Strathmore
UNIVERSITY

iLab Africa
MASTER OF SCIENCE IN DATA SCIENCE

END OF SEMESTER EXAMINATION
DSA 8301: STATISTICAL INFERENCE IN BIG DATA
JULY 19, 2023 17:00 - 20:00

Answer Question ONE and two other questions. You must show *all* work to receive *any credit.*

Question ONE (30 marks)

(a) Let X have a Poisson distribution with a variance of 3. Find $P(X = 2)$ **(5 marks)**.

(b) Let $X_1, X_2, ..., X_n$ be a random sample of size n from the distribution with $p.d.f.$

$$f(x;\theta) = \theta x^{\theta-1},\ 0 < x < 1,\ 0 < \theta < \infty.$$

Find the method of moments and the maximum likelihood estimators for θ **(10 marks)**.

(c) In a biology laboratory, students use corn to test the Mendelian theory of inheritance. The theory claims that frequencies of the four categories "smooth and yellow", "wrinkled and yellow", "smooth and purple", and "wrinkled and purple" will occur in the ratio $9 : 3 : 3 : 1$. If a student counted 124, 30, 43, and 11, respectively, for these four categories, would these data support the Mendelian theory, at $\alpha = 0.05$? Note that $\chi^2_{0.05}(3) = 7.815$ **(10 marks)**.

(d) Let X have the $p.m.f.$

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}, x = 0, 1.$$

A uniformly most powerful test for $H_0 : \theta = 0.5$ vs. $H_a : \theta < 0.5$ rejects H_0 if $Y = \sum_{i=1}^{n} X_i$ is observed to be less than or equal to a constant c. For a random sample $X_1, X_2, ..., X_n$ of size $n = 5$, find the significance level when $c = 1$ **(5 marks)**.

Question TWO (15 marks)

A historic data set on the relationship between car speed (X) and stopping distance (Y) is given in the R data *cars*. Use the R output given below to answer the following questions.

```
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -17.5791      6.7584   -2.601    0.0123
x              3.9324      0.4155    9.464  1.49e-12

Multiple R-squared: 0.6511

Analysis of Variance Table
           Df  Sum Sq  Mean Sq  F value    Pr(>F)
x           1   21186                      1.490e-12
Residuals  48   11354
```

(a) Give the values of the t-test statistics for testing $H_0 : \beta_0 = 0$ vs. $H_a : \beta_0 \neq 0$ and $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$. What are the corresponding p-values? **(2 marks)**.

(b) Give the estimate of the standard deviation of the intrinsic error **(5 marks)**.

(c) Fill in the missing entries in the ANOVA table **(5 marks)**.

(d) What proportion of the total variability of the stopping distance is explained by the regression model? **(3 marks)**.

Question THREE (15 marks)

Let $X_1, X_2, ..., X_n$ be *i.i.d.* Exponential(θ).

(a) Show that the likelihood ratio test of $H_0 : \theta = \theta_0$ vs. $H_a : \theta \neq \theta_0$, has a critical region of the form

$$\sum_{i=1}^{n} x_i \leq c_1 \text{ or } \sum_{i=1}^{n} x_i \geq c_2,$$

where c_1 and c_2 are selected appropriately **(8 marks)**.

(b) How would you modify this test so that chi-square tables can be used easily? **(7 marks)**.

Question FOUR (15 marks)

Let $X_1, ..., X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$, where σ^2 is known.

(a) Show that $Y = (X_1 + X_2 + X_3)/3$ is an unbiased estimator of μ **(2 marks)**.

(b) Find the Cramer-Rao lower bound for the variance of an unbiased estimator of μ for a general n **(5 marks)**.

(c) What is the efficiency of Y in part (a) above? **(3 marks)**.

(d) Find a sufficient statistic for μ, if it exists **(5 marks)**.

Question FIVE (15 marks)

Facilities A and B account for 60% and 40%, respectively, of the production of a certain electronic component. The two components from the two facilities are shipped to a packaging location where they are mixed and packaged. A sample of size 100 will be used to estimate the expected life time in the combined population. Use the MSE criterion to decide which of the following two sampling schemes (simple random versus stratified sampling) should be adapted, i.e. simple random sampling at the packaging location, and stratified random sampling based on a simple random sample of size 60 from facility A and a simple random sample of size 40 from facility B. [**Hint**: $\mu = 0.6\mu_A + 0.4\mu_B,\ \sigma^2 = 0.6\sigma_A^2 + 0.4\sigma_B^2 + (0.6)(0.4)(\mu_A - \mu_B)^2$ and the estimators of the mean under both sampling schemes are unbiased for μ.] **(15 marks)**.