ASSIGNMENT BOOKLET

Post Graduate Diploma in Applied Statistics

(Specialisation in Industrial Statistics/ Biostatistics)

MST-001 to MSTL-003

(Valid from 1st January, 2020 to 31st December, 2022)

It is compulsory to submit the assignments before filling the Examination Form.

School of Sciences
Indira Gandhi National Open University
Maidan Garhi, New Delhi-110068

Dear Student,

Please read the information on assignments in the Programme Guide that we have sent you after your enrolment. A weightage of 30%, as you are aware, has been earmarked for continuous evaluation, **which would consist of one tutor-marked assignment** for this course. The assignments for MST-001 to MSTL-003 have been given in this booklet.

Instructions for Formatting Your Assignments

Before attempting the assignment, please read the following instructions carefully:

1) On top of the first page of your answer sheet, please write the details exactly in the following format:

ENROLLMENT NO :……

NAME :……

ADDRESS :……

……

……

PROGRAMME CODE: ……..

COURSE CODE: …….

COURSE TITLE: ……

STUDY CENTRE: ……..……. DATE: …….……...

PLEASE FOLLOW THE ABOVE FORMAT STRICTLY TO FACILITATE EVALUATION AND TO AVOID DELAY.

2) Use only foolscap size writing paper (but not of very thin variety) for writing your answers.

3) Leave 4 cm margin on the left, top and bottom of your answer sheet.

4) Your answers should be precise.

5) This assignment is to be submitted at the Study Centre.

   We strongly suggest that you should retain a copy of your answer sheets.

6) This assignment is valid up to December 31, 2022.

7) **You cannot fill the Exam Form for this course** till you have submitted this assignment. So solve it and **submit it to your study centre at the earliest.** If you wish to appear in the **TEE, June 2022,** you should submit your TMAs by **March 31, 2022.** Similarly, if you wish to appear in the **TEE, December 2022,** you should submit your TMAs by **September 30, 2022.**

We wish you good luck.

TUTOR MARKED ASSIGNMENT

MST-001: Foundation in Mathematics and Statistics

Course Code: MST-001

Assignment Code: MST-001/TMA/2022

Maximum Marks: 100

Note: All questions are compulsory. Answer in your own words.

1. State whether the following statements are **True** or **False**. Give reason in support of youranswer: **(5×2=10)**

(a) Between any two different rational numbers there is another rational number.

(b) $\int_{-1000}^{1000} \left(x^{1001} + x^{2001} + x^{3001}\right) dx = 0$

(c) 210 is 51^{st} term of the sequence 10, 15, 20, 25, …

(d) $\begin{vmatrix} a & x & b+c \\ b & x & c+a \\ c & x & a+b \end{vmatrix} = 0$

(e) The range of the data shown in the following frequency distribution is 350**.**

Classes	200-250	250-300	300-350	350-400	400-450	450-500	500-550
Frequencies	0	7	3	8	4	0	0

2. (a) A carpenter was hired to build 192 window frames. The first day he made five frames and each day thereafter he made two more frames than he made the day before. How many days he will take to finish his job? **(4)**

(b) Set having values $\frac{1}{4}, \frac{1}{9}, \frac{1}{16}, \frac{1}{25}, \frac{1}{36}, \frac{1}{49}, \cdots$ is countable. **(3)**

(c) How many words each of three vowels and two consonants can be formed from the letters of the words INVOLUTE? **(3)**

3. (a) Express 700.1400.2100.2800.3500.42000 in terms of factorial.

(b) How many different signals are possible with 5 blue, 4 red, 3 white and 2 green flags byusing all at a time in a queue?

(c) If in a hall there are 10 randomly selected students then how many numbers of ways arethere such that all of them have different birthday. Assume that all of them have their birth day in non-leap years. **(2+4+4)**

4. Discuss the continuity and differentiability of the following function at x = 2/3. **(5+5)**

$$f(x)=\begin{cases}\left|x-\frac{2}{3}\right|, & x\neq\frac{2}{3}\\ 0, & x=2/3\end{cases}$$

5. Evaluate the following integrals: **(5+5)**

(i) $\int\left(10x^9+40x^4+3\right)\sqrt{x^{10}+8x^5+3x+5}\,dx$

(ii) $\int\frac{1}{(x-5)(x^2+4)}dx$

6. Find values of x, y and z given that

$5x+y+z=36$

$x+y+z=16$

$10x+2y+2z=72$

You are bound to use the matrix techniques to solve the given equations.

7. **(a)** Write flow charts of Cramer rule and matrix method. **(10)**

(b) Write whether the following data are discrete or continuous. Give reason in support of your answer.

i) Number of children in a family in a colony of 100 families.

ii) Number of pages in each of the 50 books having some mistake.

iii) Height of students of IGNOU who enrolled in 2021.

iv) Waiting time of metro when a person reaches metro station.

v) Monthly income of the family. **(5×2=10)**

8. **(a)** Write any 10 principles of data visualisation.

(b) What is the relation of unit on y-axis with unit on x-axis in histogram. **(10+10)**

TUTOR MARKED ASSIGNMENT

MST-002: Descriptive Statistics

Course Code: MST-002

Assignment Code: MST-002/TMA/2022

Maximum Marks: 100

Note: All questions are compulsory. Answer in your own words.

1. State whether the following statements are true or false and also give the reason in support of your answer: **(5×2=10)**

(a) If $X_1, X_2, X_3, ..., X_n$ and $Y_1, Y_2, Y_3, ..., Y_n$ are the variate values of two variables X and Y, and their geometric means are $G_1$and G_2, respectively, then geometric mean of (x_j/y_i); i = 1, 2, ... n will be (G_1/G_2).

(b) If X and Y are two independent variables and the variables U = X+ Y and V= X –Y, then the $r(U,V) = \frac{\sigma_X^2 - \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2}$

(c) If each value of X is divided by 2 and of Y is multiplied by 2, then b'_{YX} will be same as b.

(d) The mean and standard deviation of a set of values are 25 and 5, respectively. If a constant value 5 is added to each value, the coefficient of variation of the new set of values is equal to 10%.

(e) If (A) = 90, (AB) = 40, N = 150 and (β) = 80 then (αβ) = 30.

2. (**a**) The numbers 3.2, 5.8, 7.9 and 4.5 have frequencies Y, (Y + 2), (Y - 3) and (Y + 6), respectively. If the arithmetic mean is 4.876, find the value of Y and write the whole series. **(6)**

(**b**) The following is the distribution of age (in years) of 800 workers:

Age Group	No. of Workers
20 — 25	50
25 — 30	70
30 — 35	100
35 — 40	180
40 — 45	150
45 — 50	120
50 — 55	70
55 — 60	60

Find (i) Median, (ii) Quartile Deviation, and (iii) Coefficient of Quartile Deviation. **(10)**

3. (**a**) The value of Spearman's rank correlation coefficient of a set of non-repeating values was found to be 2/3. The sum of the squares of difference between the corresponding ranks was 55. Find the number of pairs. **(6)**

(b) Calculate Karl Pearson's coefficient of correlation between X and Y for the following data:

N = 12, $\sum X = 120$, $\sum Y = 130$, $\sum (X-8)^2 = 150$, $\sum (Y-10)^2 = 200$ and $\sum (X-8)(Y-10) = 50$.

