Statistical Inference

HW Author: Ali Mikaeili, Muhammad Valinezhad
Instructor: Mohammadreza A. Dehaqani

Spring 2024

Homework 4

- Feel free to use the class group to ask questions — our TA team will do their best to help out!
- Please consult the course page for important information on submission guidelines and delay policies to ensure your homework is turned in correctly and on time.
- This course aims to equip you with the skills to tackle all problems in this domain and encourages you to engage in independent research. Utilize your leanings to extend beyond the classroom teachings where necessary.
- The optional questions in this homework can also make up for the shortcomings in your previous homework, so you are encouraged to solve them or at least attempt to.

Question 1: Contingency Table (Optional)

In a clinical trial, three different treatments are being tested for their effectiveness in reducing symptoms of a chronic disease. Patients' responses are categorized as "Significant Improvement", "Moderate Improvement", or "No Improvement". The following table summarizes the outcomes observed for each treatment group.

	Treatment A	**Treatment B**	**Treatment C**
Significant Improvement	13	33	15
Moderate Improvement	7	4	11
No Improvement	18	10	14

Perform a hypothesis test to assess if there is a significant difference in the effectiveness of the treatments at a 5% significance level.

Question 2: How good is the fit?

Eighty train engines are run continuously until they stop functioning. The operational durations T in hours are recorded as follows. The sample mean is 6434. Can you verify the hypothesis that the distribution of T follows an exponential pattern? Use the 5% level of significance.

Class	Frequency	Class	Frequency
$0 < T < 2000$	11	$10000 < T < 12000$	3
$2000 < T < 4000$	21	$2000 < T < 14000$	5
$4000 < T < 6000$	19	$14000 < T < 16000$	1
$6000 < T < 8000$	9	$16000 < T < 18000$	3
$8000 < T < 10000$	4	$18000 < T$	4

Question 3: GLRT

Let $X_1, X_2, \ldots, X_n$ be i.i.d $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Consider the problem of testing

$$H_0 : \mu = 0$$
$$H_1 : \mu \neq 0$$

Show that the generalized likelihood ratio statistic for this problem simplifies to

$$\Lambda(X_1, \ldots, X_n) = \left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2 + n\bar{X}^2}\right)^{n/2}$$

Letting $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ and $T = \frac{\sqrt{n}\bar{X}}{S_X}$ (the usual one-sample t-statistic for this problem), show that $\Lambda(X_1, \ldots, X_n)$ is a monotonically decreasing function of |T|, and hence the generalized likelihood ratio test is equivalent to the two-sided t-test which rejects for large values of |T|.

Question 4: Rank-Sum

Two different types of batteries are compared by measuring how long they last in hours under continuous use. Twelve batteries of each type are tested. The lifetimes in hours are as follows:

Battery A	**Battery B**
36.1	62.5
16.6	28.2
24.6	19.9
38.5	33.9
15.6	13.3
28.3	39.4
16.0	19.3
44.7	23.7
14.3	12.7
10.8	122.0
0.7	168.0
6.5	55.0

Use a Wilcoxon rank-sum test to test the null hypothesis that the mean lifetimes are equal for the two types of batteries against the alternative that they are not. Use the 5% level of significance. Comment on any assumptions which are necessary.

Question 5: Sign-Rank (Optional)

Consider the data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f$, where f is an unknown probability density function, and the hypothesis testing problem:

$$H_0 : \text{The median of } f \text{ is } 0 \quad \text{vs.} \quad H_1 : \text{The median of } f \text{ is } \mu \text{ for some } \mu > 0.$$

(a) Explain why the Wilcoxon signed-rank statistic does not have the same sampling distribution under every $P \in H_0$. Draw a graph of a density function f with median 0, such that the Wilcoxon signed-rank statistic would tend to take larger values under f than under any density function g that is symmetric about 0.

(b) Consider the *sign statistic* S, defined as the number of values among $X_1, \ldots, X_n$ that are greater than 0. Explain why S has the same sampling distribution under any $P \in H_0$. How would you conduct a level-α test of H_0 against H_1 using the test statistic S? (Explicitly describe the rejection threshold; you may assume that for $X \sim \text{Binomial}(n, \frac{1}{2})$, there exists an integer k such that $\mathbb{P}[X \geq k] = \alpha$.)

(c) When n is large, explain why we may reject H_0 when $S > \frac{n}{2} + \sqrt{\frac{n}{4}}z(\alpha)$, where $z(\alpha)$ is the upper α point of the standard normal distribution $\mathcal{N}(0,1)$, instead of using the rejection threshold derived in part (b).

