Adeline Makokha

191199

DSA 8301 – Statistical Inference in Big Data

CAT 1

May 2025

Question 1

Show that $\frac{63}{512}$ is the probability that the fifth head is observed on the tenth independent flip of a fair coin.

Solution

Step 1: Understand the setting. This is a classic negative binomial problem where we are looking for the probability that the 5th success (head) occurs on the 10th trial.

Step 2: Count favorable arrangements. This means we must get exactly 4 heads in the first 9 flips, and the 10th flip must be a head.

$$P(\text{5th head on 10th flip}) = \binom{9}{4}(0.5)^9 \times 0.5 = \binom{9}{4}(0.5)^{10}$$

Step 3: Evaluate.

$$\binom{9}{4} = 126, \quad (0.5)^{10} = \frac{1}{1024}, \quad \Rightarrow \frac{126}{1024} = \frac{63}{512}$$

Answer: $\boxed{\frac{63}{512}}$

*T*his is a good example of how combinatories and probability work together. It's crucial to correctly count the number of sequences with 4 heads in the first 9 tosses using the binomial coefficient.

Question 2

Let $X \sim \Gamma(\alpha, \beta)$. Find:

1. Method of moment estimators of α and β.

2. Point estimates based on the dataset:

Data: 6.9, 7.3, 6.7, 6.4, 6.3, 5.9, 7.0, 7.1, 6.5, 7.6, 7.2, 7.1, 6.1, 7.3, 7.6, 7.6, 6.7, 6.3, 5.7, 6.7, 7.5, 5.3, 5.4, 7.4, 6.9

Solution

Step 1: Write the moment equations.

$$E[X] = \alpha\beta, \quad \text{Var}(X) = \alpha\beta^2$$

Step 2: Replace with sample moments.

$$\bar{X} = 6.74, \quad S^2 = 0.4432$$

Step 3: Solve for parameters.

$$\hat{\beta} = \frac{S^2}{\bar{X}} = \frac{0.4432}{6.74} \approx 0.0658, \quad \hat{\alpha} = \frac{\bar{X}^2}{S^2} = \frac{6.74^2}{0.4432} \approx 102.5$$

Answer: $\boxed{\hat{\alpha} \approx 102.5,\ \hat{\beta} \approx 0.066}$

*T*he method of moments is used, which offers a straightforward estimation technique. It may not always be as efficient as maximum likelihood, but it is computationally simpler and intuitive.

Question 3

Let $X_1, \ldots, X_{10}$ and $Y_1, \ldots, Y_{10}$ be two independent samples with:

$$E[X_i] = E[Y_i] = \mu, \quad \mathrm{Var}(X_i) = \sigma^2,\ \mathrm{Var}(Y_i) = 4\sigma^2$$

1. Show $\hat{\mu} = \alpha\bar{X} + (1-\alpha)\bar{Y}$ is unbiased.
2. Derive MSE of $\hat{\mu}$.
3. Compare $\bar{X}$ with $0.5\bar{X} + 0.5\bar{Y}$.

Solution

(a) Step 1: Show unbiasedness.

$$E[\hat{\mu}] = \alpha E[\bar{X}] + (1-\alpha)E[\bar{Y}] = \alpha\mu + (1-\alpha)\mu = \mu$$

Conclusion: The estimator is unbiased.

(b) Step 2: Compute variance.

$$\mathrm{Var}(\hat{\mu}) = \alpha^2 \cdot \frac{\sigma^2}{10} + (1-\alpha)^2 \cdot \frac{4\sigma^2}{10} = \frac{\sigma^2}{10}(\alpha^2 + 4(1-\alpha)^2)$$

This is the MSE because the estimator is unbiased.

(c) Step 3: Plug in values.

- For $\bar{X}$: $\alpha = 1 \Rightarrow \mathrm{MSE} = \frac{\sigma^2}{10}$
- For $0.5\bar{X} + 0.5\bar{Y}$: $\mathrm{MSE} = \frac{\sigma^2}{10}(0.25 + 1) = \frac{1.25\sigma^2}{10}$

Conclusion: $\bar{X}$ is more efficient. *A*ssigning equal weights to estimators from data sources

with unequal variances results in increased error. Weighting in proportion to variance is better.