(8)

4. (a) The following table shows the information as:

Statistical Measures	Advertisement Expenditure (X) (Rs. Lakhs)	Sales (Y) (Rs Lakhs)
Mean	20	100
Standard Deviation	03	12

r(X, Y) = 0.8. Then find

(i) the expected advertising expenditure of the company if sale is Rs. 125 lakhs, and

(ii) the expected sales of the company if the advertising expenditure is Rs 32 lakhs. **(8)**

(b) Given the following data:

r_{12} = 0' 8 , r_{13} = 0.6 and r_{23}= 0.4 then find (i) $r_{12.3}$ (ii) $r_{13.2}$ (iii) $r_{23.1}$ (iv) $R_{1.23}$ **(4)**

5. (a) An investigation of 23713 households was made in an urban and rural mixed locality. Of these 1618 were farmers, 2015 well to do and 770 families were having at least one graduate. Of these graduate families 335 were those of farmers and 428 were well to do; also 587 well to do families were those of farmers and out of them only 156 were having at least one graduate. Obtain all the ultimate class frequencies. **(6)**

(b) Can vaccination be regarded as a preventive measure for smallpox from the given data:

(i) Of 1482 persons in a locality exposed to smallpox, 368 in all were attacked, and

(ii) Of 1482 persons, 343 had been vaccinated and of these only 35 were attacked. **(6)**

6. (a) In a statistical study relating to the prices (in T) of two shares, X and Y, the following two regression lines were found as 8X - 10Y + 70 = 0 and 20X - 9Y - 65 = 0. The standard deviation of X = 3, then find (i) the values of X and Y, (ii) r(X, Y), and (iii) standard deviation of Y.

(12)

(b) Suppose X and Y are the two variables having the correlation coefficient 0.85. The following are the values they have:

X	Y
10	40
30	30
50	70
60	80

If two new variables X' and Y' are obtained by adding 50 to each value of X and 100 to each value of Y, respectively, calculate the correlation coefficient between X' and Y' using the above data. Also compare the results. **(8)**

7. (a) 50% of items have characteristics A and B both, 35% have A, but not B, 25% have B but not A. Show that there must be some misprints in this report. **(6)**

(b) In the given data, two frequencies are missing and its mean is found to be 1.46.

No. of Accidents (x)	**Frequencies (f)**
0	46
1	?
2	?
3	25
4	10
5	5
Total	200

Find the missing frequencies. **(10)**

TUTOR MARKED ASSIGNMENT

MST-003: Probability Theory

Course Code: MST-003

Assignment Code: MST-003/TMA/2022

Maximum Marks: 100

Note: All questions are compulsory. Answer in your own words.

1. State whether the following statements are **True** or **False** and also give the reason in support of your answer. **(5×2=10)**

(a) Sample space of a (i) random experiment tossing two coins simultaneously and (ii) One coin two times is the same.

(b) Standard deviation of a random variable X may take any real value, i.e. its value lies in the interval $(-\infty, \infty)$.

(c) If events $E_1, E_2, E_3, E_4, \cdots, E_n$ are mutually exclusive and exhaustive then $P(E_1 \cup E_2 \cup E_3 \cup \cdots \cup E_n)$ will be greater than 1/2 but less than 1.

(d) If S is sample space of a random experiment and E is an event defined on this sample space then P(S|E) = 1.

(e) If X is a random variable having range set {0, 1, 2, 3} then the set $\{x \in S : X(x) = 0\}$ is an event having at least one outcome of the random experiment.

2. There are 4 black, 3 blue and 8 red balls in an urn. Three balls are selected one by one without replacement. What is the probability that:

(i) First ball drawn is black, second one is red and third one is blue

(ii) All the three balls are of the same colour

(5+5)

3. A random 5-card poker hand is dealt from a standard deck of cards. Find the probability (in terms of binomial coefficients) of getting a flush (all 5 cards being of the same suit: do not count a royal flush, which is a flush with an ace, king, queen, jack and 10). **(10)**

4. Show that $f(x) = \left(\frac{1}{2}\right)^{x+1}$, $x = 0, 1, 2, 3, 4, 5, \cdots$ is a valid PMF for a discrete random variable. Also find out its CDF. **(10)**

5. A group of 100 people are comparing their birthdays (as usual, assume their birthdays are independent and not on February 29, etc.). Find the expected number of pairs of people with the same birthday, and the expected number of days in the year on which at least two of these people were born. **(10)**

6. Random variable X follows Beta distribution with parameters a = 3, b = 2 and has pdf

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

Find (i) CDF of X (ii) P[0<X<1/2] (iii) mean and variance of X without using direct formula for mean and variance. **(10)**

7. Consider the joint PDF for the type of customer service X (0 = telephonic hotline, 1 = Email) and of satisfaction score Y (1 = unsatisfied, 2 = satisfied, 3 = very satisfied):

	Y		
X	1	2	3
0	0	1/2	1/4
1	1/6	1/12	0

(a) Determine and interpret the marginal distributions of both X and Y.

(b) Calculate the 75 % quantile for the marginal distribution of Y.

(c) Determine and interpret the conditional distribution of satisfaction level for X = 1.

(d) Are the two variables independent?

(e) Calculate and interpret the covariance of X and Y. **(20)**

8. State Monty Hall problem and solve it. **(20)**

TUTOR MARKED ASSIGNMENT

MST-004: Statistical Inference

Course Code: MST-004

Assignment Code: MST-004/TMA/2022

Maximum Marks: 100

Note: All questions are compulsory. Answer in your own words.

1. State whether the following statements are **True** or **False**. Give reason in support of your answer: **(5×2=10)**

(a) If the probability of non rejection of H_0 when H_1 is true is 0.4 then power of the test will be 0.6.

(b) If T_1 and T_2 are two estimators of the parameter θ such that $Var(T_1) = 1/n$ and $Var(T_2) = n$ then T_1 is more efficient than T_2.

(c) A 95% confidence interval is smaller than 99% confidence interval.

(d) If the level of significance is the same, the area of the rejection region in a two-tailed test is less than that in a one-tailed test.

(e) Non parametric tests are more powerful than the parametric tests.

2. If a finite population has four elements: 6, 1, 3, 2.

(a) How many different samples of size n = 2 can be selected from this population if you sample without replacement?

(b) List all possible samples of size n = 2.

(c) Compute the sample mean for each of the samples given in part *b*.

(d) Find the sampling distribution of $\bar{x}$ and draw the histogram.

(e) Compute standard error.

(f) If all four population values are equally likely, calculate the value of the population mean μ. Do any of the samples listed in part (b) produce a value of $\bar{x}$ exactly equal to μ?

(13)

3. A study was conducted to compare the mean numbers of police emergency calls per 8-hour shift in two districts of a large city. Samples of 100 8-hour shifts were randomly selected from the police records for each of the two regions and the number of emergency calls was recorded for each shift. The sample statistics are listed here:

	Region	
	1	**2**
Sample size	100	100
Sample mean	2.4	3.1
Sample variance	1.44	2.64

Find a 90% confidence interval for the difference in the mean numbers of police emergency calls per shift between the two districts of the city. Interpret the interval. **(7)**

4. A bond proposal for school construction will be submitted to the voters at the next municipal election. A major portion of the money derived form this bond issue will be used to build schools in a rapidly developing selection of the city, and the remainder will be used to renovate and update school buildings in the rest of the city. To assess the viability of the bond proposal, a random sample of $n_1 = 50$ residents in the developing section and $n_2 = 100$ residents from the other parts of the city were asked whether they plan to vote for the proposal. The results are tabulated below

Sample Values for Opinion on Bond Proposal

	Developing Section	**Rest of the City**
Sample size	50	100
Number favoring proposal	38	65

(a) Estimate the difference in the true proportions favoring the bond proposal with a 99% confidence interval.

(b) If both samples were pooled into one sample of size n = 150, with 103 in favor of the proposal, provide a point estimate of the proportion of city residents who will vote for the bond proposal.

