Strathmore UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES
MASTER OF SCIENCE IN DATA SCIENCE

END OF SEMESTER EXAMINATION
DSA 8301: STATISTICAL INFERENCE IN BIG DATA

DATE: JULY 23RD, 2024 TIME: 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) If X has a Poisson distribution such that $P(X = 1) = P(X = 2)$, then find $P(X = 4)$ **(10 marks)**.

(b) Let $X_1, X_2, ..., X_n$ be a random sample of size n from the normal distribution with mean μ and variance σ^2. Find the method of moments and the maximum likelihood estimators for μ **(10 marks)**.

(c) Facilities A and B account for 60% and 40%, respectively, of the production of a certain electronic component. The 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 sample of size 60 from facility A and a sample of size 40 from facility B. Assume that the estimators of the expected life time under both sampling schemes are unbiased **(10 marks)**.

Question TWO (15 marks)

Let m be the median of the post-test grip strengths in the right arms of male freshmen in a study of health dynamics. A random sample of 15 such students yielded the following values:

```
35.5 44.0 45.5 46.0 48.0 52.0 52.5 53.0 54.0 57.0 57.5 58.0 58.0 65.5 71.0
```

(a) Using the Wilcoxon statistic, define a critical region that has an approximate level of $\alpha = 0.05$ for testing the hypothesis $H_0 : m = 50$ vs. $H_a : m > 50$ **(5 marks)**:

(b) What is the approximate p-value of this test? **(2 marks)**.

(c) Use the sign test to test the hypothesis in (a) above **(5 marks)**.

(d) Calculate the p-value from the sign test and compare it with that from the Wilcoxon test **(3 marks)**.

Question THREE (15 marks)

Let $X_1, X_2, ..., X_n$ be a random sample of Bernoulli trials $b(1, p)$.

(a) Show that a BCR for testing $H_0 : p = 0.9$ vs. $H_a : p = 0.8$ can be based on the statistic $Y = \sum_{i=1}^{n} X_i$ **(5 marks)**.

(b) If $C = \{(x_1, x_2, ..., x_n) : \sum_{i=1}^{n} x_i \leq n(0.85)\}$ and $Y = \sum_{i=1}^{n} X_i$, find the value of n such that $\alpha = P[Y \leq n(0.85); p = 0.9] = 0.10$ (Hint: use the normal approximation to the binomial distribution) **(3 marks)**.

(c) What is the appropriate value of $\beta = P[Y > n(0.85); p = 0.8]$ for the test given in (b) above? **(2 marks)**.

(d) Is the test in (b) a uniformly most powerful test when $H_a : p < 0.9$ **(5 marks)**.

Question FOUR (15 marks)

Let $X_1, ..., X_n$ be a random sample from Exponential(θ).

(a) Show that $Y = X_1$ is an unbiased estimator of θ **(2 marks)**.

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

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

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

Question FIVE (15 marks)

Each of two comparable classes of 15 students responded to two different methods of instruction, giving the following scores on a standardized test:

```
Class U: 91 42 62 39 55 82 67 44 51 77 61 52 76 41 59

Class V: 80 71 55 67 61 93 49 78 57 88 79 81 63 51 75
```

Use a chi-square test with $\alpha = 0.05$ to test the equality of the distributions of test scores by dividing the combined sample into three equal parts: low, middle and high [Hint: first tertile = minimum + (1/3) * range; second tertile = minimum + (2/3) * range] **(15 marks)**.