

[1+2+3]

b) It is suggested that if the regression is significant, then there is no need to have final examination. How you test the significance level of your regression analysis?

[2+2]

Answer to Q2:

(a) Simple linear regression model to predict the marks of end-term scores takes the following form:

Assume, End-term= Y and Mid-term = X

So, the simple LR model to predict the marks of end-term score looks like

$$Y = \beta X + \alpha$$

Expression for the model parameters are:

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\alpha = \overline{y} - \beta\overline{x}$$

Calculated value of the model parameters:

$\overline{x}$ = 73.2, $\overline{y}$ =70.4

$\beta =$ = 0.7651

α = 70.4 – ((0.7651 * 73.2) = 14.39

(b) The validity of the model can be done as follows.

SSE = Residual sum of the squared error

= $\sum_{i=1}^{n}(actual\ output - predicted\ output)^2$

= $\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$

= 1714.62

SST = Total corrected sum of squares

= $\sum_{i=1}^{n}(actual\ output - average\ of\ the\ output)^2$

= $\sum_{i=1}^{n}(y_i - \overline{y})^2$

= 4275.6

R^2 = $1 - \frac{SSE}{SST}$

= 0.599

3. Happiness Index (HI) is measured as *low* (L), *medium* (M), *high* (H) and *very high* (VH). A survey is conducted among a population of varied age groups and data observed are recorded in Table Q3.

Table Q3

Age-group	80-90	90-100	10-20	20-30	30-40	40-50	50-60	60-70	70-80
HI	H	VH	VH	VH	M	L	L	M	H

a) Apply a suitable correlation analysis to check if there is any correlation exists between age-group and happiness index.

[1 + 1]

b) Calculate the coefficient of determination and interpret your result.

[3+2+2+1]

Answer to Q3:

(a) For the given data, Spearman correlation analysis is applicable

Justification of the pr0posed correlation analysis: The sample data are of ordinal type. And for ordinal data, the Spearman Correlation analysis is applicable.

(b) **Calculation of coefficient of deamination:**

The contingency table form the given data

Sample#	Rank$_x$	Rank$_y$	Diff=d	d^2
1	2	4.5	-2.5	6.25
2	1	2	-1	1
3	9	2	7	49
4	8	2	6	36
5	7	6.5	0.5	0.25
6	6	8.5	-2.5	6.25
7	5	8.5	-3.5	12.25
8	4	6.5	-2.5	16.25
9	3	4.5	-1.5	2.25

Calculation of coefficient of correlation:

$$r_S = 1 - \frac{6\sum_i {d_i}^2}{n(n^2-1)} = 1 - \frac{6*119.5}{9*80} = 0.00416$$

Calculation of coefficient of determination:

The coefficient of determination is $(r^s)^2 = 0.000017$

$$t = r\sqrt{\frac{n-1}{1-r^2}} = 0.0117$$

Interpretation of the result obtained:

Almost 0% pair is correlated.

4. A survey was conducted among 500 students who are studying either in "government funded collages" (GVT) or "privately funded colleges" (PVT). The objective of the survey to see the choice of "classroom based learning" (C) over the "Internet based learning" (I). The survey results are summarized in the Table Q4.

Table Q4

Learning

		C	I	
Colleges	**GVT**	75	125	200
	PVT	60	240	300
		135	365	500

It is proposed to apply the χ^2-test to verify if there is exist any association between "colleges" and "learning".

a) Decide the null and alternate hypotheses in this case. Justify your answer. **[2]**

b) Calculate the χ^2 –value from the sample data. **[2+2+2]**

c) Test the hypothesis with 5% confidence level. **[2]**

Answer to Q4:

(a) The hypothesis of the χ^2-test is given below.

H_0 : There is no association between the attributes College and Learning

H_1 : There is an association between the attributes College and Learning

Justification of the mentioned hypothesis.

The null hypothesis assumes that there is no correlation exist between the attributes under test.

(b) **Calculation of χ^2 value from the given data.**

The contingency table showing observed and expected frequencies are shown in the form of a contingency table.