(10)

5. The following data relate to the number of items produced per shift by two workers for a number of days:

Worker A	19	22	24	27	24	18	20	19	25	
Worker B	26	37	40	35	30	40	26	30	35	45

Can it be inferred that Worker A is more stable worker compared to B by testing the variation in the item produced by them at 5% level of significance.

(10)

6. If magnitude of earthquakes recorded in a region of a country follows a distribution with parameter μ whose pdf is given below:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}, -\infty < x, \mu < \infty$$

then show that the estimators of the parameter μ using maximum likelihood and method of moments are same. **(15)**

7. A company plans to promote a new product by using one of three advertising campaigns. To investigate the extent of product recognition from these three campaigns, 15 market areas were selected and five were randomly assigned to each advertising plan. At the end of the ad campaigns, random samples of 400 adults were selected in each area and the proportions who were familiar with new product were recorded. The responses were not approximately normal. Is there a significant difference among the three population distributions from which these samples came? Use an appropriate nonparametric method to answer this question at 5% level of significance.

Campaign		
1	**2**	**3**
0.33 0.29 0.21 0.32 0.25	0.28 0.41 0.34 0.39 0.27	0.21 0.30 0.26 0.33 0.31

(15)

8. A psychology class performed an experiment to determine whether a recall score in which instructions to form images of 25 words were given differs from an initial recall score for which no imagery instructions were given. Twenty students participated in the experiment with the results listed in the table:

Student	**With Imagery**	**Without Imagery**	**Student**	**With Imagery**	**Without Imagery**
1	20	5	11	17	8
2	24	9	12	20	16
3	20	5	13	20	10
4	18	9	14	16	12
5	22	6	15	24	7
6	19	11	16	22	9
7	20	8	17	25	21
8	19	11	18	21	14
9	17	7	19	19	12
10	21	9	20	23	13

(a) What two testing procedures can be used to test for differences in the distribution of recall scores with and without imagery? What assumptions are required for the parametric procedure? Do these data satisfy these assumptions?

(b) Use both the parametric and non-parametric tests for differences in the distributions of recall scores under these two conditions.

(c) Compare the results of the tests in part b. Are the conclusions for same? If not, why not?

(20)

TUTOR MARKED ASSIGNMENT

MST-005: Statistical Techniques

Course Code: MST-005

Assignment Code: MST-005/TMA/2022

Maximum Marks: 100

Note: All questions are compulsory. Answer in your own words.

1. State whether the following statements are true or false and also give the reason in support of your answer: **(2×5=10)**

(a) The total number of all possible samples of size 2 without replacement from a population of size 7 is 21.

(b) Consecutive 3 random numbers starting from 8937 by 'middle square method' are 8937, 8699, 6726.

(c) RBD is suitable in situations where it is not possible to divide the experimental material into a number of homogeneous blocks.

(d) As we increase the sample size, representativeness of the population by the sample decreases.

(e) In a big hall, there are 50 rows and each row has 60 students. A research scholar selects 10 rows randomly and then randomly selects 15 students from each selected row. It is an example of cluster sampling procedure.

2. (a) Draw all possible samples of size 2 from the population [2, 3, 4] and verify that $E(\bar{x}) = \bar{X}$. Also find variance of $\bar{x}$. **(10)**

(b) A sample of 60 students is to be drawn from a population consisting of 600 students belonging to two villages, A and B. The means and standard deviations of their marks are give below:

Villages	**Stratum sizes (N_i)**	**Means (x_i)**	**Standard deviations**
Village A	400	60	20
Village B	200	120	80

What are the sample sizes for the two villages using proportional allocation technique? **(6)**

3. To determine the yield rate of wheat in a district of Punjab, 6 groups of 6 plots each were constructed. The data are given in the following table:

Plot No.	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**Group 6**
1	8	6	18	13	17	12
2	13	5	8	7	15	15
3	11	16	6	13	10	11
4	26	5	10	6	21	17
5	13	16	16	7	20	8
6	31	5	20	2	25	10

Select a cluster sample of 3 clusters from the above data and find its sample mean. Further, explain the procedure of two-stage sampling if we want to draw a sample of 6 plots. Which are the 6 plots in your sample? **(7)**

4. The following data relate to production in kg of three varieties P, Q, R of wheat:

P :	14	16	18		
Q :	14	13	15	22	
R :	18	16	19	15	20

Is there any significant difference among the three varieties at 5% level of significance? **(7)**

5. A researcher wants to test four diets A, B, C, D on growth rate in mice. These animals are divided into 3 groups according to their weights. Heaviest 4, next 4 and lightest 4 are put in Block I, Block II, and Block III, respectively. Within each block, one of the diets is given at random to the animals. After 15 days, increase in weight is noted, which is given in the following table:

Blocks	**Treatments/Diets**			
	A	**B**	**C**	**D**
I	12	8	6	5
II	15	12	9	6
III	14	10	8	5

Perform a two-way ANOVA to test whether the data indicates any significant difference between' the four diets due to different blocks. **(10)**

6. In the following data, two values are missing. Estimate these values by Yates method and analyse the data by suitable technique.

Treatments	**Blocks**		
	I	**II**	**III**
A	12	14	12
B	10	y	8
C	x	15	10

(12)

7. Identify the design given in the following table and then carry out the analysis:

Column	**Row**			
	I	**II**	**III**	**IV**
I	A 8	C 18	B 11	D 8
II	C 16	B 10	D 7	A 4
III	B 12	D 10	A 6	C 20
IV	D 10	A 9	C 28	B 16

(14)

8. (a) The distribution function of Pareto distribution is given by $f(x) = 1 - \left(\frac{k}{x}\right)^a$, $a > 0, 0 < k \leq x$.

Given a U~ U(0, 1), generate a random number from the above distribution, when a = 2 and k = 1. Suppose U = 0.5, then find x. **(4)**

(**b)** Generate a complete cycle for the LCG given below: $x_i = (5x_{i-1} + 3) \bmod 16$, with $x_o = 5$. A man tosses an unbiased coin ten times. Using the first ten random numbers generated above, obtain a sequence of heads and tails by taking Head (H) as $u \geq 0\text{-}5$. **(10)**

9. Times between successive crashes of a computer system were generated for a 6-month period and are given in increasing order as follows (time in hours):

1	10	20	30	40	52	63	70	80	90	100	102
130	140	190	210	266	310	530	590	640	1340		

The parameter a = 0.00435, mean = $1/\alpha$ = 230 hrs.

Use the Kolmogorov-Smirnov test to examine the goodness of fit of exponential distribution.

(10)

TUTOR MARKED ASSIGNMENT
MSTE-001: Industrial Statistics-I

Course Code: MSTE-001

Assignment Code: MSTE-001/TMA/2022

Maximum Marks: 100

Note: All questions are compulsory. Answer in your own words.

1. State whether the following statements are **True** or **False**. Give reason in support of your answer: **(5×2=10)**

(a) Statistical quality control (SQC) is a technique of process control only.

(b) Twenty pieces of different length of cloth contained 2, 4, 1, 3, 5, 4, 2, 7, 3, 5, 2, 2, 4, 5, 6, 4, 2, 1, 2, 4 defects respectively. To check the process is under control with respect to the number of defects, we should use np-chart.

(c) If the probabilities are not associated with the occurrence of different states of nature, then the situation is known as decision making under risk.

(d) In single sampling plan, if we increase acceptance number then the OC curve will be steeper.

(e) A system has four components connected in parallel configuration with reliability 0.2, 0.4, 0.5, 0.8. To improve the reliability of the system most, we have to replace the component which reliability is 0.2.

