Introduction to Statistical Inference

Instructor: Mohammadreza A. Dehaqani

Arshia Eftekharizadeh, Sepehr Karimi

Winter 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.

Solution

Step 1: Hypothesis Formulation

Null Hypothesis (H_0):
Recovery outcomes are independent of hospital type, meaning that patients' chances of full, partial, or no recovery do not depend on whether they were treated in a private or public hospital.

Alternative Hypothesis (H_a):
Recovery outcomes depend on hospital type, implying that the hospital type influences whether patients fully, partially, or do not recover.

Step 2: Expected Frequency Calculation

To determine if there is a statistically significant relationship, we compute the expected frequencies under the assumption that hospital type does not affect recovery outcomes.

The expected frequency for each category in a contingency table is calculated as:

$$E_{ij} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

Using the corrected total sample size (110) instead of 100, we apply this formula to each cell in the table. For example, the expected frequency for Fully Recovered in Private Hospital:

$$E = \frac{55 \times 60}{110} = 30.00$$

Similarly, expected frequencies for other categories are calculated.

The final expected frequency table is:

Recovery Status	Private Hospital	Public Hospital
Fully Recovered	30.00	25.00
Partially Recovered	19.09	15.91
Not Recovered	10.91	9.09

Table 2: Expected Frequency Table

Step 3: Compute Chi-Square Test Statistic

The Chi-Square statistic quantifies how much the observed frequencies deviate from the expected frequencies:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where:

- O = Observed frequency
- E = Expected frequency

This calculation is applied to each cell:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_n - E_n)^2}{E_n}$$

After applying this to all categories, we obtain:

$$\chi^2 = 0.2619$$

Step 4: Determine p-value

The p-value tells us the probability of observing this level of deviation if the null hypothesis were true.

Degrees of Freedom (df) for a contingency table is:

$$(\text{Number of Rows} - 1) \times (\text{Number of Columns} - 1) = (3 - 1) \times (2 - 1) = 2$$

Using a Chi-Square distribution table or statistical computation, the p-value corresponding to $\chi^2 = 0.2619$ with $df = 2$ is:

$$p = 0.8773$$

Step 5: Decision Rule and Conclusion

The significance level is set at $\alpha = 0.05$.

Since $p = 0.8773$ is greater than 0.05, we fail to reject the null hypothesis.

Interpretation:

- There is no statistically significant relationship between hospital type and recovery outcome.
- The differences observed in recovery rates are likely due to random variation rather than a systematic effect of hospital type.
- The independence assumption holds, meaning that whether a patient is treated in a private or public hospital does not significantly affect their likelihood of full, partial, or no recovery.

Critical Evaluation and Study Improvements

Sample Size Consideration

- A larger dataset could provide more statistical power.
- A small sample might fail to detect subtle effects.

Expected Frequency Concerns

- The expected counts are all above 5, ensuring the Chi-Square test is valid.
- If any expected frequency was below 5, an alternative test such as Fisher's Exact Test might be preferable.

Additional Factors

- The study currently only considers hospital type.
- Other variables like patient age, illness severity, treatment methods, or doctor expertise could influence recovery.

Alternative Statistical Models

- A logistic regression model could be used to predict recovery outcomes while controlling for multiple confounding factors.
- A Bayesian approach could provide insights with uncertainty quantification.

Question 2: Solving Problems

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 3: 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.

Solution

We apply the Wilcoxon signed-rank test to determine if there is a significant difference in the number of problems solved between the multiple-choice and written-response formats.

1. **Justification for Using the Wilcoxon Signed-Rank Test:**
   - The data consists of **paired observations** since each student attempts both formats.
   - We compute the differences between the two formats:
   $$d_i = X_i - Y_i = [5, 3, 3, 3, 3, 2, -1, 2, 3, 3, 4, 2]$$
   - The differences are **not all zero** and appear **reasonably symmetric**.
   - Since the normality of differences is not assumed, the **Wilcoxon signed-rank test** is appropriate.
2. **Hypotheses:**
   - H_0: There is no difference in the number of problems solved between the two formats.
   - H_A: There is a difference in the number of problems solved between the two formats.
3. **P-value:** Using statistical software, we find:
   $$p = 0.00098$$
4. **Decision:** Since $p = 0.00098 < 0.05$, we reject H_0. This indicates that there is a significant difference in student performance depending on the problem format.
5. **Conclusion:** Students performed significantly better in the **multiple-choice format** compared to the written-response format.

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.

Solution

Step 1: Hypothesis Formulation

We use a rank-based non-parametric test since the sample size is small ($n = 12$ per group), and we do not assume normality.

Null Hypothesis (H_0):
There is no difference in effectiveness between the teaching methods. The distribution of scores in both groups is similar.

