Introduction to Statistical Inference

Instructor: Mohammadreza A. Dehaqani

Arshia Eftekharizadeh, Sepehr Karimi

Fall 2024

Homework 4

- If you have any questions about the homework, don't hesitate to drop an email to the HW Authors.
- 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.

Question 1: Hospital Recovery Status Analysis

Researchers are investigating whether patient recovery outcomes are linked to the type of hospital they were treated in (Private or Public). The summary of the data is provided below:

Recovery Status	Private Hospital	Public Hospital	Total
Fully Recovered	30	25	55
Partially Recovered	20	15	35
Not Recovered	10	10	20
Total	60	50	100

Table 1: *Observed Recovery Outcomes by Hospital Type*

Hypothesis Formulation

- Null Hypothesis (H_0): Recovery outcomes are independent of hospital type.
- Alternative Hypothesis (H_A): Recovery outcomes depend on hospital type.

Independence Test

- Perform a statistical test to determine whether recovery status and hospital type are independent. Use the observed data provided in the table and an appropriate significance level (e.g., $\alpha = 0.05$).
- Compare the test statistic with the critical value or p-value to reach a conclusion.

Critical Evaluation

- Interpret the results of the test and provide insights into the relationship between recovery outcomes and hospital type.
- Address the impact of small expected frequencies, if applicable, and suggest any potential improvements to the study design.

Question 2: Multiple Choice vs Written Response

A study was conducted to determine whether students' ability to solve mathematical problems was influenced by the format of the problems presented to them. Each student was given two sets of problems: one set in a multiple-choice format and another in a written-response format. They were asked to solve as many problems as they could within a fixed amount of time. The number of problems solved in each format was recorded for each student, as shown in the table 2.

Student	**Multiple-Choice**	**Written-Response**
1	25	20
2	30	27
3	15	12
4	32	29
5	28	25
6	24	22
7	29	30
8	26	24
9	31	28
10	33	30
11	22	18
12	12	10

Table 2: Number of problems solved by students in two different formats.

Is there evidence to suggest that students perform differently depending on the format of the problems presented? Analyze the data and provide your conclusion.

Question 3: Teaching Methods and Mathematics Scores

A school is comparing two teaching methods (A and B) to determine their impact on improving mathematics scores. The post-test scores for two groups of 12 students are given below:

$$\text{Group A: } 75, 78, 82, 88, 84, 90, 77, 85, 79, 83, 80, 81$$
$$\text{Group B: } 88, 86, 90, 92, 89, 85, 87, 91, 90, 88, 89, 86$$

Ranking the Data

- Analyze the combined dataset by assigning ranks to all scores, ensuring proper treatment of tied values.

Comparison of Groups

- Evaluate the effectiveness of the two teaching methods using a rank-based statistical test. Summarize the ranks for Group A and compute the test statistic (W).

Hypothesis Framework

- Null Hypothesis (H_0): There is no difference in effectiveness between the teaching methods.
- Alternative Hypothesis (H_A): The effectiveness of the teaching methods differs.
- Utilize an appropriate approximation for the test statistic:

$$Z = \frac{W - \mu_W}{\sigma_W}, \quad \mu_W = \frac{n_A(n_A + n_B + 1)}{2}, \quad \sigma_W = \sqrt{\frac{n_A n_B(n_A + n_B + 1)}{12}}.$$

Analysis and Interpretation

- Assess the results of the hypothesis test at a significance level of $\alpha = 0.05$.
- Reflect on the role of tied ranks and external factors (e.g., classroom environment, prior knowledge) in influencing the observed outcomes.

Question 4: Education vs. Income Relationship

An economist investigates whether there is a monotonic relationship between education level (in years) and monthly income (in 1000$s).Thedatafor12individualsisprovidedbelow:$

Individual	Education (Years)	Monthly Income (×1000)
1	10	2.5
2	12	3.0
3	15	4.0
4	8	1.8
5	16	4.5
6	11	2.8
7	9	2.1
8	14	4.2
9	13	3.5
10	12	3.0
11	14	4.0
12	10	2.6

Table 3: *Education and Income Data*

Exploring Correlation

- Analyze the relationship between education and income using a rank-based method. Compute an appropriate measure of association to evaluate whether a monotonic relationship exists.
- Use the formula:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

  where d_i represents the rank differences.

Hypothesis Testing

- Test the null hypothesis (H_0): No monotonic relationship exists ($\rho_s = 0$).
- Alternative hypothesis (H_A): A monotonic relationship exists ($\rho_s \neq 0$).
- Use a suitable significance level (e.g., $\alpha = 0.05$) to draw a conclusion.