2. Twenty samples each of size 10 were inspected. The number of defectives detected in each of them is given below: 0, 1, 0, 3, 9, 2, 0, 7, 0, 1, 1, 0, 0, 3, 1, 0, 0, 2, 1, 0 Find the control limits for the number of defectives and establish quality standards for the future. Plot the graph and interpret. **(10)**

3. A manufacturer of men's jeans purchases zippers in lots of 500. The jeans manufacturer uses single-sample acceptance sampling with a sample size of 10 to determine whether to accept the lot. The manufacturer uses c = 2 as the acceptance number. Suppose 3% nonconforming zippers are acceptable to the manufacturer and 8% nonconforming zippers are not acceptable. Let incoming quality be 4%.

i) Construct an OC curve.

ii) Average outing quality **(**AOQ), if the rejected lots are screened and all defective zippers are replaced by non-defectives.

iii) Average total inspection (ATI). **(6+2+2)**

4. An office supply company ordered a lot of 400 printers. When the lot arrives the company inspector will randomly inspect 12 printers. If more than three printers in the sample are non-conforming, the lot will be rejected. If fewer than two printers are non-conforming, the lot will be accepted. Otherwise, a second sample of size 8 will be taken. Suppose the inspector finds two non-conforming printers in the first sample and two in the second sample. Also AQL and LTPD are 0.05 and 0.10 respectively. Let incoming quality be 4%.

i) What is the probability of accepting the lot at the first sample?

ii) What is the probability of accepting the lot at the second sample?

iii) Find AQL and ATI **(15)**

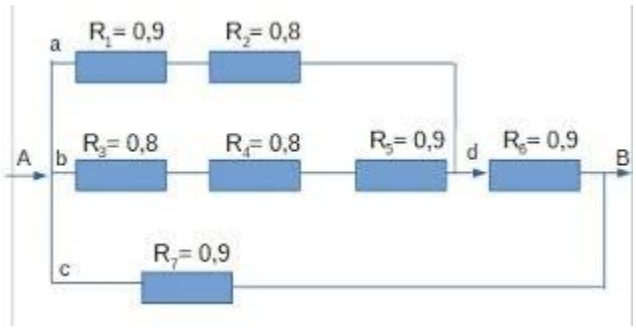
5. A two-person zero-sum game having the following payoff matrix for player A

		Player B				
		I	**II**	**III**	**IV**	**V**
Player A	**I**	2	4	3	8	4
	II	5	6	3	7	8
	III	6	7	9	8	7
	IV	4	2	8	4	3

(i) Check whether saddle point exit or not.

(ii) If saddle point does not exit then determine optimal strategies for both the manufacturers and value of the game. **(2+8)**

6. A system has seven independent components and reliability block diagram of it shown blow:



Find reliability of the system. **(10)**

7. The failure data for 40 electronic components is shown below:

Operating Time (in hours)	0-5	5-10	10-15	15-20	20-25	25-30
Number of Failures	5	7	6	4	5	4
Operating Time (in hours)	30-35	35-40	40-45	45-50	≥50	
Number of Failures	4	0	2	1	2	

Estimate the reliability, cumulative failure distribution, failure density and failure rate functions. **(15)**

8. At a call centre, callers have to wait till an operator is ready to take their call. To monitor this process, 5 calls were recorded every hour for the 8-hour working day. The data below shows the waiting time in seconds:

Time	**Sample Number**				
	1	**2**	**3**	**4**	**5**
9 a.m	8	9	15	4	11
10	7	10	7	6	8
11	11	12	10	9	10
12	12	8	6	9	12
1 p.m.	11	10	6	14	11
2	7	7	10	4	11
3	10	7	4	10	10
4	8	11	11	11	7

i) Use the data to construct control charts for mean and variability and comments about the process. If process is out of control, then calculate the revised control limits.

ii) If the specification limits as the 8±2, then find the process capability. Does it appear that the process is capable of meeting the specification requirements? **(20)**

TUTOR MARKED ASSIGNMENT
MSTE-002: Industrial Statistics-II

Course Code: MSTE-002

Assignment Code: MSTE-002/TMA/2022

Maximum Marks: 100

Note: All questions are compulsory. Answer in your own words.

1. State whether the following statements are **True** or **False**. Give reasons in support of your answers. **(5x2=10)**

(a) The solution of a transportation problem with 5 rows (supplies) and 4 columns (destinations) is feasible if number of possible allocations are 8.

(b) The moving averages of suitable period in a time-series are free from the influences of seasonal and cyclic variations.

(c) If the basic solutions for a system of equations are (- 2, 0, 1), (0, 1, 3), (- 2, 3, 0), then only (0, 1, 3) is feasible.

(d) In the stepwise selection method of multiple regression model, once a variable enters in the model then it always remains in the model.

(e) An enterprise requires 1000 units per month. The ordering cost is estimated to be 50 per order. The purchase price is 20 per unit and the carrying cost per unit is 10% of it. Then the economic lot size to be ordered is 775.

2. Use the penalty (Big M) method to solve the following LP problem: **(12)**

Minimise $Z = 5x_1 + 3x_2$

Subject to the constraints

$2x_1 + 4x_2 \leq 12$

$2x_1 + 2x_2 = 10$

$5x_1 + 2x_2 \geq 10$

$x_1, x_2 \geq 0$.

3. A company has three production facilities S_1 , S_2 and S_3 with production capacity of 7, 9 and 18 units (in 100s) per week of a product, respectively. These units are to be shipped to four warehouses D_1, D_2, D_3 and D_4 with requirement of 5, 6, 7 and 14 units (in 100s) per week, respectively. The transportation costs (in Rs) per unit between factories to warehouses are given in the table below:

	D_1	**D_2**	**D_3**	**D_4**	**Capacity**
S_1	19	30	50	10	7
S_2	70	30	40	60	9
S_3	40	8	70	20	18
Demand	5	8	7	14	34

Obtain optimal solution by the MODI method. **(12)**

4. Four professors are capable of teaching any one of four different courses. Class preparation time in hours for different topics varies from professor to professor and is given in the table below:

Professor	**A**	**B**	**C**	**D**
Linear Programming	2	15	13	4
Queuing Theory	10	4	14	15
Transportation Problem	9	14	16	13
Regression Analysis	7	8	11	9

Each professor is assigned only one course. Determine an assignment schedule so as to minimise the total course preparation time for all courses. **(10)**

5. In a railway marshalling yard, goods trains arrive at a rate of 36 trains per day. Assuming that the inter-arrival and service time distributions both follow exponential distribution with an average of 30 minutes, calculate the following:

(i) Traffic intensity

(ii) The mean queue length

(iii) Probability that the queue size exceeds **(8)**

6. Using the graphical method to minimise the time required to process Job 1 and Job 2 on five machines A, B, C, D and E, find the minimum elapsed times an idle times to complete both jobs.

Job 1	**Sequence**	A	B	C	D	E
	Time (in hours)	1	2	3	5	4
Job 2	**Sequence**	C	A	D	E	B
	Time (in hours)	3	4	2	1	5

(8)

7. A firm wants to know whether there is any linear relationship between the sales (X) and its yearly revenue (Y). The records for 10 years were examined and the following results were obtained:

$\sum X = 265$, $\sum Y = 27.73$, $SS_X = 285.6$, $SS_Y = 6.978$ and $SS_{XY} = 57.456$.

(a) Fit a regression line taking Y as the dependent variable and X as the independent variable.

(b) Test whether the sales have any effect on revenue at 5% level of significance.