Alternative Hypothesis (H_A):
The effectiveness of the two teaching methods differs, meaning that one method leads to significantly different scores.

We will use two methods to test this:

- Mann-Whitney U test (to compare distributions directly).
- Wilcoxon Rank Sum Test (Z-approximation) (a large-sample approximation of the rank-sum test).

Step 2: Ranking the Data

Step 2.1: Combine Scores from Both Groups The given data:

Group A (Teaching Method A):

75, 78, 82, 88, 84, 90, 77, 85, 79, 83, 80, 81

Group B (Teaching Method B):

88, 86, 90, 92, 89, 85, 87, 91, 90, 88, 89, 86

We merge both datasets and assign ranks to all values in ascending order.

Step 2.2: Handling Tied Ranks If scores are repeated (e.g., 88 appears multiple times), we assign the average rank for those values.

After ranking the combined dataset, we extract the ranks for each group.

$$W_A = 92.5$$

$$W_B = \text{Sum of remaining ranks}$$

Step 3: Mann-Whitney U Test

The Mann-Whitney U statistic is computed as:

$$U_A = n_A n_B + \frac{n_A(n_A+1)}{2} - W_A$$

where:

- $n_A = 12$ (size of Group A)
- $n_B = 12$ (size of Group B)
- $W_A = 92.5$ (sum of ranks for Group A)

After applying the formula, we get:

$$U_A = 14.5$$

p-value from Mann-Whitney U test:

$$p = 0.00097$$

Since $p < 0.05$, we reject H_0, indicating a significant difference between the groups.

Step 4: Z-Score Approximation (Wilcoxon Rank Sum Test)

For larger samples, the rank sum test statistic (W_A) follows a normal distribution with:

$$\mu_W = \frac{n_A(n_A+n_B+1)}{2}$$

$$\sigma_W = \sqrt{\frac{n_A n_B(n_A+n_B+1)}{12}}$$

$$Z = \frac{W_A - \mu_W}{\sigma_W}$$

After calculations:

$$Z = -3.32$$

p-value from Z-test:

$$p = 0.0009$$

Since $p < 0.05$, we again reject H_0.

Step 5: Conclusion and Interpretation

- Both tests confirm a statistically significant difference between the teaching methods.
- Since Teaching Method B consistently has higher ranks, it outperforms Method A in improving students' math scores.
- The effect size is strong, indicating a meaningful difference rather than random variation.

Critical Evaluation and Study Improvements

Effect of Tied Ranks

- Some scores were repeated (e.g., 88, 90), requiring adjustments in ranking.
- The Wilcoxon and Mann-Whitney tests correctly handle ties, ensuring an accurate result.

Possible Confounding Factors

- **Prior Knowledge:** If students had different baseline skill levels, results could be biased.
- **Teaching Environment:** Variations in teacher experience, classroom settings, and student engagement could impact outcomes.

Recommendations for Future Research

- **Increase Sample Size:** A larger dataset would improve statistical power.
- **Pre-Test and Post-Test Design:** Measuring improvements within each group before and after the teaching method would provide more robust insights.
- **Use of Effect Size Metrics:** A Cliff's delta or Rank-Biserial Correlation could quantify the practical impact of the difference.

Question 4: Education vs. Income Relationship

An economist investigates whether there is a monotonic relationship between education level (in years) and monthly income (in 1000s).$The data for 12 individuals is provided below$:

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.

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 4: *Education and Income Data*

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.

Solution

Step 1: Compute Spearman's Rank Correlation (r_s)

Since the data may not be normally distributed, we use Spearman's rank correlation coefficient to measure the strength and direction of the monotonic relationship.

Step 1.1: Assign Ranks Education Years and Monthly Income are ranked separately.
If there are tied values, their average rank is assigned.

Step 1.2: Compute Rank Differences (d_i) Compute differences between the ranks of education years and income for each individual.
Square each rank difference and sum them.

Step 1.3: Apply Spearman's Formula The formula for Spearman's rank correlation coefficient:

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

where:

- d_i = rank difference for individual i
- n = number of individuals

After computing, we obtain:

$$r_s = 0.9825$$

Step 2: Hypothesis Testing

We test:

Null Hypothesis (H_0): No monotonic relationship exists ($\rho_s = 0$).

Alternative Hypothesis (H_A): A monotonic relationship exists ($\rho_s \neq 0$).

Using $\alpha = 0.05$, we check the p-value:

$$p = 1.30 \times 10^{-8}$$

Since $p < 0.05$, we reject H_0, indicating a strong and statistically significant monotonic relationship.

Step 3: Interpretation and Evaluation

Strength of Association

- $r_s = 0.9825$ is very close to 1, meaning an extremely strong positive monotonic relationship between education and income.
- More education is strongly associated with higher income.

