Statistical Inference

Instructor: Mohammadreza A. Dehaqani

Pouya HajiMohammadi Gohari, Parmis Bathayan

Fall 2024

Homework 3 Part I

- This homework is cut into two halves. To help you focus on preparing for the midterm, weve left unrelated and heavy computational topics for the second part. This homework targets core concepts that will be in your midterm.
- If you have questions, email the HW Authors or use the class group we are ready to help!
- Please check the course page for key submission guidelines and late policy details to avoid issues.
- For computational problems, your grade depends heavily on how well you analyze your results. Always include explanations alongside your code.

Question 1: Say Hello to Neyman-Pearson

Let $X^{'} = (X_1, \cdots, X_n)$ denote a random sample from the distribution that has the pdf

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right), \; -\infty < x < \infty$$

1. Assume null and alternative hypothesis as follows:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

   Derive the likelihood ratio and identify the test statistic.

2. Express the rejection region in terms of the test statistic.
3. Compute the power under specified parameters:

$$\theta_0 = 0, \; \theta_1 = 1, \; \alpha = 0.05$$

Question 2: Wald Test

Let $X_1, \cdots, X_n \sim Poisson(\lambda)$.

1. From MLE, find $\hat{\lambda}$.
2. What is the standard deviation of the estimated point($\hat{\lambda}$)?
3. Suppose we test the following hypothesis:

$$H_0 : \lambda = \lambda_0$$
$$H_1 : \lambda \neq \lambda_1$$

   If we use the statistic:

$$\frac{\hat{\lambda} - \lambda_0}{\hat{se}}$$

   What is the distribution of this statistic under H_0? Express the rejection area in terms of the test statistic.

4. (Programming) Let λ_0, $n = 20$ and $\alpha = 0.05$. Simulate $X_1, \cdots, X_n \sim Possion(\lambda_0)$ and perform above test. Repeat many times and count how often you reject the null. How close is the Type I error rate to 0.05?

Question 3: $f_0(x)$ vs $f_1(x)$

Consider two p.d.f's $f_0(x)$ and $f_1(x)$ that are defined as follows:

$$f_0(x) = \begin{cases} \frac{3}{2} & \text{for } 0 \leq x \leq \frac{2}{3} \\ 0 & \text{otherwise} \end{cases}$$

and

$$f_1(x) = \begin{cases} 3x & \text{for } 0 \leq x \leq \frac{2}{3} \\ 0 & \text{otherwise} \end{cases}$$

Suppose that a single observation X is taken from a distribution for which the p.d.f $f(x)$ is either $f_0(x)$ or $f_1(x)$, and the following simple hypothesis are to be tested:

$$H_0: \ f(x) = f_0(x),$$
$$H_1: \ f(x) = f_1(x)$$

1. Describe a test procedure for which the following equation would be minimum:

$$\alpha + \beta$$

   where α and β are the probability of type I and II errors, respectively.

2. (Programming) Consider 1000 equally spaced decision boundaries c in the range $[0.\frac{3}{2}]$, for each decision boundary c compute the α and β. Plot the power of the test as a function of the decision boundary c and compare it to the theoretical part.

3. A medical diagnostic test is performed to detect a rare disease. The disease is serious, and a missed diagnosis (α) can result in severe consequences, while a false alarm (β) may lead to unnecessary but harmless additional tests. The test result is based on an observed value X, drawn from a distribution $f(x)$, which is either:

$$H_0: \ f(x) = f_0(x) \text{ (patient is healthy)},$$
$$H_1: \ f(x) = f_1(x) \text{ (patient has the disease)}$$

   In medical diagnostics, minimizing false negatives (β) is prioritized, as the consequences of missing a diagnosis are far more severe than unnecessary follow-up tests. To address this, we assign risks to each type of error:

   - R_1: Risk associated with α
   - R_2: Risk associated with β

   The expected risk is defined as:

$$\text{Expected Risk} = R_1\alpha + R_2\beta$$

   Provide a procedure to minimize the expected risk by selecting an optimal decision boundary c^*(Derive the expression for the optimal decision boundary c^* based on R_1 and R_2).

