ignou
THE PEOPLE'S
UNIVERSITY

Assignment Booklet
for
M.Sc. (Applied Statistics)
(MSCAST)

SCHOOL OF SCIENCES
Indira Gandhi National Open University
New Delhi - 110068

MSCAST

Assignment Booklet

M.SC. (APPLIED STATISTICS)

Valid from January 2024 to December 2024

ASSIGNMENTS

for

Semester-I &II Courses

- MST-011
- MST-012
- MST-013
- MST-014
- MST-015
- MST-016
- MST-017
- MST-018
- MST-019

It is compulsory to submit Assignments before filling the Term-End Examination (TEE) Form

School of Sciences

Indira Gandhi National Open University

Maidan Garhi, New Delhi-110068 (INDIA)

2024

Dear Learner,
Welcome to the M.Sc. (Applied Statistics) Programme.

As per the laid down guidelines of the University, you need to complete the assignment for each theory course. All the questions given in an assignment are compulsory. It is important that you should write the answers to all questions in your own words. You should remember that writing answers to the assignment questions will improve your writing skills and prepare you for the term-end examination.

This booklet includes assignments for the first and second semesters' courses:

MST-011: Real Analysis, Calculus and Geometry

MST-012: Probability and Probability Distributions

MST-013: Survey Sampling and Design of Experiments-I

MST-014: Statistical Quality Control and Time Series Analysis

MST-015: Introduction to R Software

MST-016: Statistical Inference

MST-017: Applied Regression Analysis

MST-018: Multivariate Analysis

MST-019: Epidemiology and Clinical Trials

It is compulsory to submit the assignments within the stipulated time to be eligible for appearing in the term-end examination. You will not be allowed to appear for the term-end examination for a course if you do not submit the assignment for that course within the due date. As per the University guidelines, if you appear in the term-end examination of a course without submitting its assignment, the result of the term-end examination is liable to be cancelled/ withheld.

The assignments constitute the continuous component of the evaluation process and have 30% weightage in the final grading.

Before you write the assignments, first go through the course material and then prepare the assignments carefully by following the instructions pertaining to assignments. Your responses should not be a verbatim reproduction of the textual materials provided for self-learning purposes, but it should be in your own words.

If you have any doubt or problem pertaining to the course material and assignments, contact the concerned Programme in-charge or Academic Counsellor at your Study Centre. If you still have problems, do feel free to contact us at the School of Sciences, IGNOU, New Delhi.

Wishing you all the best to successfully complete the programme.

Programme Team
MSCAST

Email: mscast@ignou.ac.in

INSTRUCTIONS

1. Read the instructions related to assignments given in the Programme Guide.
2. Please note that unless you submit the assignments contained in this booklet within the stipulated time, you would not be permitted to appear for the term-end examination.
3. You are advised to mention the following information on the first page of the assignment response sheet:

NAME: .....

ENROLLMENT NO.....

CYCLE OF ADMISSION:.....

PROGRAMME CODE:.....

ASSIGNMENT CODE:.....

COURSE CODE:.....

COURSE TITLE:.....

REGIONAL CENTRE CODE:.....

STUDY CENTRE:.....

ADDRESS: .....

.....

.....

CONTACT NUMBER:.....

DATE OF SUBMISSION:.....

You are advised to strictly follow the above format. If you do not follow this format, your assignment response sheet will be returned to you, and you will be asked for re-submission.

Note the following points before you start writing the assignments:

- Use only A-4 size paper for writing your responses. Only handwritten assignments will be accepted. Typed or printed copies of the assignments will not be accepted.
- Tie the pages after numbering them carefully.
- Write the question number for each answer.
- All questions are compulsory.
- **Keep a copy of the assignment answer sheets with you before submission for future reference.**
- Answer each assignment on separate sheets.
- It is mandatory to write all assignments neatly in your own handwriting. Write Your Name, Course Code, Enrollment No. and Cycle of admission on all the assignments in bold letters.
- Express your response in your own words. You are advised to restrict your response based on the marks assigned to it. This will also help you to distribute your time in writing or completing your assignments on time.
- The assignment must be submitted at your Study Centre.