Impact of Tied Ranks

- There were some ties in education and income.
- The Spearman formula used average ranks, which is an appropriate correction.
- The high correlation suggests ties had little impact on the result.

Limitations and Further Considerations

- Correlation does not imply causation—higher education might not directly cause higher income.
- Other factors (experience, job market conditions) could influence income.
- A larger sample across different demographics could improve reliability.

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 5: 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?

Solution

1. **Hypotheses:**

$$H_0 : \mu = 107 \quad \text{against} \quad H_1 : \mu > 107$$

2. **Test statistic:**

From the data, we calculate:

$$W^+ = 62, \quad W^- = 28, \quad W^+ + W^- = \frac{15 \cdot 16}{2} = 120$$

Using the smaller of the two sums:

$$W = \min(W^+, W^-) = 28$$

3. **Assumption:**

The population is symmetric.

4. ***p*-value:**

Since there are no ties, the normal approximation is used.

The expected value and variance of W^+ are:

$$E(W^+) = \frac{n(n+1)}{4} = \frac{15(15+1)}{4} = 60$$

$$\text{Var}(W^+) = \frac{n(n+1)(2n+1)}{24} = \frac{15(15+1)(2 \cdot 15+1)}{24} = 120$$

The standardized test statistic is:

$$Z = \frac{W^+ - E(W^+) - 0.5}{\sqrt{\text{Var}(W^+)}} = \frac{62 - 60 - 0.5}{\sqrt{120}} \approx \frac{1.5}{10.95} \approx 0.137$$

Using the standard normal table:

$$p = 1 - \Phi(Z) = 1 - \Phi(0.137) \approx 0.445$$

5. **Decision:**

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

6. **Conclusion:**

At the 0.05 level of significance, there is insufficient evidence to conclude that the mean IQ of students in the rural area 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.
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.

Solution

To compare the step-counting accuracy between two smartwatch brands, we use the **Sign Test**, assuming that the data are not normally distributed.

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

1. Calculate the Differences For each participant, compute the difference between the step counts recorded by Brand X and Brand Y:

$$\text{Difference} = \text{Brand X} - \text{Brand Y}$$

The differences are: $2, -2, -2, 3, 2, 2, -3, 3, -3, 3$.

2. Count the Signs

- Positive differences (+): 6 (for Alex, Drew, Ellis, Frankie, Harper, Kelly)
- Negative differences (−): 4 (for Bailey, Casey, Gale, Jordan)
- Zero differences: None.

Thus, $n = 10$, and the test statistic is $x = 6$ (the number of positive differences).

3. Hypotheses

$$H_0 : p_+ = \frac{1}{2} \quad \text{(No difference in accuracy between the two brands).}$$

$$H_1 : p_+ \neq \frac{1}{2} \quad \text{(A difference in accuracy exists between the two brands).}$$

4. Test Statistic Under H_0, the number of positive differences follows a binomial distribution:

$$X \sim B(n = 10, p = 0.5)$$

We calculate the two-tailed p-value:

$$P = 2 \cdot P(X \geq 6)$$

5. Calculate the p-value Using the binomial formula:

$$P(X = k) = \binom{n}{k} (0.5)^k (0.5)^{n-k}$$

We calculate the terms for $X \geq 6$:

$$P(X = 6) = \binom{10}{6} (0.5)^{10} = 210 \cdot (0.0009765625) = 0.205$$

$$P(X = 7) = \binom{10}{7} (0.5)^{10} = 120 \cdot (0.0009765625) = 0.117$$

$$P(X = 8) = \binom{10}{8} (0.5)^{10} = 45 \cdot (0.0009765625) = 0.044$$

$$P(X = 9) = \binom{10}{9} (0.5)^{10} = 10 \cdot (0.0009765625) = 0.01$$

$$P(X = 10) = \binom{10}{10}(0.5)^{10} = 1 \cdot (0.0009765625) = 0.001$$
$$P(X \geq 6) = 0.205 + 0.117 + 0.044 + 0.01 + 0.001 = 0.377$$
$$P = 2 \cdot 0.377 = 0.754$$

6. Conclusion At the 5% significance level ($\alpha = 0.05$):

- The p-value (0.754) is greater than 0.05.
- We fail to reject H_0.

Decision: There is not enough evidence to conclude a significant difference in step-counting accuracy between the two smartwatch brands.

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 7: 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.