Evaluation

- Assess the results of the test and the strength of the association.
- Consider the impact of ties in the data on the correlation measure and statistical inference.

Question 5: Mean IQ of students

In a study of academic performance in a rural area, researchers found that the median IQ of students who were 16 years of age or older was 107. The following table shows the IQs of a random sample of 15 students from another rural area.

Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
IQ	100	90	135	108	107	119	127	109
Student	**9**	**10**	**11**	**12**	**13**	**14**	**15**	
IQ	105	112	95	123	98	110	120	

Table 4: IQ scores of students from a rural area.

Assuming that the population is symmetric, could the researchers conclude, at the 0.05 level of significance, that the mean IQ of students who are 16 or older from the population of interest is higher than 107?

Question 6: Difference in SmartWatches

A study was conducted to compare the performance of two different brands of smartwatches in terms of step-counting accuracy. Ten participants were randomly selected, and the number of steps counted per minute was recorded for each participant while using both brands of smartwatches. The results are summarized in the table 5.

Participant	**Brand X**	**Brand Y**
Alex	120	118
Bailey	134	136
Casey	128	130
Drew	140	137
Ellis	145	143
Frankie	132	130
Gale	138	141
Harper	125	122
Jordan	130	133
Kelly	135	132

Table 5: Step-counting accuracy (steps per minute) for two smartwatch brands.

Assume that the step-counting accuracy data are not normally distributed. Perform the test to evaluate whether the data provide sufficient evidence at the 5% significance level to conclude that there is a difference in accuracy between the two brands of smartwatches.

Question 7: Coached and Independent group

In a study on skill acquisition in sports, nine participants were selected for a basketball free-throw experiment. Five participants were assigned to a group receiving guided training sessions with an experienced coach, while the remaining four participants practiced independently without guidance. The goal was to assess how coaching influenced their performance under similar practice conditions.

Each participant was tasked with making 10 successful free throws, and the number of attempts required to achieve this milestone was recorded. The results are as follows:

Coached Group	12	15	10	18	14
Independent Group	20	22	17	19	

Table 6: Number of attempts required to achieve 10 successful free throws.

Test if there is a significant difference in the number of attempts required between the coached and independent groups.

Question 8: Advanced Empirical Cumulative Distribution Function (eCDF)

A company records the monthly sales (in \$1000) of 50 branches:

$\{120, 135, 140, 125, 130, 150, 145, 155, 160, 125, 135, 140, 145, 150, 155, 160, 165, 170, 175, 130, 135, 140, 145, 150, 125, 130, 140, 13$

1. Derive the variance of the eCDF $F_n(x)$ using the fact that $F_n(x)$ can be expressed as the mean of n Bernoulli random variables.
2. Write a Python program to:
   (a) Compute the eCDF of the dataset.
   (b) Construct a 95% confidence band for $F(x)$ using the Dvoretzky-Kiefer-Wolfowitz inequality.

   (c) Plot the eCDF and overlay the confidence band.

3. Prove that the Dvoretzky-Kiefer-Wolfowitz inequality guarantees uniform convergence of $F_n(x)$ to $F(x)$ for large n.

4. Discuss the behavior of the confidence band as $n \to \infty$. What are the implications for large versus small sample sizes?

Question 9: Quantile-Quantile (Q-Q) Plots with Transformation

A researcher collects 100 observations on annual rainfall (in mm): $\{800, 850, 870, \ldots, 1200\}$. (Simulate these data using a Gamma distribution with shape parameter $k = 5$ and scale $\theta = 200$.)

1. Derive the theoretical quantiles for the Gamma distribution $Q(p) = F^{-1}(p)$ using its cumulative distribution function.

2. Write a Python program to:

   (a) Simulate the data.

   (b) Compute the sample quantiles.

   (c) Generate a Q-Q plot comparing the sample data against the theoretical quantiles of the Gamma distribution.

   (d) Perform a Box-Cox transformation on the data and re-generate the Q-Q plot to check if the transformation improves linearity.

3. Prove that if the transformed data perfectly matches the theoretical quantiles, the Q-Q plot becomes a straight line with slope 1 and intercept 0.

4. Analyze the effects of choosing incorrect distributional assumptions on the Q-Q plot.

Question 10: Kolmogorov-Smirnov Test with Applications

A bank tracks the waiting times (in minutes) of 150 customers at two branches:

- **Branch A:** Simulate data from Exp($\lambda = 0.1$).
- **Branch B:** Simulate data from Exp($\lambda = 0.15$).

1. Perform the two-sample Kolmogorov-Smirnov test to compare the waiting time distributions. Derive the test statistic:

$$D_n = \sup_x |F_A(x) - F_B(x)|$$

   and its asymptotic distribution under the null hypothesis.