Question 4

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.

1. Show $Y = (X_1 + X_2)/2$ is unbiased.
2. Find Cramer-Rao lower bound (CRLB).
3. Find efficiency of Y.

Solution

(a) Step 1: Unbiasedness

$$E[Y] = \frac{1}{2}(E[X_1] + E[X_2]) = \mu$$

(b) Step 2: Compute Fisher Information

$$I(\mu) = \frac{n}{\sigma^2}, \quad \text{CRLB} = \frac{\sigma^2}{n}$$

(c) Step 3: Compute efficiency

$$\text{Var}(Y) = \frac{\sigma^2}{2}, \quad \text{Efficiency} = \frac{\sigma^2/n}{\sigma^2/2} = \frac{2}{n}$$

Efficiency reflects how well an estimator uses data. Using only part of the sample reduces efficiency unless there's a strong reason to restrict the sample.

Question 5

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

$$H_0 : \theta = 1 \quad \text{vs.} \quad H_a : \theta = 2$$

Show best critical region is:

$$C = \{(x_1, ..., x_n) : \prod x_i \geq c\}$$

Solution

Step 1: Write likelihood ratio.

$$\Lambda = \frac{L(\theta = 2)}{L(\theta = 1)} = 2^n \prod x_i$$

Step 2: Apply Neyman-Pearson Lemma.

We reject H_0 for large values of the likelihood ratio $\Rightarrow$ large $\prod x_i$.
Critical Region:

$$C = \left\{ \prod_{i=1}^{n} x_i \geq c \right\}$$

A large product of x_i values favors $\theta = 2$ over $\theta = 1$. Hence this is a powerful test statistic.

Question 6

Test:

$$H_0 : \theta_2 = \theta_2' \text{ vs. } H_a : \theta_2 \neq \theta_2' \quad (\text{mean } \theta_1 \text{ unspecified})$$

Show rejection region:

$$\sum (x_i - \bar{x})^2 \leq c_1 \quad \text{or} \quad \sum (x_i - \bar{x})^2 \geq c_2$$

Solution

Step 1: Construct likelihood ratio. The statistic depends on the sample variance:

$$T = \sum (x_i - \bar{x})^2$$

Step 2: Determine behavior under H_0. If T is far from expected under H_0, reject.
Rejection Region:

$$T \leq c_1 \quad \text{or} \quad T \geq c_2$$

A two-tailed test is suitable here since both smaller and larger sample variances contradict the null hypothesis.

Question 7

Let $X_i \sim \text{Bern}(\theta)$. Test:

$$H_0 : \theta = 0.5 \quad \text{vs.} \quad H_a : \theta < 0.5$$

Reject if $Y = \sum X_i \leq c$.

1. Show this is UMP.
2. Find α when $c = 1$.
3. Find α when $c = 0$.

Solution

(a) Step 1: Use monotone likelihood ratio. Binomial distribution has MLR in Y. Hence, Karlin-Rubin theorem implies test is UMP.

(b) Step 2: Calculate significance level.

$$\alpha = P(Y \leq 1) = \frac{1+5}{32} = \frac{3}{16} = 0.1875$$

(c) Step 3: Calculate when $c = 0$.

$$\alpha = P(Y = 0) = \frac{1}{32} = 0.03125$$

This example demonstrates the use of discrete probabilities in hypothesis testing. Using a lower c leads to a more conservative test.

Question 8

Observed: 124, 30, 43, 11   Expected ratio: 9:3:3:1 $\alpha = 0.05$

Solution

Step 1: Calculate expected frequencies.

$$E_1 = 117,\ E_2 = 39,\ E_3 = 39,\ E_4 = 13$$

Step 2: Compute test statistic.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{49}{117} + \frac{81}{39} + \frac{16}{39} + \frac{4}{13} \approx 3.214$$

Step 3: Compare with critical value.

$$\chi^2_{0.95,3} = 7.815 \Rightarrow 3.214 < 7.815 \Rightarrow \text{Fail to reject}$$

Conclusion: The data supports Mendel's theoretical prediction.

Chi-square is sensitive to count discrepancies. Here, observed values are close enough to expected ones to attribute the difference to chance.