SOLUTION

The ranks for the combined sample are

Coached	12	15	10	18	14	**Independent**	20	22
17	19							
Ranks	2	4	1	6	3	**Ranks**	8	9
5	7							

We look at the ranks for the smaller sample. The Wilcoxon rank-sum test is

1. **Hypotheses:** $H_0 : \mu_x = \mu_y$ vs $H_A : \mu_x \neq \mu_y$.
2. **Test statistic:**
$$W_y = 5 + 7 + 8 + 9 = 29$$
3. **Assumption:** X_i and Y_i follow the same kind of distribution, differing only by a shift.
4. **P-value:** $2 \times \Pr(W \geq 29)$, calculated using statistical software as:
$$p = 0.0317$$
where the expected rank sum is:
$$E(W_y) = \frac{n_y(N+1)}{2} = \frac{4 \times 10}{2} = 20$$
Since $W_y = 29 > E(W_y) = 20$, the independent group required more attempts on average.
5. **Decision:** Since $p = 0.0317 < 0.05$, we reject H_0. This means that there is a significant difference in the number of attempts required between the coached and independent groups. The coached group required significantly fewer attempts, indicating that coaching had a positive impact.

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?

Solution

Step 1: Deriving the Variance of $F_n(x)$

The empirical cumulative distribution function (eCDF) is defined as:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

where $I(X_i \leq x)$ is an indicator function that takes the value 1 if $X_i \leq x$ and 0 otherwise.
Each term in the summation follows a Bernoulli distribution with:

$$P(I(X_i \leq x) = 1) = F(x)$$

Since $F_n(x)$ is the mean of n i.i.d. Bernoulli random variables, its expectation and variance are:

$$E[F_n(x)] = F(x)$$

$$\text{Var}[F_n(x)] = \frac{F(x)(1 - F(x))}{n}$$

Thus, the variance of the empirical distribution function decreases as n increases.

Step 2: Constructing and Plotting the eCDF and Confidence Bands

Using the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality, we can construct a 95% confidence band around the empirical CDF.

Step 2.1: Compute the eCDF

1. Sort the dataset.
2. Compute the proportion of observations $\leq x$ for each data point.

Step 2.2: Apply the DKW Inequality The DKW inequality provides an upper bound for the maximum deviation between the empirical CDF and the true CDF:

$$P\left(\sup_x |F_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

For 95% confidence, solving for ϵ:

$$\epsilon = \sqrt{\frac{\ln(40)}{2n}}$$

The confidence band is then:

$$F_n(x) - \epsilon \leq F(x) \leq F_n(x) + \epsilon$$

Step 2.3: Plot the eCDF with Confidence Bands

- The eCDF is plotted as a step function.
- The upper and lower confidence bounds ($F_n(x) \pm \epsilon$) are overlaid as dashed lines.

Step 3: Proof of Uniform Convergence

To prove uniform convergence, we show:

$$\sup_x |F_n(x) - F(x)| \to 0 \quad \text{as} \quad n \to \infty$$

From the DKW inequality:

$$P\left(\sup_x |F_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

For any fixed $\epsilon > 0$, as $n \to \infty$, the right-hand side converges to 0, implying:

$$\sup_x |F_n(x) - F(x)| \to 0 \quad \text{(almost surely)}$$

Thus, the eCDF converges uniformly to the true CDF as the sample size increases.

Step 4: Behavior of Confidence Bands as $n \to \infty$

As $n \to \infty$, the width of the confidence band 2ϵ shrinks. This means that the eCDF becomes a more accurate estimate of the true CDF.

- For small n, the bands are wide, indicating greater uncertainty in estimation.
- For large samples: The eCDF is a highly reliable approximation of the true CDF.
- For small samples: The confidence band is wide, meaning greater variability in estimation.

Practical Takeaway: Large sample sizes reduce uncertainty and provide a better approximation of the underlying distribution.

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.

Solution

Step 1: Deriving the Theoretical Quantiles for the Gamma Distribution

A quantile function is the inverse of the cumulative distribution function (CDF). Given a Gamma-distributed random variable:

$$X \sim \text{Gamma}(k, \theta)$$

where:

- k is the shape parameter,
- θ is the scale parameter.

Step 1.1: Cumulative Distribution Function (CDF) The Gamma CDF is:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{\gamma(k, x/\theta)}{\Gamma(k)}, & x \geq 0 \end{cases}$$

where:

- $\Gamma(k)$ is the Gamma function,
- $\gamma(k, x/\theta)$ is the lower incomplete gamma function:

$$\gamma(k, x/\theta) = \int_0^{x/\theta} t^{k-1} e^{-t} dt$$

Step 1.2: Quantile Function (Inverse CDF) The quantile function $Q_X(p)$ satisfies:

$$Q_X(p) = F_X^{-1}(p)$$

Using the reference proof, we obtain:

$$Q_X(p) = \begin{cases} -\infty, & p = 0 \\ \gamma^{-1}(k, \Gamma(k) \cdot p)\theta, & p > 0 \end{cases}$$

where $\gamma^{-1}(k, y)$ is the inverse of the lower incomplete Gamma function. This must be computed numerically in practice.