(c) Comment on the goodness of fit of the regression line. **(2+3+5)**

8. A researcher is interested in developing a linear model for the electricity consumption of a household having an AC (1.5 ton) so that she can predict the electricity consumption. For this purpose, she selects 25 houses and records the electricity consumption (in kWh), size of house (in square feet) and AC hours for one month during summers. The results obtained are:

$\hat{B}_0 = 22.38, \hat{B}_1 = 1.6161, \hat{B}_2 = 0.0144$, $SS(B_0) = 12526.08$, $SS(B_0,B_1) = 17908.47$, $SS(B_0,B_2) = 17125.23$, $SS(B_0, B_1, B_2) = 18079.0$, $\hat{\sigma}^2 = 10.53$, $SE\left(\hat{B}_1\right) = 0.17$, and $SE\left(\hat{B}_2\right) = 0.0035$.

Build a regression model by selecting appropriate regressors in the model using the Stepwise Selection method. **(10)**

9. The following table represents the sales (in thousands) of mobile sets of a shop for 16 quarters over four years:

Year	**Quarter**			
	Q_1	**Q_2**	**Q_3**	**Q_4**
2011	554	590	616	653
2012	472	501	521	552
2013	501	531	553	595
2014	403	448	460	480

(a) Compute the seasonal indices for four quarters by Simple average method.

(b) Obtain deseasonlised values. **(10)**

10. Seven successive observations on a stationary time-series are as follows:

12, 14, 13, 10, 15, 12, 15

(a) Calculate auto-covariances C_0, C_1, C_2, C_3 and C_4.

(b) Calculate auto-correlation coefficients r_1, r_2, r_3 and r_4.

(c) Plot the correlogram. **(10)**

TUTOR MARKED ASSIGNMENT
MSTE-003: Biostatistics-I

Course Code: MSTE-003
Assignment Code: MSTE-003/TMA/2022
Maximum Marks: 100

Note: All questions are compulsory. Answer in your own words.

1. State whether the following statements are **True** or **False**. Give reason in support of your answer: **(2×5=10)**

(a) Suppose A is the exposure and B is a confounding factor for outcome C, then there will be a path from A to C via B.

(b) Doing exercise may also be a regimen.

(c) In clinical trials, a control only may be: treatment or no treatment.

(d) In Greville's method, the central death rate is more in the life table than the population.

(e) In a slope ratio assay, both regression lines have common slope.

2. Explain with examples:

(a) Various sources of demographic data in India

(b) Types of bioassays

(c) LD50 and ED50

(5+5+5)

3.(a) Differentiate between complete and abridged life tables.

(b) The data on population and number of deaths for different age groups of Districts A and B in the year 2001 were collected in the following table:

Age Group (Years)	District A		District B	
	Population	No. of Deaths	Population	No. of Deaths
0 - 5	55,300	385	51,165	805
5 –15	109,125	410	98,170	510
15-35	1,72,050	675	1,68,450	790
35-50	1,15,600	1625	1,40,625	990
50 & above	2,65,775	3288	2,40,900	2485

Calculate standardised death rate by direct method, taking population of District A as the standard population.

(5+5)

3. In a parallel-line assay, total 18 guinea pigs, 4 guinea pigs each from 3 different breeds were selected and classified into 4 groups for each breed. Two groups were administered with two doses of the standard preparation and remaining two groups with two doses of the test preparation. The responses of these doses are recorded in the following table:

	Dose (Standard)			**Dose (Test)**		
Breed	**10 (in µL)**	**15 (in µL)**	**20 (in µL)**	**5 (in µL)**	**10 (in µL)**	**15 (in µL)**
1	25	42	55	20	43	64
2	23	47	52	23	42	66
3	22	38	58	24	44	67

(i) Determine the dose-response regression models for both preparations.

(ii) Test whether the dose-response curves of both preparations are linear and parallel to each other or not.

(iii) Interpret whether the relative potency and its confidence interval can be computed or not.

(8+15+2)

5.(a) If D^+ and D^- denote presence and absence, respectively, of a disease and T^+ and T^- denote test result as positive and negative, respectively, then on the basis of the following information:

	D^+	D^-	Total
T^+	145	2000	2145
T^-	15	48000	48015
Total	160	50000	50160

Find: (i) sensitivity (ii) specificity (iii) positive and negative predictive values.

(b) Obtain sample size for the following given information:

$\delta = 0.04,\ \pi_1 = 0.72,\ \pi_2 = 0.84,\ \alpha = 0.01,\ \beta = 0.25$ **(5+5)**

6.(a) Explain design and analysis of data of case control study in detail.

(b) Creatinine excretion is a parameter of kidney function. Generally speaking, lower values indicate better health. This depends on body weight. A researcher conducted a study on creatinine excretion in test group and control group to find the efficacy of a new drug. The subjects were randomly divided. He included 100 subjects in each group but for this exercise consider only 10 subjects in each group. The data obtained on creatinine level in these 10 subjects are as follows:

Test group:	16.6	19.8	17.1	17.0	15.6	20.3	24.7	18.5	17.6	22.0
Control group:	23.2	22.0	21.9	16.7	14.2	23.2	24.8	25.5	28.1	21.8

Do you think that creatinine excretion was really lower in the test group on average?

(15+5)

7. Suppose you try a regimen A on 1000 subjects and regimen B on 1600 subjects. Results of the trial show that efficacies of regimen A and B are 76% and 82% respectively. Suppose doctor determines 4% as superiority margin. Can you consider regimen B as superior to regimen A.

(10)

TUTOR MARKED ASSIGNMENT

MSTE-004: Biostatistics-II

Course Code: MSTE-004

Assignment Code: MSTE-004/TMA/2022

Maximum Marks: 100

Note: All questions are compulsory. Answer in your own words.

1. State whether the following statements are **True** or **False**. Give reason in support of your answer: **(2×5=10)**

(a) The value of sensitivity of the following results of a diagnostic test is 0.85.

Disease	Test result		Total
	+	–	
Present	170	30	200
Absent	20	280	300

(b) For the following cohort study, the relative risk for the lung cancer among smokers is 3.5.

	Lung Cancer	No Lung Cancer	Total
Smokers	100	1220	1320
Non-smokers	50	2260	2310

(c) The logit link function is $\log\left[-\log(1-\pi)\right]$.

(d) We define three indicator/dummy variables for a regressor variable with three categories.

(e) Left censoring occurs whenever the exact time of occurrence of an event is not known.

2. Differentiate between Chi-square tests for association and homogeneity of proportions. Also mention the assumptions of these tests.

(10)

3. A random sample of 250 patients was selected and their workout timing and diabetes status were recorded. The following table shows the workout timing and severity of diabetes:

Workout (in minutes)	Severity of diabetes status		
	Low	Moderate	High
0 −15	06	27	19
15 to 30	08	36	17
30 to 45	21	45	33
≥ 45	14	18	06

Test at 5% level of significance whether workout habit and diabetes are associated with to each other or not.

(10)

4.(a) Explain the assumptions underlying multiple linear regression model.

(b) Suppose a researcher wants to evaluate the effect of cholesterol on the blood pressure. The following data on serum cholesterol (in mg/dL) and systolic blood pressure (in mm/Hg) were obtained for 15 patients to explore the relationship between cholesterol and blood pressure:

S. No.	Cholesterol (mg/dL)	SBP (mm/Hg)
1	300	150
2	410	270
3	380	210
4	530	310
5	570	350
6	490	310
7	340	210
8	320	150
9	280	110
10	550	320
11	340	220
12	350	170
13	410	260
14	390	230
15	450	270

(i) Fit a linear regression model using the method of least squares.