2. Write a Python program to:

   (a) Simulate the two datasets.

   (b) Compute the empirical CDFs for Branch A and Branch B.

   (c) Compute the Kolmogorov-Smirnov statistic and compare it to the critical value.

   (d) Visualize the CDFs and highlight the maximum deviation.

3. Prove that the Kolmogorov-Smirnov test is distribution-free under the null hypothesis.

4. Discuss the sensitivity of the Kolmogorov-Smirnov test to large differences in the tails of the distributions.

Question 11: (Optional) KDE with Bandwidth Selection

A factory records the lifetimes (in hours) of 200 machine parts: $\{50, 60, 65, 80, \ldots, 300\}$. (Simulate these data using an exponential distribution with rate $\lambda = 0.01$.)

1. Derive the asymptotic mean integrated squared error (AMISE) for a kernel density estimate $\hat{\pi}_h(x)$ with a Gaussian kernel.
2. Write a Python program to:
   (a) Simulate the dataset.
   (b) Compute the kernel density estimate using bandwidths $h = 1, 5, 10$.
   (c) Implement Silverman's rule of thumb to select an optimal bandwidth.
   (d) Plot the KDEs for each bandwidth and overlay the histogram of the data.
3. Prove that the optimal bandwidth h^* minimizes the AMISE for smooth distributions.
4. Discuss the implications of bandwidth selection in cases where the true density is multimodal.

Question 12: (Optional) Outlier Impact on Steel Strength Analysis

A team of materials scientists is testing the strength (in MPa) of a newly developed alloy to ensure it meets manufacturing standards. The recorded strength values from 20 samples are:

$$\{220, 225, 230, 240, 250, 260, 270, 275, 280, 285, 290, 295, 300, 310, 315, 320, 325, 330, 335, 340\}.$$

1. Define the 10% trimmed mean and explain its advantage in reducing sensitivity to extreme outliers.
2. Calculate the 10% trimmed mean of the dataset and compare it with the arithmetic mean.
3. A measurement error introduces a severe outlier: 450. Recompute the arithmetic mean and 10% trimmed mean, and discuss the robustness of the trimmed mean against outliers.
4. Prove that as the trimming percentage $\alpha \to 50\%$, the trimmed mean converges to the sample median.

Question 13: (Optional) Robust Weight Analysis

A packaging company monitors the weight (in grams) of products to maintain consistency. For one batch, the recorded weights are:

$$\{95, 100, 105, 110, 115, 95, 100, 105, 110, 115, 300, 305, 310, 315, 320\}.$$

1. Define an M-estimate as the minimizer of the loss function:
   $$y^* = \arg\min_y \sum_{i=1}^{n} \psi(x_i, y),$$
   where $\psi(x_i, y) = |x_i - y|$ for the median and $\psi(x_i, y) = (x_i - y)^2$ for the mean.
2. Derive the M-estimates for the mean and median of the dataset.
3. Introduce a robust weighting function $\psi(x_i, y) = |x_i - y|^p$ and explain how $p > 1$ affects the sensitivity to extreme weights.
4. Discuss how adjusting the weights downplays the influence of extreme weights (e.g., above 300) and improves the robustness of the M-estimates.

Question 14: (Optional) Market Volatility Analysis Using KS Test

A financial analyst is comparing the daily returns of two stock indices, S&P500 and NASDAQ, to assess market volatility. The daily returns are assumed to follow:

- S&P500: $N(\mu = 0.001, \sigma = 0.02)$,
- NASDAQ: $N(\mu = 0.0015, \sigma = 0.03)$.

1. (Data Simulation) Simulate 250 daily returns for each stock index.
2. Compute the empirical cumulative distribution functions (eCDFs) for both indices.
3. Perform a Kolmogorov-Smirnov test to determine if the distributions differ significantly.
4. Plot the eCDFs and highlight the maximum vertical deviation between them.
5. Test the null hypothesis H_0: "The two indices have the same return distribution" at $\alpha = 0.05$. Indicate whether the null hypothesis is rejected and visualize the critical region.

Question 15: (Optional) Success Rates of Promotional Campaigns

A marketing team compares the success rates of two promotional campaigns to determine which is more effective. The first campaign targeted 200 participants, of whom 120 made a purchase. The second campaign targeted 250 participants, of whom 140 made a purchase.

1. Write a Python program to test whether there is a significant difference in the success rates of the two campaigns.
2. Compute and display the confidence interval for the difference in proportions.
3. Incorporate a continuity correction in the calculations.
4. Based on the analysis, conclude whether the success rates of the two campaigns are significantly different at the $\alpha = 0.05$ level.