Learning

		C	I	
Colleges	**GVT**	75 (54)	125 (146)	200
	PVT	60 (81)	240 (219)	300
		135	365	500

The formula for the χ^2-value is:

$$\chi^2 = \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \text{ , } o_{ij} = \text{Observed frequency and } e_{ij} = \text{Expected frequency}$$

The calculated value of χ^2-value in this case is:

$$\chi^2 = \frac{(75-54)^2}{54} + \frac{(125-146)^2}{146} + \frac{(60-81)^2}{81} + \frac{(240-219)^2}{219}$$

$= 8.16 + 3.02 + 5.44 + 2.01 = 18.63$

Testing the null hypothesis with 5% confidence level

(c) Degree of freedom for the given sample is:

$v = (r-1)\times(c-1) = 1$

The critical value of χ^2 from the χ^2-test statistical table in this case is:

Critical value is = 3.841

Inference about the null hypothesis:

As $|\chi^2| > 3.841$

Reject the null hypothesis that means class room based learning is not equal to internet based learning.

5. It is claimed that an automobile is driven on the average more than 20,000 kilometres per year. To test the claim, a random sample of 100 automobile owners is asked to keep a record of the kilometres they travel. The random sample showed an average of 23500 kilometres and a standard deviation of 3900 kilometres. It is planned to test the above with parametric based hypothesis testing. Assume 1% confidence level.

a) Mention the hypotheses that you should consider. Justify the hypothesis you have proposed. **[2]**

b) Calculate the test statistic and decide the critical region for rejecting the hypothesis. **[2+2]**

c) Decide the sample statistics. **[2]**

d) Check if the null hypothesis be rejected. **[2]**

Answer to Q5:

(a) Null hypothesis $H_0 : \mu = 20{,}000$

Justification: The problem is to infer the population mean as 20,000

Alternate hypothesis $H_1 : \mu > 20{,}000$

Justification: Alternate hypothesis is that population mean is more than 20,000

(b) This hypothesis comes under the case of t-test.

Reason: Population standard deviation is unknown.

(c) The critical region from the statistical table is:

$t = 2.365$

The test value from the sample statistics is given below:

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{23500-20000}{3900/\sqrt{100}} = 8.974$$

(d) Decision from the hypothesis testing is concluded below with justification.
H_0 is rejected that means an automobile is driven on the average more than 20,000 kilometres per year.

6. In a test paper, there are two parts with 10 and 40 marks in them. Marks in two parts are denoted by the random variable *X*. Another random variable *Y* denotes the number of students who have attended the test. Table Q6 shows the joint mass probability distribution function *f(x,y) = P(X=x, Y=y)*.

Table Q6

		Y			
	f(x,y)	10	20	30	$f_X(x)$
X	**10**	0.25	0.25	0	0.5
	40	0	0.25	0.25	0.5
	$f_Y(y)$	0.25	0.5	0.25	

(a) Calculate the covariance. Interpret the result signifying what the covariance implies.

[2+1]

(b) Calculate the coefficient of correlation. Express the meaning of the result.

[3+1]

(c) Calculate the coefficient of determination. How do you interpret the result?

[2+1]

Answer to Q6:

(a) The formula for the Covariance calculation:

$$\text{Covariance} = \sum_{x,y}(x-\mu_x)(y-m\mu_y).f(x,y)$$

The value of the covariance for the given data:

$$\mu_x = \left(\sum_x x\right).f_x = 25$$
$$\mu_y = \left(\sum_y y\right).f_y = 20$$

Covariance = 56.25

(b) Interpretation of the result:

The variance between the attributes X and Y is 56.25

The formula for the coefficient of correlation relevant to the problem is:

$$\rho = \frac{Cov\ (X, Y)}{S_x . S_y}$$

(c) The value of the coefficient of correlation for the given data:

$S_x = \sqrt{\Sigma(x_i - \mu_x)^2/(2-1)}$ =
= $\sqrt{(10-25)^2(40-25)^2}$ = 21.21

$S_y = \sqrt{\Sigma(y_i - \mu_y)^2/(3-1}) =$
= $\sqrt{(10-15)^2(20-15)^2\ (30-15)^2/2}$ = 6.12

$\rho = \frac{56.25}{21.21 * 6.12} = 0.43$

Interpretation of the result:
Since ρ is positive, we can conclude that there is a positive correlation. Further, the value of ρ is very close to 0, which implies that the correlation is very weak.

The formula for the coefficient of determination is:
Coefficient of determination = ρ^2

The value of the coefficient of determination for the given data:

ρ^2 = 0.186

Interpretation of the result:

18.6% of the variations are correlated.