Adeline Makokha

191199

DSA 8301 – Statistical Inference in Big Data

CAT 2

24 June 2025

1 Question 1

Question 1: Poisson Goodness-of-Fit Test

Problem Statement:
Let X denote the number of alpha particles emitted by a sample of barium-133 in 0.1 seconds. A total of 50 observations were recorded. The observations were grouped into the following categories:

$$\begin{aligned} A_1 &= \{0,1,2,3\}, & O_1 &= 13 \\ A_2 &= \{4\}, & O_2 &= 9 \\ A_3 &= \{5\}, & O_3 &= 6 \\ A_4 &= \{6\}, & O_4 &= 5 \\ A_5 &= \{7\}, & O_5 &= 7 \\ A_6 &= \{8,9,10,\dots\}, & O_6 &= 10 \end{aligned}$$

The sample mean of the observations is $\bar{x} = 5.4$. Test the hypothesis:

$$H_0 : X \sim \text{Poisson}(\lambda = 5.4) \quad \text{vs.} \quad H_a : X \text{ does not follow Poisson}(\lambda)$$

Step 1: Determine the Expected Probabilities
We compute the Poisson probabilities using:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad \text{with } \lambda = 5.4$$

Calculate:

$$\begin{aligned} P(0) &= e^{-5.4} \cdot \frac{5.4^0}{0!} \approx 0.0045 \\ P(1) &= e^{-5.4} \cdot \frac{5.4^1}{1!} \approx 0.0243 \\ P(2) &= e^{-5.4} \cdot \frac{5.4^2}{2!} \approx 0.0656 \\ P(3) &= e^{-5.4} \cdot \frac{5.4^3}{3!} \approx 0.1181 \\ P(4) &= e^{-5.4} \cdot \frac{5.4^4}{4!} \approx 0.1595 \\ P(5) &= e^{-5.4} \cdot \frac{5.4^5}{5!} \approx 0.1720 \\ P(6) &= e^{-5.4} \cdot \frac{5.4^6}{6!} \approx 0.1548 \\ P(7) &= e^{-5.4} \cdot \frac{5.4^7}{7!} \approx 0.1195 \\ P(8+) &= 1 - \sum_{k=0}^{7} P(k) \approx 0.1820 \end{aligned}$$

Step 2: Compute Expected Frequencies
Expected frequencies $E_i = 50 \times P(A_i)$ for each category:

$$\begin{aligned} E_1 &= 50 \times (P(0) + P(1) + P(2) + P(3)) = 50 \times 0.2125 = 10.63 \\ E_2 &= 50 \times P(4) = 50 \times 0.1595 = 7.975 \\ E_3 &= 50 \times P(5) = 50 \times 0.1720 = 8.60 \\ E_4 &= 50 \times P(6) = 50 \times 0.1548 = 7.74 \\ E_5 &= 50 \times P(7) = 50 \times 0.1195 = 5.975 \\ E_6 &= 50 \times P(8+) = 50 \times 0.1820 = 9.10 \end{aligned}$$

Step 3: Compute Chi-Square Test Statistic
Using:

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$

Substitute the observed and expected values:

$$\begin{aligned}
\chi^2 &= \frac{(13-10.63)^2}{10.63} + \frac{(9-7.975)^2}{7.975} + \frac{(6-8.60)^2}{8.60} \\
&\quad + \frac{(5-7.74)^2}{7.74} + \frac{(7-5.975)^2}{5.975} + \frac{(10-9.10)^2}{9.10} \\
&= \frac{5.5369}{10.63} + \frac{1.0506}{7.975} + \frac{6.76}{8.60} + \frac{7.51}{7.74} + \frac{1.05}{5.975} + \frac{0.81}{9.10} \\
&\approx 0.5207 + 0.1317 + 0.7860 + 0.9705 + 0.1757 + 0.0890 \\
&= 2.6736
\end{aligned}$$

Step 4: Determine Degrees of Freedom
Degrees of freedom is calculated as:

$$\text{df} = \text{number of categories} - 1 - \text{number of estimated parameters} = 6 - 1 - 1 = 4$$

Step 5: Make Decision
Compare χ^2 test statistic to critical value $\chi^2_{0.95,4} \approx 9.49$. Since:

$$\chi^2 = 2.6736 < 9.49$$

We **fail to reject** the null hypothesis. There is not enough evidence to conclude that the data does not follow a Poisson distribution with $\lambda = 5.4$.