You will have to submit your solved assignments at the Study Centre allotted to you before the due date as set by the University.

It is desirable to keep with you a photocopy of the assignment(s) submitted by you.

***You will have to submit the assignments at the Study Centre by the due date as specified by the University.**

Due Date of Submission*: **As specified by IGNOU.**

*Please note that last date of submission may be changed by the University. Please check IGNOU website for updated information regarding due date of assignment submission.

TUTOR MARKED ASSIGNMENT

MST-011: Real Analysis, Calculus and Geometry

Course Code: MST-011

Assignment Code: MST-011/TMA/2023

Maximum Marks: 50

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

1. Solve the following problems.

(a) A ball is thrown in an upward direction. If the variable x represents the velocity of the ball when it strikes the ground. Classify variable x as discrete or continuous. Justify your answer with a proper explanation.

(b) In R we have a built-in data set "trees". A screenshot of the first four rows together with the R code to obtain it is given as follows. To get more detail about this data set you can run ?trees command on R console.

```
> head(trees, 4)
  Girth Height Volume
1   8.3     70   10.3
2   8.6     65   10.3
3   8.8     63   10.2
4  10.5     72   16.4
```

Note that all the three variables of this data set are numeric. So, assuming each row of this data set is a point in 3-dimension. Find the distances between the points corresponding to the first and the third rows using the Manhattan and Chebyshev distance formula.

(c) Find the equation of a line passing through points A(2, 3, 5) and B(5, 8, 9). Also, find the coordinates of a point on this line which is at a distance of 10 units from point A opposite to the side of point B.

(d) Give an example of a set which is convex but not affine. Justify your claim with a proper explanation.

(2 + 3 + 2 + 3)

2. (a) Test the convergence of the series $\frac{3}{5.9.7}x + \frac{5}{10.12.11}x^2 + \frac{7}{15.15.15}x^3 + \frac{9}{20.18.19}x^4 + \cdots,\ x > 0.$

(b) If $f : [0, 5] \to \mathbb{R}$ be a function defined by $f(x) = x^2 + 2x + 1,\ x \in [0, 5]$. Show that f is Riemann integrable using both definitions. Also, verify that the results of both definition match.

(10 + 10)

3. (a) Evaluate the integral $\iint_D e^{4x+5y}\,dx\,dy$, where $D = \{(x, y) : x \geq 0,\ y \geq 0,\ x + y \leq 1\}$, by considering D as a region of Type I and then as a region of Type II.

(b) Evaluate the integral $\int_0^4 \frac{dx}{\sqrt{4x - x^2}}$ using beta and gamma functions. **(14 + 6)**

TUTOR MARKED ASSIGNMENT

MST-012: Probability and Probability Distributions

Course Code: MST-012

Assignment Code: MST-012/TMA/2023

Maximum Marks: 100

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

1. (a) Suppose two friends Anjali and Prabhat trying to meet for a date to have lunch say between 2 pm to 3 pm. Suppose they follow the following rules for this meeting:

- Each of them will reach either on time or 10 minutes late or 20 minutes late or 30 minutes late or 40 minutes late or 50 minutes late or 1 hour late. All these arrival times are equally likely for both of them.
- Whoever of them reaches first will wait for the other to meet only for 10 minutes. If within 10 minutes the other does not reach, he/she leaves the place and they will not meet.

Find the probability of their meeting.

(b) In the study learning material (SLM), you have seen many situations where Poisson distribution is suitable and discussed some examples of such situations. Create your own example for a situation other than those that are discussed in SLM. If you denote your created random variable by X then find the probability that X is less than 2.

(8 + 17)

2. (a) In an election there are two candidates. Being a statistician, you are interested in predicting the result of the election. So, you plan to conduct a survey. Using the learning skill of this course answer the following question. How many people should be surveyed to be at least 90% sure that the estimate is within 0.03 of the true value?

