BStat (Hons) II Year

2025-26 First Semester

Statistical Methods III

Final Examination

Duration: 3.5 hours | 25 November 2025 | Maximum marks: 100

This examination is open book, open notes. Access to any electronic device is not allowed during the examination. All answers, including numerical ones, should be simplified as much as possible. RMMR tables may be used. The entire question paper is worth 115 marks. The maximum you can score in 100 marks.

1. Consider the R code given below.

```
nsim <- 5000; n <- 50; theta0 <- 10
theta1list <- NULL; theta2list <- NULL
for (i in 1:nsim) {
  x <- runif(n, min=0, max=theta0)
  theta1 <- max(x)
  theta2 <- 2 * mean(x)
  theta1list <- c(theta1list, theta1)
  theta2list <- c(theta2list, theta2)
}
mean((theta0-theta1list)^2); mean((theta0-theta2list)^2)
hist(theta1list); hist(theta2list)
```

   (a) Describe in words what the programmer attempted to do through this code. [3]

   (b) Explain, with theoretical justification, what you expect the numerical part of the output to be. [4+4]

   (c) Explain, with theoretical justification, what you expect the graphical part of the output to be. [2+2]

   [Total 15]

2. Let $Y_1, Y_2, \ldots, Y_n$ be lifetimes of n light bulbs, regarded as an iid sample from a mixture of two populations. Each Y_i is a sample from the exponential distribution with rate λ_1 with probability p and a sample from the exponential distribution with rate λ_2 with probability $1-p$, for $0 < p < 1$. The MLEs of the parameters p, λ_1 and λ_2 have to be obtained from the data. You may assume $\lambda_1 < \lambda_2$.

   (a) 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. [3]

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

   (c) Write down the M-step of the EM algorithm explicitly. [2]

   (d) Suggest how you can obtain a suitable set of 'initial iterates' of the parameters for the 'complete data' version of the problem. [1]

   [Total 10]

Please turn over

3. Suppose Z is a random variable representing the inherent strength of an item that will fail if the environmental stress Y is greater than Z. The strength has distribution F. The stress is given by $Y = \beta^T X$, which is a linear combination of a vector X of stress factors. Let I be the indicator of $Z < Y$, namely the event of failure.

   (a) Obtain the conditional expectation of I given $X = x$, in terms of the distribution F. [2]

   (b) Identify the distribution function F, which makes the model obtained in part (a) same as the logistic regression model. [2]

   (c) Identify the distribution function F, which makes the model obtained in part (a) same as the probit regression model. [2]

   (d) By using the result of part (c), provide a verbal interpretation of the regression coefficient of a particular explanatory variable in the probit regression model. [4]

[Total 10]

4. The data $X_1, X_2, \ldots, X_n$ are presumed to be independent and identically distributed as exponential with rate parameter λ. You have to test the null hypothesis $\mathcal{H}_0 : \lambda = 1$ against the alternative hypothesis $\mathcal{H}_1 : \lambda > 1$.

   (a) If the statistic $T = \min\{X_1, X_2, \ldots, X_n\}$ is used to test $\mathcal{H}_0$ against $\mathcal{H}_1$, and the size of the test is required to be α, what should be the critical region of the test? [5]

   (b) Derive an expression for the power of the above test. [3]

   (c) For which value of λ are the probabilities of the two types of error for the test of part (a) equal to one another? [2]

   (d) Can you suggest an alternative test with size α? Explain. [3]

   (e) Write an R program to compare the powers of the two tests when $\lambda = 2$ [5]

   (f) Comment on the powers of the two tests when the sample size n is large. [2]

[Total 20]

5. It was found from a survey project carried out by BStat students a few years ago that students of class IX in the Baranagar area spent 4.5 hours per week, on average, on the phone. The current students of the same BStat class suspect that now the mean is higher. They ask four hundred randomly chosen students of class IX in the Baranagar area how many hours per week they spend on the phone. The sample mean is found to be 4.75 hours with a sample standard deviation of 2.0. Report the p-value of a suitable test of hypothesis to answer the underlying question, after listing suitable assumptions. [10]

6. The data $X_1, X_2, \ldots, X_n$ are presumed to be independent and identically distributed as normal with mean 0 and variance σ^2.

   (a) Derive the score test for the null hypothesis $\mathcal{H}_0 : \sigma^2 = \sigma_0^2$ against the alternative hypothesis $\mathcal{H}_1 : \sigma^2 > \sigma_0^2$. [9]

   (b) Describe an alternative test based on the exact distribution of an equivalent statistic. [3]

   (c) Describe a realistic problem that can be solved with either of the above tests. [3]

[Total 15]

Please turn over

7. Suppose that you have independent samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ of the normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively. Suppose the null hypothesis $\mathcal{H}_0 : \mu_1 = \mu_2$ has to be tested against the alternative hypothesis $\mathcal{H}_1 : \mu_1 \neq \mu_2$.

   (a) What is the non-centrality parameter of the independent samples t-test statistic under the alternative hypothesis? [3]

   (b) Is the power of the independent samples t-test greater than its size? Explain. [6]

   (c) Suppose the paired data t-test is used for this testing problem after pairing X_i with Y_i for $i = 1, 2, \ldots, n$. Do you expect it to have greater power than the independent samples t-test of identical size? Explain. [6]

   [Total 15]

8. The examiner of a 100 page thesis sample checks 10 randomly selected pages for minor errors. The number of minor errors that he finds on those pages are: 2, 1, 1, 0, 2, 2, 3, 3, 2, 0. Assuming that the number of minor errors in the different pages of the manuscript are independent and have the Poisson distribution with mean λ, derive an asymptotic confidence interval of the total number of minor errors in the manuscript by using the Wald test statistic as pivot, so that the coverage probability is approximately $1 - \alpha$. [10]

9. A number of taxi drivers, who are exposed to more noise on the right side while driving, were tested for moderate level of hearing loss. Out of the sample of 100 drivers, 27 had no loss in either ear, 42 had loss only in the right ear, 16 had loss only in the left ear and the remaining 15 had loss in both the ears. The question of interest is whether there is greater hearing loss in the right ear. Answer this question by suitably formulating it as a test of hypotheses, after making reasonable and clear assumptions. [10]