Question 6: t and Wilcoxon meet

(a) In a clinical trial, a new drug is tested to see if it will lower the cholesterol levels of patients. The following information is obtained from the treatment and control groups:

	Treatment Group	**Control Group**
Sample Size	$n_x = 7$	$n_y = 8$
Sample Mean	$\bar{x} = 6.34286$	$\bar{y} = 8.0625$
Sample Variance	$s_x^2 = (1.40102)^2$	$s_y^2 = (1.10446)^2$

Test if the cholesterol levels of patients in the treatment group are lower than those in the control group using the two-sample t-test.

(b) Repeat the test in part (a) without the assumption of equal variance.

(c) From a group of nine rats available for a study of transfer of learning, five were selected at random and were trained to imitate leader rats in a maze. They were then placed together with four untrained control rats in a situation where imitation of the leaders enabled them to avoid receiving an electric shock. The results (the number of trials required to obtain ten correct responses in ten consecutive trials) were as follows:

Trained Rats	78	64	75	45	82
Controls	110	70	53	51	

Test if there is a difference in the number of trials required between the trained rats and the controls using the Wilcoxon rank sum test.

Question 7: Binomial Test

1. The proportion of defective items produced by a factory is 0.1. As part of the quality control procedure, a random sample of 12 items is inspected daily.
   (a) State the distribution of X, the number of defective items in a random sample of size 12.
   (b) Find the probability that there are fewer than two defective items on a particular day.
   (c) A new manager introduces work practices which are expected to reduce the proportion of defective items produced. After a settling-in period, he asks for a random sample of 200 items to be inspected for defects. Test the effectiveness of the new work practices if it is found that there are only 11 defective items in the sample. [Hints: Let p be the proportion of defective items produced after the new work practices are introduced. Set up appropriate null and alternative hypotheses concerning p.]
2. Suppose that the probability of success at each repetition of an experiment is 0.6. Perform a one-sided test of $H_0 : p = 0.6$ vs. $H_1 : p > 0.6$ if 40 successes are observed in 50 trials. Comment on your answers.

Question 8: KS

You are given two samples representing the test scores of students from two different streams who are considering joining a Computer Science (CS) program. The sample sizes are $n = 60$ and $m = 50$ for streams A and B, respectively. We want to test whether the distributions of test scores for the two streams are different at a 5% significance level using the Kolmogorov-Smirnov test.

The empirical distribution functions for the two samples are given as follows:

- For stream A: $F_{A,n}(X)$
- For stream B: $F_{B,m}(X)$

The maximum absolute difference between these two empirical distribution functions is calculated to be $D = 0.25$.

1. **Calculate the critical value D_α at the 5% significance level.**
2. **Determine whether the null hypothesis that the two samples come from the same distribution should be rejected.**
3. **Explain your conclusion based on the results of the Kolmogorov-Smirnov test.**

Question 9: To be Normal or Not to Be

Imagine we want to analyze the age of male and female fans of a small new band. We want to test the hypothesis whether male fans are older than female fans. We collected a sample of 50 male fans and 50 female fans. Now, we want to analyze the age distribution to determine if there is a significant age difference between male and female fans.

Men's Age: 52, 18, 27, 12, 24, 17, 68, 25, 12, 9, 51, 44, 42, 34, 44, 15, 21, 66, 61, 32, 31, 20, 6, 13, 34, 38, 45, 17, 16, 15, 36, 21, 29, 21, 29, 9, 33, 15, 37, 27, 31, 15, 57, 37, 27, 31, 38, 27, 60, 23

Women's Age: 36, 49, 20, 31, 51, 31, 15, 16, 39, 70, 52, 16, 39, 34, 18, 34, 30, 18, 26, 18, 25, 16, 39, 49, 22, 37, 39, 21, 16, 63, 45, 43, 17, 28, 29, 23, 42, 23, 28, 55, 41, 18, 23, 8, 13, 26, 13, 27, 28, 18

Tasks:

1. Make a histogram of the original data for the ages of men and women.
2. Show whether the distribution is normal or not using statistical methods.
3. Can we use parametric tests in this problem? Examine all of the assumptions.
4. Can we transform the data to the normal distribution? Implement it and recheck the distribution.
5. If the parametric requirements are met, perform a parametric test and analyze the results.
6. Perform a non-parametric test on the original data and compare the results with the parametric one. Is there any difference? Compare the powers of two tests.
7. Which test is more appropriate in this situation? Why?