(b) Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables from the uniform distribution on the interval [12, 20]. If $\overline{X}_n$ denotes the sample mean of the first n random variables of the sequence $X_1, X_2, X_3, \ldots$, and $\overline{X}_n \xrightarrow{p} a$. Find the value of a. **(15 + 10)**

3. Explain the procedure of assigning probability in a continuous world of probability theory. **(25)**

4. (a) If $X \sim \text{Gamma}(\theta, \alpha)$ and $Y \sim \text{Gamma}(\theta, \beta)$ be two independent gamma distributions and $U = \frac{X}{X+Y}$ and $V = X + Y$ then find the distribution of U.

(b) A hospital specialising in heart surgery. In 2022 total of 2000 patients were admitted for treatment. The average payment made by a patient was Rs 1, 50,000 with a standard deviation of Rs 25000. Under the assumption that payments follow a normal distribution, answer the following questions.

(i) The number of patients who paid between Rs 1,40,000 and Rs 1,70,000.

(ii) The probability that a patient bill exceeds Rs 1,00,000.

(iii) Maximum amount paid by the lowest paying one-third of patients.

(10 + 15)

TUTOR MARKED ASSIGNMENT

MST-013: Survey Sampling and Design of Experiments-I

Course Code: MST-013

Assignment Code: MST-013/TMA/2023-24

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) $V_{opt}(\bar{x}_{st})$ lies between $V_{prop}(\bar{x}_{st})$ and $V_{Random}(\bar{x}_{st})$.

(b) The total number of all possible samples of size 3 without replacement from a population of size 7 is 21.

(c) While analysing the data of a 5 × 5 Latin Square design the d.f. for ESS is equal to 16.

(d) In a Two-way Analysis of Variance test with 5 observations per cell having 4 blocks and 4 treatments the degree of freedom for the total variation is 64.

(e) The probability of selection of a sample of n from the population by SRSWOR is $1/N$.

2(a) A sample of 100 employees is to be drawn from a population of collages A and B. The population means and population mean squares of their monthly wages are given below:

Village	N_i	$\bar{X}_i$	S_i^2
Collage A	400	60	20
Collage B	200	120	80

Draw the samples using Proportional and Neyman allocation techniques and compare. Obtain the sample mean and variances for the Proportional Allocation and SRSWOR for the given information. Then Find the percentage gain in precision of variances of sample mean under the Proportional Allocation over the that of SRSWOR.

(15)

(b) A population consists of 10 villages with a total of 212 households. The second column of the accompanying table shows the number of households corresponding to each village. Select a PPS with replacement sample of 6 villages by using the Cumulative Total method:

Village	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
No. of Households	35	28	20	25	30	19	10	12	18	15

(10)

3(a) In a population of size N = 5, the values of the population characteristics are 1, 3, 5, 7, 9, a sample of size 2 is drawn. Verify that $\bar{y}$ is an unbiased estimate of $\bar{Y}$ and $V(\bar{y})$ is equal to

$$V(\bar{y}) = \frac{N-n}{N.n} S^2 .$$ **(12)**

(b) In order to compare the mileage yields of 3 kinds of Gasoline, several tests were run, and the following results were obtained:

Gasoline A:	19	21	20	18	21	21
Gasoline B:	23	20	22	20	24	23
Gasoline C:	20	17	21	19	20	17

Carry out the Analysis of Variance test and test whether there is significant differences between the average mileage of 3 kinds of gasoline at 5% level of significance.

(8)

4(a) A manurial trial with six levels of Farmyard Manure (FYM) was carried out in a randomised block design with 4 replications at the experimental station Junagarh with a new study the rate of decomposition of organic matters in soil and its synthetic capacity in soil on cotton crop. The yield per plot in kg for different levels of FYM and replications is given below:

Cotton Yield Per Plot (in Kg)

Levels of FYM	**Replications**			
	I	**II**	**III**	**IV**
1	6.90	4.60	4.40	4.81
2	6.48	5.57	4.28	4.45
3	6.52	7.60	5.30	5.30
4	6.90	6.65	6.75	7.75
5	6.00	6.18	6.50	5.50
6	7.90	7.57	6.80	6.62

Carry out the Analysis of Variance test and draw the conclusions. **(15)**

(b) For the given data the yield of the treatment 2 in 3rd block is missing. Estimate the missing value and analyse the data.