Step 2: Simulation, Empirical Quantiles, and Q-Q Plot

Step 2.1: Simulating Data from a Gamma Distribution

- Generate 100 observations from a Gamma$(5, 200)$ distribution.

Step 2.2: Compute Sample Quantiles

- Sort the empirical data and compute its empirical quantiles.

Step 2.3: Compute Theoretical Quantiles

$$Q_X(p) = \gamma^{-1}(5, \Gamma(5) \cdot p) \cdot 200$$

Compare empirical quantiles to theoretical values.

Step 2.4: Generate a Q-Q Plot

- Plot sample quantiles vs. theoretical quantiles.
- If the data follows the assumed distribution, points should lie on a straight line.

Step 2.5: Box-Cox Transformation and Q-Q Plot

- Apply Box-Cox transformation to improve normality:

$$X_{\text{transformed}} = \begin{cases} \frac{X^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(X), & \lambda = 0 \end{cases}$$

- Generate a new Q-Q plot to check if the transformation improves linearity.

Step 3: Proof - Why a Perfect Q-Q Plot is a Straight Line

Step 3.1: Given Condition From the Q-Q plot definition, we compare:

- E_k: The empirical quantile at rank k.
- Q_k: The theoretical quantile at rank k.

If the empirical data perfectly follows the theoretical distribution:

$$E_k = Q_k \quad \forall k = 1, 2, \ldots, n$$

This means that each Q-Q plot point is:

$$(Q_k, E_k) = (Q_k, Q_k)$$

Step 3.2: Line Equation For any point (Q_k, Q_k), let:

$$x = Q_k \quad \text{(theoretical quantile)}, \quad y = E_k \quad \text{(empirical quantile)}$$

Since $E_k = Q_k$, we obtain:

$$y - x = 0 \Rightarrow y = x$$

Thus, all Q-Q points satisfy the equation of a straight line.

Step 3.3: Q-Q Plot Definition The Q-Q plot consists of all ordered pairs:

$$\{(Q_k, E_k) \mid k = 1, 2, \ldots, n\}$$

Since $E_k = Q_k$ for all k, this simplifies to:

$$\{(Q_k, Q_k) \mid k = 1, 2, \ldots, n\}$$

Step 3.4: Subset Relation Since every point satisfies $y = x$, we conclude:

$$\{(Q_k, Q_k) \mid k = 1, 2, \ldots, n\} \subseteq \{(x, y) \mid y = x\}$$

This shows that all Q-Q points lie exactly on the 45-degree reference line.

Step 3.5: Line Properties The line equation $y = x$ in the (x, y)-plane has:

- Slope $= 1$
- Intercept $= 0$

Thus, in a perfect Q-Q plot, the empirical quantiles match the theoretical quantiles exactly, forming a straight line with slope 1 and intercept 0.

Step 4: Effects of Incorrect Distributional Assumptions

If the wrong theoretical distribution is assumed, the Q-Q plot deviates from the ideal straight line:

Non-Linearity

- The Q-Q plot deviates from a straight line, indicating a mismatch between the empirical and theoretical distributions.

Curvature in the Tails

- If the assumed distribution underestimates extreme values, the Q-Q plot curves upward.
- If it overestimates extreme values, the Q-Q plot curves downward.

Different Slope

- A slope $\neq 1$ suggests a difference in spread (variance) between the sample and assumed distribution.

Shifted Intercept

- A nonzero intercept means the assumed distribution is miscentered.

Conclusion: Analyzing the Q-Q plot helps diagnose distributional assumptions and whether transformations (e.g., Box-Cox) are needed.

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.

Solution

Step 1: Kolmogorov-Smirnov (KS) Test and Derivation of the Test Statistic

The Kolmogorov-Smirnov (KS) test is a non-parametric test used to compare two empirical distributions. It is based on the maximum absolute difference between the empirical CDFs of two datasets.

Step 1.1: Definition of the KS Test Statistic For two independent samples:

- **Branch A:** $X_1, X_2, \ldots, X_n \sim \text{Exp}(\lambda_A)$, where $\lambda_A = 0.1$.
- **Branch B:** $Y_1, Y_2, \ldots, Y_m \sim \text{Exp}(\lambda_B)$, where $\lambda_B = 0.15$.

Define the empirical cumulative distribution functions (eCDFs):

$$F_A(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x), \quad F_B(x) = \frac{1}{m}\sum_{i=1}^{m} I(Y_i \leq x)$$

The KS test statistic is:

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

where $\sup_x$ denotes the supremum (maximum absolute difference) over all x.

Step 1.2: Asymptotic Distribution Under the Null Hypothesis Under the null hypothesis (H_0):

$$F_A(x) = F_B(x) \quad \forall x$$

i.e., the two distributions are identical.
For large n and m, the KS statistic follows the Kolmogorov distribution:

$$\Pr(D_n > d) \approx 2e^{-2n_{\text{eff}}d^2}$$

where:

$$n_{\text{eff}} = \frac{nm}{n+m}$$

The critical value is derived from:

$$D_n > \frac{c(\alpha)}{\sqrt{n_{\text{eff}}}}$$

where $c(\alpha)$ is a tabulated critical value depending on the significance level α.