Question 4: Uniformly Most Powerful

Let $X_1, X_2, \cdots, X_n$ be a random sample from a distribution with pdf $f(x;\theta), 0 < x < 1$, zero elsewhere, where $\theta > 0$.Show the likelihood ratio has the statistic $\prod_{i=1}^{n} X_i$. Use this to determine the UMP[1] test for $H_0 : \theta = \theta^{'}$ against $H_1 : \theta < \theta^{'}$, for fixed $\theta^{'}$.

Question 5: Generalized Likelihood Ratio(Optional)

1. Let $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$ be independent random samples from the distributions $N(\theta_1, \theta_3)$ and $N(\theta_2, \theta_4)$, respectively. Show that the likelihood ratio for testing $H_0 : \theta_1 = \theta_2, \theta_3 = \theta_4$ against all alternatives is given by

$$\frac{\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\right]^{n/2}\left[\frac{\sum_{j=1}^{m}(y_j-\bar{y})^2}{m}\right]^{m/2}}{\left\{\left[\sum_{i=1}^{n}(x_i-u)^2+\sum_{j=1}^{m}(y_j-u)^2\right]/(n+m)\right\}^{(n+m)/2}},$$

   where $\mu = \frac{n\bar{x}+m\bar{y}}{m+n}$.

2. Let the independent random variables X and Y have distributions that are $N(\theta_1, \theta_3)$ and $N(\theta_2, \theta_4)$, where the means θ_1 and θ_2 and common variance θ_3 are unknown. If $X_1, X_2, \cdots, X_n$ and $Y_1, Y_2, \cdots, Y_m$ denote independent random samples from these distribution. Show that the likelihood ratio for testing $H_0 : \theta_1 = \theta_2$, unspecified, and θ_3 unspecified, can be based on the test statistic T(t-distribution) with n+m-2 degrees of freedom.

Question 6: Error Types

One generates a number X from a uniform distribution on the interval $[0, \theta]$. A hypothesis test is performed to test $H_0 : \theta = 2$ against $H_A : \theta \neq 2$ by rejecting H_0, if $X \leq 0.1$ or $X \geq 1.9$.

1. Compute the probability of a type I error. Illustrate the rejection region on a plot of the uniform probability density function when $\theta = 2$.

2. Compute the probability of a type II error if the true value of θ is 2.5. Show the calculations and illustrate how the shift in θ affects the overlap between the rejection and acceptance regions.

3. Suppose we change the decision rule to reject H_0 if $X \leq c_1$ or $X \geq c_2$, where c_1 and c_2 are adjustable thresholds. Derive appropriate values for c_1 and c_2 such that the type I error probability is exactly $\alpha = 0.05$. Explain how the new α affects the power of the test.

4. Discuss the impact of increasing θ on the type II error probability. Provide a mathematical argument to support your answer.

5. Derive an upper bound for the type I error probability α using an appropriate inequality studied in the course (e.g., for distributions with finite variance). Compare this bound to the exact value of α from part 3 and discuss its significance.

6. Identify and use an appropriate inequality to derive an upper bound for the type II error probability β when $\theta = 2.5$. If needed, feel free to search the web for relevant inequalities, and justify your choice. Discuss how such bounds are useful in hypothesis testing, especially when exact calculations are impractical. **(Optional)**

[1] Uniformly Most Powerful

Question 7: Warm Up t-test

Given a two-sample t-test to determine if there is a significant difference between the means of two populations:

$$\text{Null Hypothesis: } H_0 : \mu_1 = \mu_2, \quad \text{Alternative Hypothesis: } H_1 : \mu_1 \neq \mu_2$$

Follow the instructions below to prove that H_0 is rejected if and only if the confidence interval for the difference between the sample means does not include zero.

(a) Derive the test statistic assuming $H_0 : \mu_1 = \mu_2$, and identify the conditions under which H_0 is rejected.

(b) Derive the confidence interval formula for the difference $\mu_1 - \mu_2$, and explain the role of the critical value in determining the interval's endpoints.

(c) Prove that the confidence interval excludes zero if and only if the t-test rejects H_0. Provide a clear explanation of the equivalence between these two approaches and interpret the result in the context of hypothesis testing.