Treatments	**Blocks**			
	I	**II**	**III**	**IV**
1	105	114	108	109
2	112	113	Y	112
3	106	114	105	109

(10)

5(a) The following data is the data pertaining to a feeding trial on sheep. Treatments

A: Grazing only
B: Grazing + Maize Supplements
C: Grazing + Maize + Protein Supplement P_1
D: Grazing + Maize + Protein Supplement P_2
E: Grazing + Maize + Protein Supplement P_3

Layout and Wool Yield (100 gm) is given as:

32(D)	33(E)	30(C)	28(B)	24(A)
51(C)	45(D)	41(A)	45(E)	29(B)
41 (E)	29 (A)	24 (B)	36 (D)	35 (C)
38 (B)	39(C)	42(E)	23(A)	37(D)
38(A)	24(B)	21(D)	29(C)	26(E)

Analyse the design with appropriate method and calculate the Critical Difference for the treatment mean yield. **(15)**

(b) In a class of Statistics, total number of students is 30. Select the linear and circular systematic random samples of 12 students. The age of 30 students is given below:

Age:	22	25	22	21	22	25	24	23	22	21	20
	21	22	23	25	23	24	22	24	24	21	20
	23	21	22	20	20	21	22	25			

(5)

TUTOR MARKED ASSIGNMENT

MST-014: Statistical Quality Control and Time Series

Course Code: MST-014

Assignment Code: MST-014/TMA/2023

Maximum Marks: 100

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

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

(i) The R- chart is suitable when subgroup size is greater than 10.

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

(iii) If the effect of summer and winter is not constant on the sale of AC then we use the additive model of the time series.

(iv) If a researcher wants to find the relationship between today's unemployment and that of 5 years ago without considering what happens in between then the partial autocorrelation is the better way in comparison to autocorrelation.

(v) 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.

(b) Differentiate between the autoregressive and moving average models of time series. **(10)**

2(a) 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. Find

(i) Probability of accepting a lot of incoming quality 0.04.

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

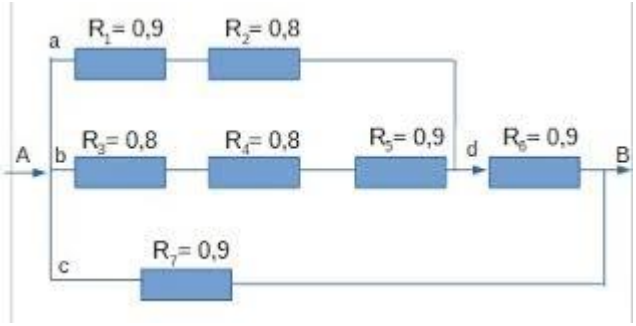
(iii) Average total inspection (ATI). **(6+2+2)**

(b) 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? **(2+8)**

3(a) A system has seven independent components and reliability block diagram of it shown as follows:



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

(b) 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. **(10)**

4. 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 comments about the process. If process is out of control, then calculate the revised control limits.

(ii) Construct the CUSUM chart when the process is under control and draw the conclusion about the process.

(iii) If the specification limits as the 8±2, then calculate the process capability index C_{pk} and impetrate the result.

(iv) Also find the percentage of calls lie outside the specification limits assuming that calls follow the normal distribution.

(20)

5(a) Consider the time series model

$$y_t = 10 + 0.5y_{t-1} - 0.8y_{t-2} + \varepsilon_t$$

where $\varepsilon_t \sim N[0,1]$

(i) Is this a stationary time series?

(ii) What are the mean and variance of the time series?

(iii) Calculate the autocorrelation function.

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

(b) The marketing manager of a company recorded the number of mobiles sold quarterly for which are given in the following table:

Quarter / Year	Q_1	Q_2	Q_3	Q_4
2018	48	41	60	65
2019	58	52	68	74
2020	60	56	75	78

(i) Find the quarterly seasonal indexes for the mobile sold using the ratio to trend method.

(ii) Do seasonal forces significantly influence the sale of mobile? Comment.