(ii) Construct the normal probability plot for the data on serum cholesterol and systolic blood pressure.

(iii) Test the significance of the fitted regression model.

(5+15)

4. Write a short note on the following:

(i) Polytomous logistic models

(ii) Poisson regression

(iii) Kaplan and Meier method

(12)

6. The following data on diagnosis of coronary heart disease (where 0 indicating absence and 1 indicating presence), serum cholesterol (in mg/dl), resting blood pressure (in mmHg) and weight (in kg) were obtained for 80 patients to explore the relationship of coronary heart disease with cholesterol and weight.

S. No.	Serum Cholesterol (mg/dl)	Weight (kg)	Number of Patients having CHD	Total Number of Patients
1	420	60	10	20
2	450	68	15	30
3	400	54	4	15
4	510	74	2	10
5	480	62	1	5

(i) Fit a multiple logistic model for the dependence of coronary heart disease on the average serum cholesterol and weight considering $\hat{\beta}_0^0 = 4.279$, $\hat{\beta}_1^0 = -0.035$ and $\hat{\beta}_2^0 = 0.172$ as the initial values of the parameters (solve only for one Iteration).

(ii) Test the significance of the fitted model using Hosmer-Lemeshow test at 5% level of significance.

(12+8)

7.(a) Describe censoring and differentiate between different types of censoring with the help of examples which are not considered in Block 4 of MSTE-004.

(b) A study was conducted on 185 patients aged more than 45 years which are followed until the time of death or up to 10 years, whichever comes first. The patients have different covariates: age, gender (male/female), systolic blood pressure, smoking (yes/no), total serum cholesterol and diabetes (yes/no). The objective of this study is to determine which covariate influences the survival time. An analysis is conducted to investigate differences in all-cause mortality between men and women participating in the study. Suppose we obtain the following results after applying the Cox regression hazard model analyses:

Risk Factor	Parameter Estimate	SE
Age	0.150	0.010
Gender	0.450	0.150
Systolic Blood Pressure	0.015	0.008
Smoking	0.650	0.170
Total Serum Cholesterol	0.002	0.004
Diabetes	-0.350	0.250

(i) Obtain hazard ratio and interpret the results.

(ii) Find the 99% confidence interval for the hazard ratio.

(iii) Test whether the covariates are significant or not at 1% level of significance.

(8+10)

TUTOR MARKED ASSIGNMENT
MSTL-001: Basic Statistics Lab

Course Code: MSTL-001

Assignment Code: MSTL-001/TMA/2022

Maximum Marks: 100

Note:

1. **All questions are compulsory.**
2. **Solve the following questions in MS Excel.**
3. **Take the screenshots of the final output/spreadsheet.**
4. **Paste all screenshots in the assignment booklets with all necessary interpretation and steps.**

1. The production of semiconductors needs a lot of water to cool down the equipment and clean silicon wafers. To study the water wastage during the production of semiconductor chips, the data of the water required for past 100 days in two plants: Plant A and Plant B are given in the following table:

Day	Plant A (in '000 litres)	Plant A (in '000 litres)	Day	Plant A (in '000 litres)	Plant A (in '000 litres)
1	616	670	**51**	910	1092
2	656	766	**52**	866	1040
3	780	910	**53**	672	806
4	728	850	**54**	704	846
5	814	650	**55**	810	972
6	648	756	**56**	854	1024
7	748	872	**57**	638	766
8	780	910	**58**	740	890
9	624	730	**59**	854	1024
10	642	750	**60**	650	780
11	764	892	**61**	684	712
12	814	950	**62**	632	760
13	656	766	**63**	752	902
14	678	790	**64**	702	842
15	858	1030	**65**	786	942
16	632	760	**66**	624	748
17	624	702	**67**	722	866
18	900	1080	**68**	752	902
19	684	728	**69**	642	722
20	726	872	**70**	664	742
21	912	1096	**71**	738	886
22	868	1042	**72**	786	942
23	674	810	**73**	632	760
24	708	850	**74**	652	784
25	812	976	**75**	852	1022
26	856	1028	**76**	626	752
27	640	768	**77**	692	696
28	744	892	**78**	894	1072
29	856	1028	**79**	652	722
30	652	782	**80**	722	866

31	774	716	**81**	908	1090
32	636	762	**82**	862	1036
33	754	906	**83**	668	802
34	704	846	**84**	702	842
35	788	946	**85**	808	970
36	626	752	**86**	850	1022
37	724	870	**87**	634	762
38	754	906	**88**	738	886
39	632	726	**89**	850	1022
40	622	746	**90**	646	776
41	740	890	**91**	704	710
42	788	946	**92**	630	756
43	636	762	**93**	750	900
44	656	788	**94**	700	840
45	854	1026	**95**	782	940
46	630	756	**96**	654	746
47	708	698	**97**	718	862
48	896	1076	**98**	750	900
49	624	726	**99**	628	720
50	724	870	**100**	702	740

Answer the followings:

i) Which plant has more wastage of water?

ii) Which plant shows greater variability in the wastage of water?

iii) Determine the correlation between the wastage in both plants.

iv) Compute suitable width of the class intervals for both plants.

v) Construct the continuous frequency distribution for both plants. **(25)**

2. The number of employees (in hundreds) and the revenues (in ₹ hundred crores) of 20 companies were recorded to access the relationship between the revenue generated and strength of the employees. The data are given in the following table:

S. No.	No. of Employees (in '00)	Revenue (in ₹ '00 crores)
1	165	335
2	550	425
3	330	345
4	550	415
5	275	325
6	330	360
7	385	360
8	330	365
9	440	410
10	385	375
11	275	335
12	495	390
13	495	395
14	440	395
15	495	400
16	440	425
17	220	320
18	330	420

19	440	425
20	330	370

Compute the Spearman's rank correlation coefficient between the number of employees and the revenues of the companies. **(25)**

3. For the data given in Question 1, compare the water wastage of both plants to get the answers of the following questions:

i) Is there enough evidence that the average water wastage of Plant A is more than the average water wastage of Plant B at 5 % level of significance?

ii) Are the variances of the distributions of water wastage of Plants A and B equal at 5 % level of significance?

4. Suppose that a production house wants to evaluate the popularity of eight books on a particular subject. The customer service manager of the production house hires seven evaluators with varying experience in that subject to review the books. To reduce the effect of the variability from evaluator to evaluator, she uses a randomised block design, with evaluators serving as the blocks. The eight books are the groups of interest.

The seven evaluators assigned to each of the eight books in a random order. A rating scale from 0 (low) to 100 (high) is used. The following table summarises the results:

	Books							
Evaluators	A	B	C	D	E	F	G	H
1	68	59	80	72	66	57	78	79
2	75	73	86	74	73	71	84	85
3	74	65	88	78	72	63	80	86
4	78	61	85	74	76	59	83	84
5	82	64	90	82	80	62	88	89
6	76	66	92	84	74	64	78	90
7	75	73	86	74	73	71	84	82

The effect of each evaluator of eight books is normally distributed with approximately equal variances.

i) Analyse the design at 5% level of significance.

ii) Is the average popularity of the eight books significantly different? If the difference between the averages popularity of the eight books is significant, do the pair-wise comparison between them. **(25)**

TUTOR MARKED ASSIGNMENT
MSTL-002: Industrial Statistics Lab

Course Code: MSTL-002

Assignment Code: MSTL-002/TMA/2022

Maximum Marks: 100

Note:

1. **All questions are compulsory.**
2. **Solve the following questions in MS Excel.**
3. **Take the screenshots of the final output/spreadsheet.**
4. **Paste all screenshots in the assignment booklets with all necessary interpretation and steps.**