Step 2: Python Implementation (Simulation and KS Test)

A Python program is used to:

1. **Simulate Data:**
   - Generate 150 samples for Branch A ($\lambda = 0.1$).
   - Generate 150 samples for Branch B ($\lambda = 0.15$).
2. **Compute Empirical CDFs:**
   - Compute $F_A(x)$ and $F_B(x)$ using the sorted data.
3. **Compute KS Statistic:**
   - Find $D_n = \sup_x |F_A(x) - F_B(x)|$.
   - Compare D_n with the critical value.
4. **Visualize Results:**
   - Plot the empirical CDFs.
   - Highlight the maximum deviation D_n on the graph.

Step 3: Proof That the KS Test is Distribution-Free Under H_0

A test is distribution-free if its null distribution does not depend on the underlying distribution.
Under H_0, the two samples are drawn from the same continuous distribution $F(x)$.
The empirical CDFs $F_A(x)$ and $F_B(x)$ converge to $F(x)$.
By the Glivenko-Cantelli theorem:

$$\sup_x |F_n(x) - F(x)| \to 0 \quad \text{almost surely}$$

Transform each sample into uniform values:

$$U_i = F(X_i), \quad V_i = F(Y_i)$$

where $U_i \sim \text{Uniform}(0,1)$ and $V_i \sim \text{Uniform}(0,1)$.
Since the KS test is based on:

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

and the transformation does not change its structure, the distribution of D_n under H_0 remains the same regardless of $F(x)$.

Thus, the KS test is distribution-free under H_0, meaning its significance level does not depend on the original distribution.

Step 4: Sensitivity of the KS Test to Tail Differences

The KS test is most sensitive to central differences and less sensitive to differences in the tails.

Case 1: Small Sample Size

- If n, m are small, the KS test may not detect subtle tail differences.
- Large tail discrepancies might be ignored since the supremum focuses on the largest difference, which is often near the center of the distribution.

Case 2: Large Sample Size

- As $n, m \to \infty$, the test is more powerful and can detect even small differences.
- The tail behavior becomes significant in large samples.

Alternative Tests for Tail Sensitivity

- **Cramér–von Mises Test:** More sensitive to small overall differences.
- **Anderson-Darling Test:** Places more weight on the tails.
- **Quantile-Quantile (Q-Q) Plot:** Useful for visualizing tail differences.

Thus, while the KS test is effective, it may not be the best choice if tail behavior is a primary concern.

Question 11: (Optional) Kernel Probability Density Estimates 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.

Solution

Step 1: Deriving the Asymptotic Mean Integrated Squared Error (AMISE) for a Kernel Density Estimate

The Kernel Density Estimate (KDE) is a non-parametric estimator for the probability density function (PDF) of a dataset. Given a sample $X_1, X_2, \ldots, X_n$, the KDE is defined as:

$$\hat{\pi}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where:

- $K(x)$ is the kernel function (typically Gaussian: $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$),
- h is the bandwidth that controls smoothness.

The Asymptotic Mean Integrated Squared Error (AMISE) is used to evaluate the quality of KDE:

$$AMISE(h) = \frac{R(K)}{nh} + \frac{1}{4} h^4 R(f'')$$

where:

- $R(K) = \int K^2(x)dx$ is a constant depending on the kernel function,
- $R(f'') = \int (f''(x))^2 dx$ quantifies the smoothness of the true density $f(x)$,
- The first term $\frac{R(K)}{nh}$ is the variance component, decreasing with larger h,
- The second term $\frac{1}{4} h^4 R(f'')$ is the bias component, increasing with h.

Step 1.1: AMISE for the Gaussian Kernel For a Gaussian kernel, $K(x)$ is the standard normal PDF, and:

$$R(K) = \frac{1}{2\sqrt{\pi}}$$

Thus, the AMISE equation simplifies to:

$$AMISE(h) = \frac{1}{2\sqrt{\pi} nh} + \frac{1}{4} h^4 R(f'')$$

Step 2: Python Implementation of KDE with Different Bandwidths

A Python program is used to:

1. **Simulate the Dataset:**
   - Generate 200 samples from an Exp($\lambda = 0.01$) distribution.
2. **Compute the KDE with Different Bandwidths:**
   - Use a Gaussian kernel with bandwidths $h = 1, 5, 10$.
3. **Implement Silverman's Rule of Thumb:**
   - Optimal bandwidth h^* is estimated using Silverman's formula:

$$h^* = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}}$$

   where $\hat{\sigma}$ is the standard deviation of the dataset.
4. **Plot KDEs and Histogram:**
   - Visualize the KDEs with different bandwidths and overlay the histogram.