(iii) Also find the deseasonalised values. **(10)**

TUTOR MARKED ASSIGNMENT

MST-015: INTRODUCTION TO R SOFTWARE

Course Code: MST-015
Assignment Code: MST-015/TMA/2023
Maximum Marks: 50

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

1. Attempt the following:

(a) Write the output of the following statements:

(i) `rep(x=c(T,F,T,F), times=c(2,1,2,3))`

(ii) `5%/%3; diag(3)`

(b) Differentiate between the use of the `sep` and `collapse` arguments of the `paste()` function.

(c) Write R commands to create a bar plot of the following data by using arguments of the used function for filling up the bars and to give labels to the axis:

5, 10, 8, 7, 8, 5, 8, 7, 5, 8, 9, 6, 8, 8, 8

(d) Check whether the given loop is finite or infinite. If infinite, do the necessary changes in the written loop to make it finite.

```
x<-0
repeat{
print(x^2)
x<-x+1
if(x<5) print(x) }
```

(1×3+2=5)

2. The following data relates to the number of items produced per shift by two workers for a number of days.

Worker A	19	22	24	27	24	18
Worker B	26	37	40	35	NA	NA

(a) Write R command to create a list named `LT` with worker's data. Also, after creating the list, do the following tasks:

(i) Use a suitable loop function to compute the mean of number of items produced by each worker in a single line command.

(ii) Extract the worker A data from it by using two different approaches.

(b) Write R command to create a data frame named `DF` with worker's data and do the following tasks:

(i) Use suitable function to remove `NA` from the data and then create a scatter plot.

(ii) Write the known data obtained in step (i) to a .txt file named "WORK".

(8+7=15)

3. Write R commands to:

(a) Create a function to compute ranks (in case of tied ranks) of the given data.

(b) Create a date object named `Ddata` consisting of the following dates.

26Jan2023, 15Aug2023, 02Oct2023, 05Sep2023

(c) Create an array of two dimension with following elements.

$$\begin{pmatrix} -2 & 4 \\ 0 & 1 \\ 9 & 2 \end{pmatrix}$$

Also, extract the row shown in the rectangular box.

(d) Create the graph of the following function.

$$f(x) = |x|,\ -5 \le x \le 5$$

(4x3+3=15)

4. **(a)** Create following two matrices A and B with following elements.

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix},\ B = \begin{pmatrix} -3 & 1 \\ 2 & -1 \end{pmatrix}$$

Write R commands to do the following tasks:

(i) Multiply the two matrices.

(ii) Combine the two matrices row-wise.

(iii) Create a function that computes the following expression:

A^2+3*B

(b) Create a data frame named `RData` consisting of the following data:

x	y	z	w
0.04	0.16	0.53	A
0.82	0.87	0.84	A
0.32	0.65	0.99	A
0.39	0.83	0.42	A
0.31	0.93	0.78	A
0.83	0.31	0.41	A
0.73	0.74	0.88	A
0.32	0.39	0.50	B
0.60	0.85	0.68	B
0.65	0.28	0.86	B
0.55	0.95	0.77	B
0.53	0.35	0.32	B
0.91	0.01	0.91	B
0.84	0.81	0.37	B

Write R commands to:

(i) Compute the group wise means of `x`, `y` and `z` according to the groups defined by `w` column using apply family function.

(ii) Sort `RData` according the `y` column of it.

(8+7=15)

MST-017: Applied Regression Analysis

Course Code: MST-017

Assignment Code: MST-017/TMA/2024

Maximum Marks: 100

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

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

(i) We define three indicator variables for an explanatory variable with three categories.

(ii) If the coefficient of determination is 0.833, the number of observations and explanatory variables are 12 and 3, respectively, then the Adjusted R^2 will be 0.84.

(iii) For a simple regression model fitted on 15 observations, if we have $h_{ii} = 0.37$, then it is an indication to trace the leverage point in the regression model.

(iv) In a regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, if H_0: $\beta_1 = 0$ is not rejected, then the variable X_1 will remain in the model.

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

(b) Write a short note on the problem of multicollinearity and autocorrelation. **(10)**

2(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 regression model fitted on serum cholesterol and systolic blood pressure.

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

(5+15)

3. For the data given in **Question 2(b)**, obtain the followings:

(i) Diagonal of the hat matrix and also check the leverage points if any.
(ii) Cook's Distances, DFFITS and DFBETAS. Also verify the influence points if any.