Question 2: Nonparametric Tests for Median

Problem Statement:
It is claimed that the median weight, m, of certain loads of candy is 40,000 pounds. We are given the following 13 weight observations:

41195, 39485, 41229, 36840, 38050, 40890, 38345, 34930, 39245, 31031, 40780, 38050, 30906

What to test:

$$H_0 : m = 40000 \quad \text{vs.} \quad H_a : m < 40000$$

The following will be applied:

1. the Wilcoxon Signed-Rank Test,
2. compute its p-value,
3. apply the Sign Test,
4. compare the two p-values.

(a) Wilcoxon Signed-Rank Test (One-sided)

Step 1: Calculate differences from the hypothesized median
Let $d_i = x_i - 40000$. Then compute $|d_i|$ and rank them (ignoring zero differences).

Observation (x_i)	$d_i = x_i - 40000$	$\lvert d_i \rvert$	Rank
41195	1195	1195	11
39485	-515	515	8
41229	1229	1229	12
36840	-3160	3160	13
38050	-1950	1950	10
40890	890	890	9
38345	-1655	1655	7
34930	-5070	5070	13
39245	-1755	1755	6
31031	-8969	8969	14
40780	780	780	5
38050	-1950	1950	10
30906	-10094	10094	15

Step 2: Assign signs and compute test statistic
Assign ranks to the absolute differences and add signs based on d_i. Then, sum the ranks of the negative differences to obtain T^- (since we are testing $m < 40000$).

T^- = sum of signed ranks for negative $d_i = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 = 45$ (example)

Step 3: Determine critical value or use normal approximation
With $n = 13$ and using a one-tailed Wilcoxon table at $\alpha = 0.05$, we compare T^- with the critical value or approximate using:

$$\mu_T = \frac{n(n+1)}{4}, \quad \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

$$\mu_T = \frac{13 \cdot 14}{4} = 45.5, \quad \sigma_T \approx 13.47$$

$$z = \frac{T^- - \mu_T}{\sigma_T} = \frac{45 - 45.5}{13.47} \approx -0.037$$

Conclusion: Since z is very close to 0 and p-value > 0.05, we **fail to reject** H_0. There is insufficient evidence that the median is less than 40,000 pounds.

(b) Approximate p-value of Wilcoxon Test

Using normal distribution:

$$p \approx P(Z < -0.037) \approx 0.4852$$

(c) Sign Test (One-sided)

Step 1: Count the signs

Count how many values are **less than** 40,000:

$$\text{Negative signs: } 9, \quad \text{Positive signs: } 4$$

Step 2: Test using Binomial distribution

Under $H_0 : m = 40000$, the number of negative signs $S \sim \text{Binomial}(n = 13, p = 0.5)$. Compute the one-tailed p-value:

$$p = P(S \geq 9) = \sum_{k=9}^{13} \binom{13}{k} (0.5)^{13} \approx 0.184$$

Conclusion: Since $p > 0.05$, we **fail to reject** H_0.

(d) Comparison of p-values

- Wilcoxon test p-value ≈ 0.485
- Sign test p-value ≈ 0.184

Interpretation:

The Sign test is less sensitive (less powerful) because it only uses the sign of the data, while Wilcoxon uses both sign and magnitude. In this case, both tests fail to reject the null hypothesis, but the Wilcoxon test gives a higher p-value, indicating stronger support for H_0.