Step 3: Proof That the Optimal Bandwidth Minimizes AMISE

To find the optimal bandwidth h^*, we minimize $AMISE(h)$.

Step 3.1: Take the Derivative of AMISE The AMISE function:

$$AMISE(h) = \frac{R(K)}{nh} + \frac{1}{4}h^4 R(f'')$$

Taking the derivative with respect to h:

$$\frac{d}{dh}AMISE(h) = -\frac{R(K)}{nh^2} + h^3 R(f'')$$

Setting $\frac{d}{dh}AMISE(h) = 0$ for minimization:

$$\frac{R(K)}{nh^2} = h^3 R(f'')$$

Solving for h:

$$h^* = \left(\frac{R(K)}{nR(f'')}\right)^{\frac{1}{5}}$$

This formula provides the optimal smoothing bandwidth for minimizing AMISE.

Step 4: Implications of Bandwidth Selection for Multimodal Distributions

The choice of bandwidth h in KDE significantly affects the estimated density function.

Case 1: Small Bandwidth (h too low)

- KDE is too sensitive, capturing random noise.
- Produces overfitting with too many peaks.
- Not ideal for generalizing patterns.

Case 2: Large Bandwidth (h too high)

- KDE over-smooths the data.
- Can miss important structure in multimodal distributions.
- Fails to capture separate peaks in the true density.

Case 3: True Density is Multimodal

- A small h is necessary to detect separate peaks.
- However, if h is too small, spurious peaks (from noise) appear.
- Adaptive methods, such as variable bandwidth KDE, can improve multimodal estimation.

Alternative Approaches for Multimodal Densities

- Silverman's Rule is optimal for unimodal distributions but may oversmooth multimodal data.
- Cross-validation methods can fine-tune h based on multimodal characteristics.

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.

Solution

Step 1: Definition of the 10% Trimmed Mean and Its Advantages

The 10% trimmed mean is a robust measure of central tendency where the lowest and highest 10% of values are removed before computing the mean.

Advantages of the Trimmed Mean

- **Reduces Sensitivity to Outliers:** Unlike the arithmetic mean, which is affected by extreme values, the trimmed mean removes them before calculation.
- **More Robust in Skewed Distributions:** If a dataset has a long tail or a few extreme values, the trimmed mean provides a better measure of central tendency.
- **Better for Small Sample Sizes:** When dealing with limited data points, extreme values can skew results. The trimmed mean mitigates this impact.

Step 2: Compute the 10% Trimmed Mean and Compare It to the Arithmetic Mean

Given Data

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

Sample Size: $n = 20$

10% Trimming

- Remove the smallest 10% (2 values): 220, 225
- Remove the largest 10% (2 values): 335, 340
- Remaining dataset:

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

Calculations Arithmetic Mean:

$$\text{Mean} = \frac{\sum X_i}{n} = \frac{220 + 225 + 230 + \cdots + 340}{20} = 284.75$$

10% Trimmed Mean:

$$\text{Trimmed Mean} = \frac{\sum X_i}{n_{\text{trimmed}}} = \frac{230 + 240 + 250 + \cdots + 330}{16} = 285.94$$

Conclusion

- The trimmed mean (285.94) is close to the arithmetic mean (284.75), confirming that the dataset has no extreme outliers.

Step 3: Effect of an Outlier on the Mean and Trimmed Mean

New Dataset with Outlier (450 Introduced)

$$\{220, 225, 230, 240, 250, 260, 270, 275, 280, 285, 290, 295, 300, 310, 315, 320, 325, 330, 335, 340, 450\}$$

Sample Size: $n = 21$

10% Trimming

- Remove the smallest 10% (2 values): 220, 225
- Remove the largest 10% (2 values): 340, 450
- Remaining dataset:

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

Calculations New Arithmetic Mean:

$$\text{New Mean} = \frac{\sum X_i}{n} = \frac{220 + 225 + 230 + \cdots + 450}{21} = 292.62$$

New 10% Trimmed Mean:

$$\text{Trimmed Mean} = \frac{\sum X_i}{n_{\text{trimmed}}} = \frac{230 + 240 + 250 + \cdots + 330}{16} = 288.82$$

Impact of the Outlier

- The arithmetic mean increased significantly from 284.75 to 292.62 due to the outlier.
- The trimmed mean remained stable, changing only slightly from 285.94 to 288.82.

Conclusion

- The trimmed mean is robust against extreme values, while the arithmetic mean is sensitive to them.