(8+12)

4 A company conducted a study on its employees to see the relationship of several variables with an employ's IQ. For this purpose, fifteen employees were selected and an IQ as well as five different personality tests were given to them. Each employ's IQ was recorded along with scores on five tests. The data are shown in the following table:

Employee	Test 1	Test 2	Test 3	Test 4	Test 5	IQ
1	83	80	78	77	67	99
2	73	85	67	80	63	92
3	81	80	71	81	68	94
4	96	86	82	83	56	99
5	84	73	75	75	68	94
6	72	74	71	67	59	79
7	84	79	84	84	69	97
8	54	86	61	69	53	92
9	86	85	79	78	76	94
10	42	71	60	80	56	86
11	83	72	72	78	74	98
12	63	86	65	85	56	83
13	69	76	64	85	61	98
14	81	84	65	79	64	96
15	50	85	71	65	75	76

Determine the most appropriate regression model for the employee's IQ using stepwise approach at 5 % level of significance and interpret the results. Does the final regression model satisfy the linearity and normality assumptions? **(20)**

5. 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)

TUTOR MARKED ASSIGNMENT

MST-016: Statistical Inference

Course Code: MST-016

Assignment Code: MST-016/TMA/2024

Maximum Marks: 100

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

1. (a) State whether the following statements are **True** or **False**. Give reason in support of your answer:

(i) If X_1, X_2, X_3, X_4 and X_5 is a random sample of size 5 taken from an Exponential distribution, then estimator T_1 is more efficient than T_2.

$$T_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5},\ T_2 = \frac{X_1 + 2X_2 + 3X_3 + 4X_4 + 5X_5}{15}$$

(ii) 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.

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

(iv) If the probability density function of a random variable X follows F-distribution is

$$f(x) = \frac{1}{(1+x)^2}, x \geq 0$$

then degrees of freedom of the distribution will be (2,2).

(v) A patient suffering from fever reaches to a doctor and suppose the doctor formulate the hypotheses as

H_0: The patient is a chikunguniya patient

H_1: The patient is not a chikunguniya patient

If the doctor rejects H_0 when the patient is actually a chikunguniya patient, then the doctor commits type II error.

(5×2=10)

(b) Describe the various forms of the sampling distribution of ratio of two sample variances.

(10)

2(a) A baby-sister has 6 children under her supervision. The age of each child is as follows:

Child	Age (in years)
Sonu	2
Lavnik	4
Chiya	3
Amam	3
Avishi	4
Ridhi	5
Sidhi	3

(i) What is the form of population of age of children?

(ii) Prepare the sampling distribution of sample mean when sample size is 2.

(iii) Is the shape of the sampling distribution normal?

(iv) Calculate the mean and standard error of the sampling distribution. **(20)**

3. The department of transportation has mandated that the average speed of cars on interstate highways be no more than 70 km per hours in order. To check that the people follow it or not, a researcher took a random sample of 186 cars and found that the average speed was 72 km per hours with a standard deviation 0.6 km per hours.

(a) Construct the interval around the sample mean that would contain the population mean 95% of the time.

(b) If the researcher wants to test that the true mean speed on its highways is 70 km per hours or less with 95% confidence then

(i) State null and alternative hypotheses.

(ii) Name the test which is suitable in this situation and why?

(iii) Calculate the value of test statistic and critical value.

(iv) Draw the conclusion on the basis of the applied test. **(20)**

4(a) A sample of 500 shops was selected in a large metropolitan area to determine various information concerning consumer behaviour. One question, among the questions, asked, was "Do you enjoy shopping for clothing?" Out of 240 males 136 answered yes. Out of 260 females, 224 answered yes. Find 95% confidence interval for the difference of the proportions for enjoys shopping for clothing. **(10)**

(b) An engineer conducted an experiment to compare two metals: iron and copper, as bonding agents for an alloy material. Components of the alloy were bonded using the metals as bonding agents, and the pressures required to break the bonds were measured. The data for the breaking pressures are given in the following table:

