Statistical Inference

Instructor: Mohammadreza A. Dehaqani

Muhammad Valinezhad

Fall 2024

Homework 2

- This homework includes 13 questions; the first 7 are available now so you can get started early.
- Part 9 of Q7 is a bonus question, offering extra credit equal to two quizzes. **have until Friday night to submit the bonus section.**
- 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.
- Please note that for computing questions, a major part of your grade is based on analyzing your results, so be sure to include explanations along with your code.
- 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 learnings to extend beyond the classroom teachings where necessary.
- As we mentioned in class, you'll have a quick (5 minute) in-person (or virtual!) hand-in session to help us check your understanding of the work you've submitted. For each assignment, about 25 students will be randomly chosen by an algorithm designed to ensure fairness. This algorithm will make sure you only present around 2 times during the term to keep things stress-free. However, if we notice inconsistencies between your work and what you present, the algorithm will adjust, increasing the chances you'll be selected again. Think of it as a dynamic process that adapts based on your performance—ensuring everyone gets a fair shot!

Q1: Oil Pipeline Pressure Monitoring

An engineer is monitoring the pressure inside an oil pipeline. Due to varying flow rates and environmental conditions, the pressure in the pipeline fluctuates slightly with time. The true average pressure of the pipeline is unknown. Pressure measurements, $X_1, X_2, \ldots, X_n$, satisfy the following model:

$$X_i = \mu + \epsilon_i$$

where μ is the unknown true average pressure, and ϵ_i represents random error. The errors are i.i.d. with mean 0 and unknown standard deviation σ.

The pipeline's pressure is measured 100 times. The recorded mean pressure is 75,348 Pascals, with a standard deviation of 25 Pascals.

(a) Construct an approximate 95% confidence interval for μ.

(b) The interval in part (a) was constructed for one of the following purposes. Indicate which is correct and explain why:

i) To estimate the average of the 100 pressure measurements and give ourselves some room for error in the estimate.

ii) To estimate the true average pressure of the pipeline and give ourselves some room for error in the estimate.

iii) To provide a range in which 95 of the 100 pressure measurements are likely to have fallen.

iv) To provide a range in which 95% of all possible pressure measurements are likely to fall.

Which of (i)-(iv) are false? Explain why they are false.

(c) Sketch the histogram of the 100 pressure measurements, including the mean and SD, or explain why this is not possible.

(d) If the engineer wants to ensure that the average pressure is within 1 Pascal of the true pressure, how many pressure measurements should be recorded for 95% confidence?

Q2: Manufacturing Quality Control

A quality control engineer is studying the strength of a batch of 625 industrial springs. The strength of the springs follows a normal distribution, and the engineer wants to monitor the proportion of springs that exceed a certain strength, which would make them too rigid and unusable. Based on a sample of 625 springs, the engineer constructs a 99% confidence interval for the mean strength of the springs, which ranges from 126.45 N (Newtons) to 128.55 N.

Springs with a strength above 140 N are considered defective.

(a) Construct an approximate 90% confidence interval for the percentage of springs in the batch that are defective (i.e., have a strength greater than 140 N).

(b) Explain whether the confidence interval for the percentage of defective springs can be computed based on the given information.

Q3: Ancient War between Persians and Greeks

Recall that the Law of Large Numbers (LLN) holds if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}S_n - \mathbb{E}\left(\frac{1}{n}S_n\right)\right| > \epsilon\right) = 0,$$

where $S_n = X_1 + X_2 + \cdots + X_n$, and the X_i's are i.i.d. random variables.

Imagine an ancient war between the Persians and the Greeks. The Persian army launches attacks on Greek fortresses. There are several strategic routes to each fortress, and each route has a probability p of being blocked by Greek defenses, meaning that no Persian soldiers can reach the fortress through that route. The routes fail independently. If a route is blocked, all soldiers sent along that route are lost. The Persian army does not know which routes will be blocked ahead of time.

For each of the following battle strategies, determine whether the Law of Large Numbers holds when S_n is defined as the total number of soldiers successfully reaching the fortresses out of n soldiers sent. Answer YES if the Law of Large Numbers holds, or NO if not, and give a brief justification of your answer. (Whenever convenient, you can assume that n is even.)

(a) Each soldier is sent through a completely different route to the fortress.

(b) The soldiers are split into $n/2$ pairs. Each pair is sent through its own route (i.e., different pairs are sent through different routes).

(c) The soldiers are split into two groups of $n/2$. All the soldiers in each group are sent through the same route, and the two groups are sent through different routes.

(d) All the soldiers are sent through one route.