1. A Company wants to maintain the quality of bottling process which uses a particular brand of machine to fill 100 ml sanitizer spray bottle. During each shift, a sample of 10 bottles is selected (2 hours apart) and the volume of the each filled sanitizer spray bottle (in ml) is determined. In this regards, total 25 subgroups consisting of a sample of 10 bottles in each subgroup were selected. The following table lists the measurements from 25 consecutive shifts:

Sample No.	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5	Obs. 6	Obs. 7	Obs. 8	Obs. 9	Obs. 10
1	99.46	100.12	99.73	99.56	99.46	99.73	99.46	100.12	99.73	99.56
2	100.95	100.00	99.66	100.06	100.95	99.66	100.95	100.00	99.66	100.06
3	99.84	99.43	99.62	99.73	99.84	99.62	99.84	99.43	99.62	99.93
4	99.85	99.26	99.77	99.56	99.85	99.77	99.85	99.86	99.77	99.76
5	99.66	100.12	99.91	100.02	99.66	99.91	99.66	100.12	99.91	100.02
6	99.82	100.06	99.87	100.18	99.82	99.87	99.82	100.06	99.87	100.18
7	99.86	99.66	99.46	99.52	99.86	99.46	99.86	99.66	99.86	99.72
8	99.87	99.62	100.12	99.62	99.87	100.12	99.87	99.62	100.12	99.62
9	99.85	99.67	100.13	100.07	99.85	100.13	99.85	99.67	100.13	100.07
10	99.72	99.52	99.85	99.71	99.72	99.85	99.72	99.52	99.85	99.71
11	99.88	100.00	100.26	99.87	99.88	100.26	99.88	100.00	99.86	99.87
12	99.65	100.06	100.15	99.93	99.65	100.15	99.65	100.06	100.15	99.93
13	99.46	99.70	99.43	99.87	99.86	99.63	99.66	99.70	99.83	99.87
14	99.91	99.60	99.75	100.22	99.91	99.75	99.91	99.60	99.75	100.22
15	100.04	100.06	99.82	99.85	100.04	99.82	100.04	100.06	99.82	99.85
16	100.15	100.10	99.93	99.62	100.15	99.93	100.15	100.10	99.93	99.62
17	100.13	99.56	98.81	99.25	100.13	98.81	100.13	99.56	98.81	99.25
18	99.82	100.05	100.08	99.75	99.82	100.08	99.82	100.05	100.08	99.75
19	99.90	100.10	99.91	100.15	99.90	99.91	99.90	100.10	99.91	100.15
20	99.50	99.76	99.95	99.56	99.50	99.95	99.50	99.76	99.95	99.56
21	99.88	100.12	100.00	100.25	99.88	100.00	99.88	99.92	100.00	100.05
22	99.87	99.67	99.51	99.71	99.87	99.51	99.87	99.67	99.51	99.71
23	100.31	99.91	99.60	99.96	100.31	99.60	100.31	99.91	99.60	99.96
24	99.66	99.46	99.91	99.72	99.66	99.91	99.66	99.46	99.91	99.72
25	99.90	100.02	99.70	100.07	99.90	99.70	99.90	100.02	99.70	100.07

Construct suitable control charts for variability as well as for average to infer whether the process of bottling is under statistical control or not. If it is out-of-control, also plot the revised control charts, if necessary. **(25)**

2. An automobile manufacturing company examined the cars of a particular model to identify the number of defects during the final inspection stage. The total number of inspected cars were recorded for last 35 days along with the number of defects. The results are given in the following table:

Days	Total Inspected Cars	Number of Defects
1	45	6
2	35	2
3	40	1
4	30	2
5	40	5
6	25	1
7	35	3
8	30	4
9	30	6
10	40	2
11	35	11
12	35	1
13	25	3
14	40	7
15	45	2
16	40	6
17	30	5
18	35	2
19	30	7
20	45	3
21	28	1
22	38	4
23	33	11
24	33	1
25	43	4
26	38	2
27	38	5
28	28	4
29	43	6
30	48	1
31	43	3
32	33	4
33	38	3
34	33	2
35	48	5

Construct a suitable control chart for the number of defects to check whether the process is under statistical control or not. Also plot the revised control charts, if necessary. **(25)**

3. A Mobility-as-a-Service (MaaS) provider company conducted a study to check the relationship of several variables with its weekly commuters. For this purpose, thirty cities were selected and the number of weekly commuters were recorded along with other variables like: the average petrol price (in ₹), population of the city, monthly income of commuters (in ₹), average parking rates per month (in ₹). The data are given in the following table:

City	Number of Weekly Commuters	Average Petrol Price	Population of City (in '000)	Average Monthly Income of Commuters (in '00)	Average Monthly Parking Rates (in ₹)
1	17700	75	1900	580	1000
2	17540	75	1890	620	1000
3	17620	75	1880	640	1200
4	16260	77	1878	650	1200
5	16180	77	1850	655	1200
6	16340	77	1840	658	1400
7	16580	77	1825	820	1500
8	16020	82	1825	860	1500
9	15940	82	1820	880	1500
10	15892	82	1805	920	1600
11	15780	85	1810	963	1600
12	14820	95	1800	1057	1600
13	14660	95	1795	1133	1700
14	14660	96	1795	1160	2000
15	14580	96	1790	1180	2100
16	14420	96	1730	1183	2100
17	13380	97	1740	1265	2100
18	10070	120	1735	1300	2200
19	13220	102	1730	1325	2500
20	13540	102	1720	1380	2600
21	13700	102	1715	1401	3000
22	12100	107	1705	1450	3100
23	11124	113	1690	1500	3300
24	10900	125	1695	1520	3500
25	11108	114	1690	1560	3500
26	13668	104	1700	1600	3800
27	13780	90	1710	1620	4000
28	12108	118	1790	1590	3700
29	14668	108	1800	1630	4000
30	14780	94	1810	1650	4200

Now determine the most appropriate regression model for the number of weekly commuters using stepwise approach at 5 % level of significance and interpret the results. Does the final regression model satisfy the linearity and normality assumptions? **(25)**

4. A division of climatic change is interested in analysing the pattern of the CO_2 concentrations in the air of a particular state in past years and then forecasting the CO_2 concentrations for the upcoming years. The monthly mean CO_2 concentrations ppm (parts per million) mixing ratio in dry air from January 2006 through December 2019 were recorded. The following monthly data are given in the following table for past 14 years:

	Year													
Month	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
January	299.16	299.62	300.60	301.67	303.71	305.00	306.77	308.32	309.90	310.41	312.87	313.88	315.04	316.33
February	299.94	300.40	301.60	302.17	304.23	305.63	307.26	309.41	310.70	311.68	313.59	314.62	315.70	316.82
March	300.41	300.87	302.57	303.86	305.54	306.93	309.10	310.69	311.70	312.04	314.11	316.23	316.38	318.29
April	301.72	302.18	303.72	305.07	306.79	307.95	309.88	311.51	312.65	314.27	316.07	316.62	318.38	319.91
May	302.13	302.92	303.68	305.82	306.95	308.05	310.47	312.02	313.28	314.92	316.38	317.53	318.93	319.41
June	301.09	302.43	303.17	305.12	307.00	308.27	310.31	311.54	312.42	314.40	315.78	316.87	318.26	318.74
July	300.10	300.85	301.96	303.81	305.37	306.64	308.41	309.88	311.02	313.16	313.96	315.00	316.44	316.92
August	298.14	299.01	299.80	301.56	303.47	304.68	306.74	307.75	308.97	311.11	311.71	312.86	314.55	315.03
September	296.36	297.51	297.98	300.30	301.46	302.77	305.07	306.05	306.84	309.11	309.86	311.55	313.21	313.69
October	296.29	297.41	297.57	300.22	301.29	303.09	305.07	306.13	307.00	309.15	310.13	311.57	312.67	313.15
November	297.23	298.25	299.20	301.37	302.76	304.29	306.22	307.45	308.20	310.38	311.84	313.11	314.09	314.57
December	298.55	299.97	300.58	302.49	303.80	305.76	307.38	308.85	309.63	312.02	313.32	314.49	315.27	315.75

a) Compute the seasonal indices using ratio to moving average method.

b) Obtain the deseasonalised values and then fit a linear trend line to the average annual CO_2 concentrations using least squares method.

c) Convert the annual least-squared trend equation to a monthly trend equation.

d) Use the monthly trend equation and seasonal indices to forecast the CO_2 concentrations for all twelve months of 2021.

e) Plot the original data, deseasonalised data and trend values.