S. No.	Breaking Pressure	
	Iron	Copper
1	72.7	73
2	69.6	67.2
3	83.4	75.3
4	78.9	61.4
5	75	74
6	71.6	69.5
7	85.7	69.8
8	73.5	73.8
9	70.4	68
10	84.2	76.1

If the breaking pressures for both iron and copper are normally distributed, are the variances of the distributions of the breaking pressure of iron and copper equal at 5 % level of significance? **(10)**

5. (a) Complete the following table, one is done for you:

S. No.	Test For	Name of the Test	Test Statistic
1	Population mean when population variance is known and population is normal	Z-test	$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$
2	Population mean when population variance is unknown and population is normal		
3	Difference of two population means when samples are paired, and population of differences follows normal distribution.		
4	Difference of two population means when samples are independent, and population of differences follows normal distribution.		
5	Population variance when the population is normal distributed		
6	Population variance when the population is not normal distributed		

(10)

(b) Describe the following:

(i) Curve of F-distribution

(ii) Mean Squared Error

(10)

TUTOR MARKED ASSIGNMENT

MST-018: Multivariate Analysis

Course Code: MST-018
Assignment Code: MST-018/TMA/2024-25
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) The covariance matrix of random vectors $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ is symmetric.

(b) If $\underset{\sim}{X}$ is a p-variate normal random vector, then every linear combination $\underset{\sim}{c}'\underset{\sim}{X}$, where $\underset{\sim}{c}_{p\times 1}$ is a scalar vector, is also p-variate normal vector.

(c) The trace of matrix $\begin{pmatrix} 3 & -2 \\ -2 & 6 \end{pmatrix}$ is 9.

(d) If a matrix is positive definite then its inverse is also positive definite.

(e) If $\underset{\sim}{X} \sim N_2\left(\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$ and $\underset{\sim}{Y} \sim N_2\left(\begin{pmatrix} -1 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$, then

$$\underset{\sim}{X} + \underset{\sim}{Y} \sim N_2\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

2 (a) Let $\underset{\sim}{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ has the following joint density function

$$f(x_1, x_2) = \begin{cases} 4x_1x_2 & , 0 < x_1 < 1, 0 < x_2 < 1, \\ 0 & , \text{otherwise.} \end{cases}$$

Find the marginal distributions, mean vector and variance-covariance matrix. Also, comment on the independence of X_1 and X_2.

(b) Let $\underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X}^{(1)} \\ \underset{\sim}{X}^{(2)} \end{pmatrix} \sim N_4\left(\underset{\sim}{\mu}, \Sigma\right)$, where $\underset{\sim}{\mu} = \begin{pmatrix} -4 \\ 1 \\ 4 \\ 0 \end{pmatrix}$ and $\Sigma = \left(\begin{array}{cc|cc} 2 & 2 & 3 & 0 \\ 2 & 1 & 2 & 0 \\ \hline 3 & 2 & 2 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right)$. Find the

$E\left(\underset{\sim}{X}^{(2)} \middle| \underset{\sim}{X}^{(1)} = \underset{\sim}{x}^{(1)}\right)$ and $\text{Cov}\left(\underset{\sim}{X}^{(2)} \middle| \underset{\sim}{X}^{(1)} = \underset{\sim}{x}^{(1)}\right)$. **(10×2=20)**

3 (a) Let $\underset{\sim}{X}$ be a 3-dimensional random vector with dispersion matrix

$$\Sigma = \begin{pmatrix} 4 & -2 & 0 \\ -2 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Determine the first principal component and the proportion of the total variability that it explains.

(b) Let $\underset{\sim}{X} \sim N_4(\mu, \Sigma)$, where $\underset{\sim}{\mu} = \begin{pmatrix} 3 \\ -2 \\ 1 \\ -2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & -2 & 5 \end{pmatrix}$. Check the independence of the (i) X_2 and X_1 (ii) (X_2, X_4) and (X_1, X_3) (iii) (X_1, X_2) and (X_3, X_4).

(10×2=20)

4 (a) Consider the following data of 11 samples on 8 variables by Anscombe, Francis J. (1973):

x_1	x_2	x_3	x_4	y_1	y_2	y_3	y_4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.10	5.39	12.50
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

If the vector $\underset{\sim}{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$ and $\underset{\sim}{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$, then obtain the sample covariance matrix between $\underset{\sim}{x}$ and $\underset{\sim}{y}$.