Step 4: Proof That the Trimmed Mean Converges to the Median as $\alpha \to 50\%$

The trimmed mean is defined as:

$$\bar{X}_{\text{trim}} = \frac{1}{(1-2\alpha)n} \sum_{i=\lfloor \alpha n \rfloor + 1}^{n - \lfloor \alpha n \rfloor} X_i$$

Step 4.1: As Trimming Increases

- $\alpha = 10\% \Rightarrow 10\%$ of values are removed from both ends.
- $\alpha = 25\% \Rightarrow 25\%$ of values are removed.
- $\alpha = 50\% \Rightarrow$ Only the middle values remain.

Step 4.2: Two Cases If n is Odd

- Only one value remains after trimming:

$$\text{Trimmed Mean} = X_{\frac{n+1}{2}} = \text{Median}$$

If n is Even

- Two middle values remain, and their average is taken:

$$\text{Trimmed Mean} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} = \text{Median}$$

Thus, mathematically:

$$\lim_{\alpha \to 50\%} \bar{X}_{\text{trim}} = \text{Median}(X)$$

Step 4.3: Verify with Data Median of Original Data:

$$\text{Median} = \frac{290 + 295}{2} = 287.5$$

As trimming removes more extreme values, the trimmed mean approaches 287.5, confirming the proof.

Final Answers

- **Arithmetic Mean (Original Data)** = 284.75
- **10% Trimmed Mean (Original Data)** = 285.94
- **Arithmetic Mean (With Outlier 450)** = 292.62
- **10% Trimmed Mean (With Outlier 450)** = 288.82

Proof Conclusion As $\alpha \to 50\%$, the trimmed mean equals the median. The median of the original data is 287.5, which is what the trimmed mean approaches.

Question 13: (Optional) Robust Weight Analysis in Packaging Industry

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.

Solution

Step 1: Definition of M-Estimates

An M-estimate is a robust statistic obtained by minimizing a loss function over a dataset. Given n observations $x_1, x_2, \ldots, x_n$, the M-estimate of central tendency y^* is defined as:

$$y^* = \arg\min_y \sum_{i=1}^{n} \psi(x_i, y)$$

where $\psi(x_i, y)$ represents a loss function that determines the influence of each observation on y^*.

Common Choices for $\psi(x_i, y)$ For the Median:

$$\psi(x_i, y) = |x_i - y|$$

- This corresponds to minimizing the sum of absolute deviations.
- The solution is the median of the dataset.

For the Mean:

$$\psi(x_i, y) = (x_i - y)^2$$

- This corresponds to minimizing the sum of squared errors.
- The solution is the arithmetic mean.

Step 2: Compute M-Estimates for the Mean and Median

Given Dataset

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

Compute the Arithmetic Mean (Minimizing $(x_i - y)^2$)

$$\text{Mean} = \frac{\sum x_i}{n} = \frac{95 + 100 + 105 + 110 + 115 + 95 + 100 + 105 + 110 + 115 + 300 + 305 + 310 + 315 + 320}{15}$$

$$= 179.67$$

Compute the Median (Minimizing $|x_i - y|$) Sort the dataset:

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

The middle value (8th value) is:

$$\text{Median} = 110$$

Conclusion

- The mean (179.67) is highly influenced by the extreme values (300–320).
- The median (110) remains stable, demonstrating robustness to outliers.

Step 3: Effect of a Robust Weighting Function

A generalized robust loss function is:

$$\psi(x_i, y) = |x_i - y|^p$$

where $p \geq 1$ controls the sensitivity to large deviations.

When $p = 1$ (Absolute Loss, Median Minimization)

- Weights all errors equally, treating large and small deviations the same.
- The solution is the median.

When $p = 2$ (Squared Loss, Mean Minimization)

- Amplifies large deviations, giving more weight to outliers.
- The solution is the arithmetic mean, which is highly sensitive to outliers.

For $p > 2$ (Higher-Order Loss Functions)

- Further amplifies the impact of extreme values.
- Leads to even greater sensitivity to outliers, making the estimate less robust.

Conclusion

- Choosing $p = 1$ (absolute error) is more robust.
- Larger p makes estimates more sensitive to outliers.

Step 4: Impact of Weight Adjustments on Robustness

Large Outliers (300–320) Strongly Influence the Mean

- The arithmetic mean shifts upward due to these high values.
- Using $p = 2$ (squared loss) further exaggerates this effect.

Downweighting Extreme Values

- By choosing a function where large deviations have less impact (e.g., using a Huber loss function or a weighted M-estimate), we reduce their influence.
- For example, a capped loss function could ignore large errors after a threshold.

Robust M-Estimators Reduce the Effect of Large Deviations

- The median is robust because it completely ignores magnitude differences beyond ranking.
- Trimmed means and Huber estimators partially reduce the impact of large values.

Question 14: (Optional) Stock Market Volatility Analysis Using Kolmogorov-Smirnov 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.