Important Note: For this question, you might need to go beyond the tests of normalcy you learned in class. You are expected to do your own research and find the best one if needed.

Question 10: QQ Time

Use either R or Python to answer the following questions.

In this section, you will generate random samples from various distributions and create Q-Q (quantile-quantile) plots to compare these samples against theoretical distributions. You will need to write Python or R code to perform these tasks and then answer questions based on the generated plots.

Instructions

1. **Generate Random Samples:**
   (a) Create a random sample from a normal distribution with a mean of 0 and a standard deviation of 1.
   (b) Create a random sample from a right-skewed distribution using an exponential distribution.
   (c) Create a random sample from a left-skewed distribution by negating an exponential distribution.
   (d) Create a random sample from an under-dispersed distribution by truncating a normal distribution with a mean of 0 and a standard deviation of 0.5 within the range [-1, 1].
   (e) Create a random sample from an over-dispersed distribution by combining two normal distributions, one with a mean of -2 and the other with a mean of 2, both with a standard deviation of 1.
2. **Create Q-Q Plots:**
   Generate Q-Q plots for each of the samples you created. Compare the normal sample to a normal distribution, the right-skewed sample to an exponential distribution, the left-skewed sample to an exponential distribution, and both the under-dispersed and over-dispersed samples to a normal distribution.

Questions

1. **Describe the purpose of a Q-Q plot. What information does it provide about the data?**
2. **Explain the process of generating the `under_dispersed_data` sample. Why is this data considered under-dispersed?**
3. **The Q-Q plot for the `right_skewed_data` uses an exponential distribution for comparison. Why is an exponential distribution appropriate for this data, and what would you expect the Q-Q plot to look like?**
4. **Examine the Q-Q plot for `over_dispersed_data`. This data is generated from a mixture of normal distributions. Describe how you would expect this Q-Q plot to differ from the Q-Q plot of the `normal_data`.**
5. **What characteristics of the Q-Q plots would indicate that the data does not follow the theoretical distribution? Provide examples based on the distributions used in the problem.**
6. **Write the Python code required to generate the samples and create the Q-Q plots as described.**

Question 11: Random Number Generation (Optional)

Given a seed $x_0 = 15$, generate 30 random numbers using the following linear congruential generator (LCG) formula:

$$x_n = 3x_{n-1} \mod 31$$

1. Calculate the sequence of 30 random numbers starting with $x_0 = 15$.

2. Normalize these numbers by dividing each by 31.
3. Verify the sequence using the provided normalized sequence:

   $0.45161, 0.35484, 0.06452, 0.19355, 0.58065, 0.74194, 0.22581, 0.67742, 0.03226, 0.09677, 0.29032, 0.87097, 0.61290, 0.83871$

Question 12: Kernel Density Estimation (Optional)

Use either R or Python to answer the following questions.

You are given a sample of $n = 50$ observations from a continuous random variable. Your task is to estimate the probability density function (PDF) of this random variable using kernel density estimators (KDE).

1. **Understanding the KDE:**

   Given the sample $\{X_1, X_2, \ldots, X_n\}$, the kernel density estimator for the PDF $f(x)$ is defined as:

   $$\hat{f}(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)$$

   where:

   - K is the kernel function.
   - h is the bandwidth or smoothing parameter.

2. **Choice of Kernel Function:**

   For this problem, use the Gaussian kernel, defined as:

   $$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

3. **Bandwidth Selection:**

   Use the rule-of-thumb bandwidth selection method, also known as Silverman's rule of thumb, given by:

   $$h = 1.06\hat{\sigma}n^{-1/5}$$

   where $\hat{\sigma}$ is the sample standard deviation.

4. **Estimating the Density:**

   Given the sample data $\{X_1, X_2, \ldots, X_n\}$, use the above formulae to estimate the PDF at specific points.

Tasks

1. **Calculate the bandwidth h using Silverman's rule of thumb.**
2. **Using the Gaussian kernel, compute the kernel density estimate $\hat{f}(x)$ at the points $x = 5$, $x = 10$, and $x = 15$. The sample data are given below:**

   $X = [3, 7, 8, 12, 13, 14, 15, 16, 18, 20, 3, 6, 9, 11, 13, 14, 15, 17, 19, 21, 2, 4, 7, 10, 12, 13, 16, 17, 20, 22, 1, 5, 8, 11, 14, 15, 16, 18,$

3. **Plot the kernel density estimate $\hat{f}(x)$ over the range $x = 0$ to $x = 25$.**