(8+6+5+4+2)

TUTOR MARKED ASSIGNMENT
MSTL-003: Biostatistics Lab

Course Code: MSTL-003

Assignment Code: MSTL-003/TMA/2022

Maximum Marks: 100

Note:

1. **All questions are compulsory.**
2. **Solve the following questions in MS Excel.**
3. **Take the screenshots of the final output/spreadsheet.**
4. **Paste all screenshots in the assignment's booklets with all necessary interpretation and steps.**

Q1(a) A random sample of 440 patients of cardiology department of a hospital was taken and their workout timing and severity of heart disease status were recorded. The following table shows the workout timing and severity of heart disease:

Workout (in minutes)	Severity of Heart Disease				
	Low	Mild	Moderate	High	Very High
No workout	5	13	26	21	23
0- 15	6	15	19	19	21
15 to 30	16	17	14	16	12
30 to 45	18	17	13	11	9
45 to 60	20	19	15	13	7
≥ 60	16	22	6	5	6

Test at 5% level of significance whether workout habit and heart disease are associated with to each other or not.

(b) To study the association between the diabetic patients and their family history of diabetes, the following data were obtained on 70 subjects.

Diabetes in Family	Diabetes in Subject		Total
	Yes	No	
Yes	14	3	**17**
No	3	50	**53**
Total	**17**	**53**	**70**

Which test is appropriate in this situation? Check whether the diabetes runs with generations in families or not at 5% level of significance using appropriate test.

(10+15)

Q2 A researcher is interested to check the relationship between the serum creatinine (in mg/dL) with the weight (in kg) and gender (0 if female and 1 if male). The data were collected from the hospital records to examine the contribution of these variables to serum creatinine. A total of 40 patients were sampled and the data are shown in the following table:

S. No.	Serum Creatinine	Weight	Gender	S. No.	Serum Creatinine	Weight	Gender
1	0.7	46	1	**21**	1.1	55	1
2	1.3	65	1	**22**	0.9	55	0
3	1	59	1	**23**	0.9	62	0
4	1.5	84	0	**24**	1.1	65	0
5	1.7	91	1	**25**	0.8	54	0
6	1.5	78	1	**26**	0.5	45	0
7	1	53	0	**27**	0.6	45	0
8	0.7	49	1	**28**	1	62	0
9	0.5	42	0	**29**	0.5	40	0
10	1.6	87	0	**30**	0.9	58	0
11	1.1	53	1	**31**	1.3	65	1
12	0.8	54	0	**32**	1.1	58	1
13	1.3	65	1	**33**	1.4	67	1
14	1.1	61	1	**34**	0.8	42	1
15	1.3	71	0	**35**	1.6	81	1
16	1.3	71	0	**36**	1.8	92	1
17	1.2	68	0	**37**	1.5	80	0
18	1.3	65	1	**38**	1.7	91	1
19	1	55	1	**39**	0.8	55	0
20	1.2	66	0	**40**	0.5	33	0

(i) Prepare a scatter plot to get an idea about the relationship among the variables.

(ii) Fit a linear regression model and its related analysis at 1% level of significance.

(iii) Does the fitted regression model satisfy the linearity and normality assumptions?

(iv) Also, draw both fitted regression lines on the scatter plot.

(5+5+10+5)

Q3 A hypothetical data of 40 patients on age (in years), weight (in kgs) and systolic blood pressure (in mm/hg) denoting 1 for high SBP and 0 for normal SBP are given in the following table:

S. No.	Age	Weight	SBP	S. No.	Age	Weight	SBP
1	52	60	0	**21**	47	48	0
2	56	68	1	**22**	42	45	0
3	51	54	0	**23**	45	57	0
4	63	74	1	**24**	56	83	1
5	54	62	0	**25**	49	63	0
6	51	67	0	**26**	56	94	1
7	51	66	0	**27**	55	87	1
8	54	65	0	**28**	53	67	0
9	59	71	1	**29**	65	70	1
10	51	89	1	**30**	44	70	0
11	56	72	1	**31**	48	54	0

12	55	72	1	**32**	61	79	1
13	46	57	0	**33**	45	85	0
14	42	54	0	**34**	63	98	1
15	52	63	0	**35**	49	78	0
16	65	67	1	**36**	65	80	1
17	50	67	0	**37**	60	70	1
18	42	53	0	**38**	53	98	1
19	50	68	1	**39**	41	53	0
20	39	55	0	**40**	50	70	1

For this data:

(i) Fit a multiple logistic regression model.

(ii) Test the significance of the individual model coefficients β_1 and β_2 at 5% level of significance.

(iii) Obtain the 95% confidence intervals for β_1 and β_2.

(iv) Determine the Nagelkerke pseudo R-squared.

(11+6+4+4)

Q4 A clinical study was conducted on individuals with advanced stage of Hepatocellular Carcinoma to test three lines of Treatments: T1, T2 and T3. Thirthy-six patients with stage III Hepatocellular Carcinoma who agreed to take part in the experiment were randomly allocated one of three line of Treatments T1, T2 and T3. The primary outcome was mortality, and patients were monitored for up to 60 months (5 years) after recruitment. The data (in months) so obtained are given as follows:

Patient ID	Survival time	Outcome	Treatment	Patient ID	Survival time	Outcome	Treatment
ID001	14	Died	T3	**ID019**	50	Died	T1
ID002	27	Unknown	T1	**ID020**	54	Unknown	T3
ID003	37	Unknown	T3	**ID021**	57	Died	T2
ID004	44	Died	T1	**ID022**	60	Survived	T3
ID005	27	Died	T2	**ID023**	20	Died	T1
ID006	29	Died	T3	**ID024**	22	Unknown	T2
ID007	50	Died	T2	**ID025**	11	Unknown	T2
ID008	31	Died	T1	**ID026**	12	Unknown	T1
ID009	54	Died	T2	**ID027**	57	Unknown	T3
ID010	32	Died	T2	**ID028**	60	Survived	T3
ID011	32	Unknown	T2	**ID029**	44	Died	T1
ID012	60	Unknown	T1	**ID030**	47	Unknown	T2

ID013	2	Unknown	T2	**ID031**	32	Died	T2
ID014	42	Died	T3	**ID032**	34	Died	T1
ID015	42	Unknown	T2	**ID033**	17	Died	T2
ID016	60	Died	T3	**ID034**	6	Died	T1
ID017	60	Survived	T3	**ID035**	50	Unknown	T3
ID018	47	Died	T3	**ID036**	14	Unknown	T2

For this data,

(i) Construct Kaplan and Meier survival curves.

(ii) Test whether there is a significant difference between the survival distributions of the patients under all treatments at 5% level of significance.

(10+15)