INDIAN STATISTICAL INSTITUTE

BStat (Hons) I Year

Final Examination

2024-25 Second Semester

Statistical Methods II

Duration: 3 hours | 5 May 2025 | Maximum marks: 100

This examination is closed book, closed notes. Calculators are NOT allowed during the examination. Please answer all the questions.

1. A geo-referenced map is a digital map or image that has been aligned with real-world geographic coordinates, allowing it to be used in a Geographic Information System (GIS) or other mapping software. This means that each point on the map or image has a corresponding location on the Earth's surface. Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are the pairs of horizontal and vertical pixel coordinates of a selection of n points on a digital image that has not been geo-referenced, and $(T_1, N_1), (T_2, N_2), \ldots, (T_n, N_n)$ are the pairs of latitude and longitude of the same n points on a geo-referenced map. Assume that the latitude-longitude pair (T, N) of a point with pixel coordinates (X, Y) should ideally be

$$\begin{pmatrix} T \\ N \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix},$$

for some unknown parameters a, b, c, d, e and f. Explain, as specifically as possible, how multiple regression can be used to estimate these parameters and how those parameters can be used to obtain the latitude-longitude pair of any pixel in the said image. [13+2=15]

2. A few lines of R code are given below.

```
densty <- function(x) (x^1.5)*exp(-x)/gamma(2.5)
xlist <- NULL
for (i in 1:100) {
    U1 <- runif(1); U2 <- runif(1); Y <- qexp(U1, rate=.2)
    threshold <- densty(Y)/(4*dexp(Y,rate=0.2))
    if (U2<threshold) xlist <- c(xlist,Y)
}
xlist
```

   (a) What has the programmer tried to do with the code?

   (b) Assuming that there was no error in the coding, what do you expect the length of the output to be?

   (c) Please check carefully if there is any error in the method or the code and report what exactly you have checked and found. [5+5+10=20]

(Please turn over)

3. The image below is a partial snapshot from the R documentation on a multivariate data set.

BostonHousing {mlbench} R Documentation

Boston Housing Data

Description

Housing data for 506 census tracts of Boston from the 1970 census. The dataframe BostonHousing contains the original data by Harrison and Rubinfeld (1979). the dataframe BostonHousing2 the corrected version with additional spatial information (see references below).

Usage

```
data(BostonHousing, package="mlbench")
data(BostonHousing2, package="mlbench")
```

Format

The original data are 506 observations on 14 variables, medv being the target variable:

crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	nitric oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per USD 10,000
ptratio	pupil-teacher ratio by town
b	$1000(B - 0.63)^2$ where B is the proportion of blacks by town
lstat	percentage of lower status of the population
medv	median value of owner-occupied homes in USD 1000's

The data set was analysed with the R code reproduced below.

```
library(mlbench)
data("BostonHousing")
lmedv <- log(BostonHousing$medv)
lmbh <- lm(lmedv~crim+zn+chas+nox+rm+dis+rad+tax, data=BostonHousing)
summary(lmbh)
```

A part of the resulting output is reproduced below.

```
    Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7851908  0.1381711  12.920  < 2e-16 ***
crim         -0.0149619  0.0014733 -10.155  < 2e-16 ***
zn            0.0007408  0.0005807   1.276    0.203
chas1         0.1933774  0.0455130   4.249 2.56e-05 ***
nox          -0.7932819  0.1257563  -6.308 6.22e-10 ***
rm            0.2739415  0.0175257  15.631  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2555 on 500 degrees of freedom
Multiple R-squared:  0.613,Adjusted R-squared:  0.6092
F-statistic: 158.4 on 5 and 500 DF,  p-value: < 2.2e-16
```

Interpret the different parts of the output in plain language, with specific reference to the variables as described in the documentation and the code. [20]

(Please turn over)

6. Consider the following data set.

   1.146, 2.224, 1.193, 1.015, 1.617, 1.588, 1.710, 2.055.

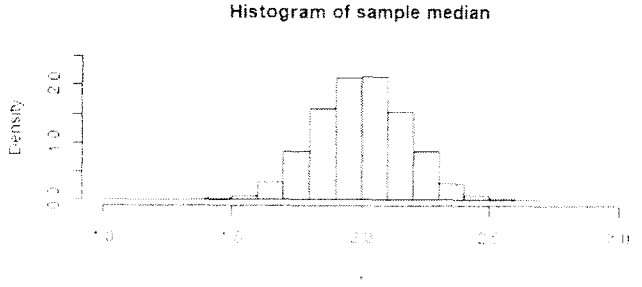
   It has been claimed that the data is a sample from the population with probability density function

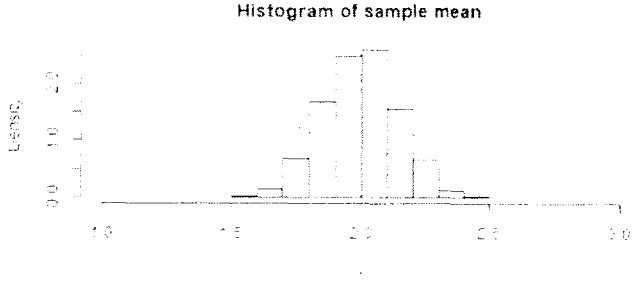
$$f(x) = \begin{cases} \frac{1}{x^2} & \text{for } x \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

   Draw a free hand sketch of the QQ plot for checking the fit of the hypothesized distribution to this data set and comment on the fit. [15+5=20]

7. Give an example of a distribution for which the Maximum Likelihood Estimator of a parameter, computed a sample from that distribution, differs from the method of moments estimator of that parameter. [5]

4. A student of Statistical Methods II was asked to simulate repeated samples of the form $(X_1, X_2, \ldots, X_{100})$ such that $X_1 - 3, X_2 - 3, \ldots, X_{100} - 3$ are independent and identically distributed with the (central) student's t distribution having 3 degrees of freedom. She generated the data, computed the sample mean and the sample median for each simulated sample of size 100, and drew the following histograms from 1000 simulation runs.





Scrutinize the plots and assess her work, with reasons. [10]

5. Consider the following summary of a count data set with sample size 1000.

Value of count	0	1	2	3	4	Greater than 4	Total
Frequency	136	240	281	186	99	58	1000

A student expects that the Poisson distribution will fit the data well. Accordingly, he computes the sample mean of the original data (sample size 1000) and uses that mean to compute the cell probabilities according to the Poisson distribution. The sample mean is 2.061 and the probabilities turn out to be 0.1272,0.2623,0.2704,0.1859,0.0958 and 0.0584. He then uses the following R code to carry out the chi-squared test for the goodness of fit of the Poisson distribution and obtains the result shown below.

```
> chisq.test(c(136,240,281,186,99,58),
                  p = c(0.1272,0.2623,0.2704,0.1859,0.0958,0.0584))
Chi-squared test for given probabilities
data:  c(136, 240, 281, 186, 99, 58)
X-squared = 3.0299, df = 5, p-value = 0.6954
```

(a) Explain how the three items reported on the last line of the output are calculated from the given data, without doing the actual computations.

(b) Is there any deficiency in the analysis? Comment. [6+4=10]

(Please turn over)