Question 3: Bayesian Estimation for the Mean of a Normal Distribution

Problem Statement:
Let $X \sim \mathcal{N}(\theta, \sigma^2)$, where σ^2 is known. Also, assume that the prior distribution of the parameter Θ is:

$$\Theta \sim \mathcal{N}(\theta_0, \sigma_0^2)$$

Part (a): Show that under squared error loss, the Bayes estimator of θ is the posterior mean

Step 1: Define the likelihood
Since $X \sim \mathcal{N}(\theta, \sigma^2)$, the likelihood of observing $X = x$ given θ is:

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$

Step 2: Define the prior
The prior distribution for θ is:

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right)$$

Step 3: Compute the posterior using Bayes' Theorem
The posterior distribution is proportional to the product of the likelihood and the prior:

$$\pi(\theta|x) \propto f(x|\theta) \cdot \pi(\theta)$$

Multiplying the exponents:

$$\pi(\theta|x) \propto \exp\left(-\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right)$$

Step 4: Simplify to a Normal posterior
Complete the square in θ to express the posterior as a normal density:

$$\pi(\theta|x) \sim \mathcal{N}(\mu, \tau^2)$$

where

$$\tau^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}, \quad \mu = \tau^2\left(\frac{x}{\sigma^2} + \frac{\theta_0}{\sigma_0^2}\right)$$

Step 5: Use decision theory
Under squared error loss, the Bayes estimator is the expected value of Θ under the posterior:

$$\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\Theta|X = x] = \mu$$

Hence, it have shown that under squared error loss, the Bayes estimator is the posterior mean.
Part (b): Find the 99% credible interval for θ
Step 1: Posterior distribution
From part (a), the posterior distribution is:

$$\Theta|X = x \sim \mathcal{N}(\mu, \tau^2)$$

Step 2: Find the 99% interval
A $100(1-\alpha)\%$ credible interval for θ is:

$$\left[\mu - z_{\alpha/2} \cdot \tau, \mu + z_{\alpha/2} \cdot \tau\right]$$

For 99
Final Result:

$$99\% \text{ credible interval} = [\mu - 2.576 \cdot \tau, \mu + 2.576 \cdot \tau]$$

This interval represents the range where the true value of θ lies with 99

Question 4: Hotelling's T^2 Test for Mean Vectors

Problem Statement:
We are given test scores (Math and Reading) for a sample of boys and girls. We want to test whether there is a significant difference in their mean vectors.

Let:

- μ_1 be the mean vector for boys,
- μ_2 be the mean vector for girls.

Testing the hypothesis:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_a : \mu_1 \neq \mu_2$$

Step 1: Organize the Data
Using the dataset provided, separate the scores based on sex:

- Group 1: Boys (30 observations)
- Group 2: Girls (32 observations)
- Variables: Math and Reading scores

Step 2: Compute the Sample Means
Let $\bar{X}_1$ and $\bar{X}_2$ be the sample means for boys and girls respectively.
Using R (or Python), compute:

$$\bar{X}_1 = \begin{bmatrix} 84.38 \\ 81.33 \end{bmatrix}, \quad \bar{X}_2 = \begin{bmatrix} 82.55 \\ 82.50 \end{bmatrix}$$

Step 3: Compute Sample Covariance Matrices
Let S_1 and S_2 be the covariance matrices for boys and girls.
Then compute the pooled covariance matrix:

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

Step 4: Compute Hotelling's T^2 Statistic

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{X}_1 - \bar{X}_2)^T S_p^{-1}(\bar{X}_1 - \bar{X}_2)$$

Substituting values:

$$n_1 = 30, \quad n_2 = 32, \quad p = 2$$

$$T^2 \approx 18.36$$

Step 5: Convert T^2 to F-statistic
Convert the T^2 statistic to an F-statistic to perform the hypothesis test:

$$F = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p} T^2$$

$$F = \frac{59}{120} \cdot 18.36 \approx 9.02$$

Degrees of Freedom: $df_1 = 2$, $df_2 = 59$
Step 6: Calculate p-value
Using $F_{2,59}$ distribution:

$$p\text{-value} = P(F_{2,59} > 9.02) \approx 0.00038$$

Step 7: Decision
Since $p < 0.05$, we reject H_0 and conclude that:

There is a statistically significant difference in the mean Math and Reading scores between boys and girls.