Source: Anscombe, Francis J. (1973). Graphs in statistical analysis. *The American Statistician*, **27**, 17–21. doi: 10.2307/2682899.

(b) Obtain the maximum likelihood estimator of the mean vector and variance-covariance matrix of the multivariate normal distribution.

(15+10=25)

5 (a) Define the following:

(i) Covariance Matrix

(ii) Mahalanobis D^2

(iii) Hotelling's T^2

(iv) Clustering

(v) Relationship between (ii) and (iii).

(b) If $\underset{\sim}{X} \sim N_3(\underset{\sim}{\mu}, \Sigma)$ with $\underset{\sim}{\mu} = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 5 & 3 & 0 \\ 3 & 3 & -2 \\ 0 & -2 & 5 \end{pmatrix}$. Then find the joint distribution of $X_1 + 2X_2$, $2X_1 - X_2$ and X_3.

(15+10=25)

TUTOR MARKED ASSIGNMENT

MST-019: Epidemiology and Clinical Trials

Course Code: MST-019

Assignment Code: MST-019/TMA/2024

Maximum Marks: 50

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

1. (a) In the natural history of a disease define: total preclinical phase, detectable pre-clinical phase and clinical phase. Suppose on 10 am on 25.01.2010 a disease A onset you biologically. Suppose test of the disease A can detect it exactly after completion of 1000 days of biologically onset. Suppose signs and symptoms develop exactly after completion of 1010 days of biologically onset. Suppose you consult doctor exactly after 1015 days of biologically onset of the disease. Suppose outcome the treatment is cure. What is duration of (i) total preclinical phase (ii) detectable pre-clinical phase (iii) clinical phase.

(b) If p denotes proportion and q denotes odds then prove that $q = \frac{p}{1-p}$. Find the range of q. If the odds of smokers in a study are 0.25 then find the proportion of smokers in the study. Assume that each subject of the study is either a smoker or non-smoker.

(c) A trial is conducted in which some people with disease X were randomly allocated into two groups. First group was advised to do some morning walk for 30 minutes and take light food each day and second group was given one injection and one 100 mg tablet once a day to control disease X. The injection can cause loose motion in some cases and 100 mg tablet has no side effect. At the end of two months, 90% of group I and 80% of group II had recovered from disease X.

i) What are the regimens for group I and group II in this trial?

ii) What are the efficacies in group I and group II?

iii) What are the safety issues in group I and group II in this trial?

(4 + 4 + 2)

2. (a) What is the basic difference between (a) cross-sectional (b) cohort and (c) case control study designs. Explain with suitable examples of each design.

(b) RTPCR test is applied on 300 covid patients and 200 non-covid patients. The results of the test are shown as follows.

		Disease status		
		Yes (D^+)	No (D^-)	Total
Result of RTPCR test	T^+	480	40	520
	T^-	120	360	480
	Total	600	400	1000

What are the sensitivity and specificity of the test? Also, determine the positive and negative values of the test.

(15 + 5)

3. (a) In usual notations you are given the following information

$\delta = 0.05,\ \pi_1 = 0.75,\ \pi_2 = 0.65,\ \bar{\pi} = 0.7,\ \alpha = 0.05,\ \beta = 0.20$

Find the sample size. If power of the test is 95% instead of 80%, then find new sample size.

(b) If you are told that haemoglobin (Hb) level in blood measures the iron content. The normal level in healthy people is around 15 g/dL. Most Indian women have less and some have even less than 8 g/dL. They are called anemic. Iron supplementation is given to increase this level. Suppose one supplementation increase the mean Hb level in anemics by 3.2 g/dL, after one month of use and the other by 3.6 g/dL, in an equivalence trial on 500 women each. The respective SD's of the of the increases are 0.52 and 0.72 g/dL. The doctors determine that the supplementations can be clinically equivalent if the difference between the increases by two supplementations does not exceed 0.52 g/dL. Can these two supplementations be considered clinically equivalent at 5% level of significance?

. **(10 + 10)**