BStat (Hons) II Year

2025-26 Second Semester

Statistical Methods II

Mid-semester Examination

Duration: 2 hours 30 minutes 11 September 2025 Maximum marks: 100

This examination is open book, open notes. Access to any electronic device is not allowed during the examination. Answer all questions. All answers should be simplified as much as possible.

1. Suppose $X_1, X_2, \ldots, X_n$ is a sample of size n from the normal distribution $N(\sigma, \sigma^2)$. Consider the estimator $\hat{\sigma}_c^2 = c\bar{X}^2$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

   (a) For what value of c is $\hat{\sigma}_c^2$ an unbiased estimator of σ^2? [3]

   (b) For what value of c does $\hat{\sigma}_c^2$ have the smallest possible mean squared error? [5]

   (c) Derive the Cramer-Rao lower bound for the variance of an unbiased estimator of σ^2. [5]

   (d) Does the variance of the unbiased estimator of σ^2 obtained in part (a) reach the bound obtained in part (c), either for finite n or as n goes to infinity? Explain. [4+1]

   (e) Can you suggest another estimator with variance reaching the Cramer-Rao lower bound as n goes to infinity? Explain. [1+1]

   [Total 20]

2. Consider the R code given below.

```
n <- 201; sigma <- 2; theta <- 1
delta <- 2 * rbinom(n, size = 1, prob = 0.5) - 1
epsilon <- rexp(n, rate = 1/sigma) * delta
y <- theta + epsilon
thetaold <- 1; thetanew <- 2
while (abs(thetanew - thetaold) > 1e-10 * thetaold) {
  thetaold <- thetanew
  thetanew <- sum (y/abs(y - thetaold)) / sum (1/abs(y - thetaold))
  }
thetanew
```

   (a) Describe the distributions of `epsilon` and `y` as per the code, with justification. [3+1]

   (b) The while loop implements an Iteratively Reweighted Least Squares (IRLS) procedure. Identify the parameter being estimated and describe it as an attribute of the distribution of `y` mentioned in part (a). [2]

   (c) Identify, with justification, the criterion minimized through the above procedure. [7]

   (d) Describe (in words) the stopping criterion implemented in the code. [2]

   (e) What is the probability that the estimation error is less than 0.05? Explain. [5]

   [Total 20]

3. Write an R code to obtain a nonparametric bootstrap estimate of the standard error of the sample median. [10]

Please turn over

4. You have the sample $X_1, X_2, \ldots, X_n$ from the distribution of X given θ, which is $N(\theta, \sigma^2)$ for some fixed and known σ^2. The prior distribution of θ is $N(\mu, \tau^2)$. What is the posterior distribution of θ given the data? Explain. [10]

5. Let $Y_1, Y_2, \ldots, Y_n$ be iid samples from the mixed normal distribution with pdf

$$f(y|p, \overset{\sigma_1^2}{\cancel{\mu_1, \mu_2}}, \sigma_2^2) = \frac{p}{\sigma_1}\phi\left(\frac{y}{\sigma_1}\right) + \frac{1-p}{\sigma_2}\phi\left(\frac{y}{\sigma_2}\right), \quad -\infty < y < \infty,$$

for $0 < p < 1$, $\sigma_1, \sigma_2 > 0$, where ϕ is the standard normal pdf. The MLEs of the parameters p, σ_1^2 and σ_2^2 have to be obtained from the data.

   (a) Explain why the three parameters are not uniquely identifiable from the data. State a condition on the parameters that makes them identifiable. [1+1]

   (b) Write down an iterative step of the steepest ascent method for obtaining the MLEs of the three parameters. [5]

   (c) Formulate a suitable 'complete data' version of this problem for the purpose of using the EM algorithm to obtain the MLEs of the three parameters. [4]

   (d) Write down the E-step of the EM algorithm explicitly. [4]

   (e) Write down the M-step of the EM algorithm explicitly. [3]

   (f) Suggest how you can obtain a suitable set of 'initial iterates' of the parameters for the 'complete data' version of the problem. [2]

[Total 20]

6. Consider the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where the binary response y_i is assumed to follow the logistic regression model

$$y_i|x_i \overset{\text{ind}}{\sim} Bernoulli\left(\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right), \quad i = 1, 2, \ldots, n.$$

The parameters of the model are sought to be estimated by Fisher's scoring method with iterative steps of the form

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + \boldsymbol{I}^{-1}(\boldsymbol{\beta}) \sum_{i=1}^{n} \left(y_i - \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}\right) \boldsymbol{x}_i \Bigg|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(r)}}.$$

   (a) Explain the notations $\boldsymbol{\beta}$ and $\boldsymbol{x}_i$ and derive an expression for $\boldsymbol{I}(\boldsymbol{\beta})$ specifically for the present problem. [1+1+2]

   (b) Express $\boldsymbol{I}(\boldsymbol{\beta}^{(r)})$ in the form $\boldsymbol{X}^T \boldsymbol{W}^{(r)} \boldsymbol{X}$ for a suitable matrix $\boldsymbol{X}$ having two columns and a diagonal weight matrix $\boldsymbol{W}^{(r)}$, which you have to identify. [5]

   (c) If the iterative step is written as

$$\boldsymbol{\beta}^{(r+1)} = \left(\boldsymbol{X}^T \boldsymbol{W}^{(r)} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}^{(r)} \boldsymbol{z}^{(r)},$$

   explain what $\boldsymbol{z}^{(r)}$ must be. [4]

   (d) The above model is fitted to the data of the 23 initial flights of the space shuttle Challenger, where x_i and y_i are, respectively, the launch-time air temperature in Fahrenheit and the indicator of at least one failure of a primary O-ring in the ith flight. The fitted values of β_0 and β_1 are 15.0429 and –0.2322, respectively. Interpret β_1. What is the probability of at least one primary O-ring failure at $31°F$? By what factor will the odds of at least one failure change if the temperature increases to $60°F$? [2+3+2]

[Total 20]