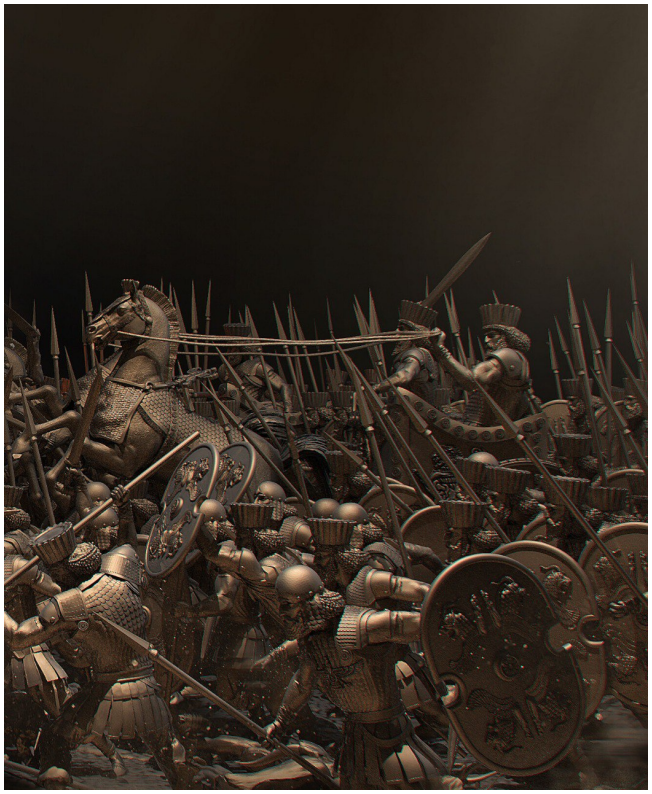

Figure 1: War

Q4: Estimating π by Throwing Darts

Imagine a square dartboard with a circle inscribed inside it, as shown in the figure below. Every dart you throw always lands somewhere within the square. The probability that the dart lands inside the circle is proportional to the area ratio of the circle to the square, which is $\frac{\pi}{4}$.

Now, let X_i be a random variable that takes the value 1 if the i-th dart lands within the circle, and 0 if it lands outside the circle. Using this setup, we can estimate π. Specifically, how many dart throws are required to ensure that the estimation error is no more than 0.01, with a probability of at least 95%? (You do not need to calculate the exact number of throws but provide the numerical expression.)

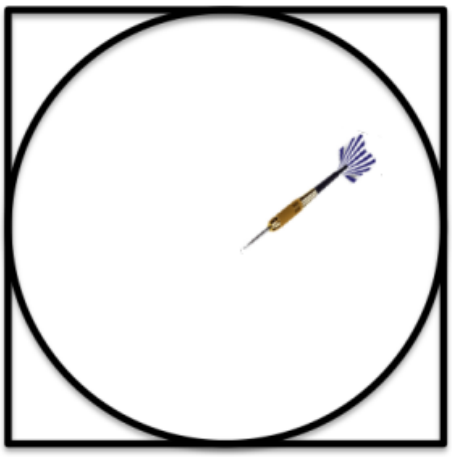

Figure 2: Dartboard with an inscribed circle.

Q5: Parameter Estimation for an Exponential Distribution

Consider a random sample $X_1, X_2, \ldots, X_n$ of size n drawn from a distribution with the following probability density function (PDF):

$$f(x;\alpha) = \frac{x}{\alpha^2} e^{-x/\alpha}, \quad x > 0, \quad \alpha > 0.$$

(a) Derive the maximum likelihood estimator (MLE) for the parameter α. Using the following data set, compute the estimate for α:

$$x_1 = 0.25, \quad x_2 = 0.75, \quad x_3 = 1.50, \quad x_4 = 2.50, \quad x_5 = 2.00.$$

(b) Find the method of moments (MoM) estimator for α. Using the same data set provided above, calculate the estimate for α.

Q6: Roulette Simulation and Profit Analysis

Roulette is a popular casino game played with a wheel that has numbered slots colored red, black, or green. In American roulette, the wheel has 38 slots: 18 red slots, 18 black slots, and 2 green slots labeled "0" and "00". Players can place various types of bets, including betting on whether the outcome will be a red or black slot.

In this exercise, we focus on a simple bet: betting on black.

Figure 3: Roulette game

If you place a bet on black and the outcome is indeed black, you win and double your money. However, if the outcome is red or green, you lose the amount you bet. For example, if you bet 1 dollar on black and win, you gain 1 dollar. If you lose, you forfeit your 1-dollar bet.

Because of the two green slots, the probability of landing on black (or red) is slightly less than $\frac{1}{2}$, specifically $\frac{18}{38} = \frac{9}{19}$.

Consider the following tasks to simulate this game and analyze the expected outcomes of betting on black:

1. Write a function that simulates this game for N rounds, where each round consists of betting 1 dollar on black. The function should return your total earnings S_N after N rounds.

2. Use Monte Carlo simulation to study the distribution of total earnings S_N for $N = 10, 25, 100, 1000$. For each N, simulate 100,000 rounds and plot the distribution of total earnings. Analyze whether the distributions appear similar to a normal distribution and observe how the expected values and standard errors change with N.

3. Repeat the previous simulation but for the average winnings $\frac{S_N}{N}$ instead of S_N. For each N, plot the distribution of average winnings and examine the changes in expected values and standard errors with different values of N.($N = 10, 25, 100, 1000$)

4. Calculate the theoretical expected values and standard errors of S_N for each N, and compare these theoretical values with your Monte Carlo simulation results. Report any differences between the theoretical and simulated values for each N.

5. Use the Central Limit Theorem (CLT) to approximate the probability that the casino loses money when you play $N = 25$ rounds, and verify this approximation using a Monte Carlo simulation.

6. Plot the probability that the casino loses money as a function of N for values N ranging from 25 to 1000. Discuss why casinos might encourage players to continue betting in light of these results.

Q7: Predicting the Outcome of the 2016 USA Presidential Election

In 2012, data scientists, including Nate Silver, accurately predicted the U.S. presidential election outcomes by aggregating data from multiple polls. By combining poll results, they provided more precise estimates than a single poll could achieve.

In this exercise, we aim to predict the result of the 2016 U.S. presidential election by analyzing polling data and aggregating results

Figure 4: Election

The data for this exercise is in a CSV file named `2016-general-election-trump-vs-clinton.csv`. Note that some rows may represent subgroups (e.g., voters affiliated with specific parties) and contain `NaN` values in the "Number of Observations" column. Exclude such rows from your calculations to avoid errors.

Question 1: Let X_i be a random variable where:

- $X_i = 1$ if the i-th voter supports the Democratic candidate.
- $X_i = 0$ if the i-th voter supports the Republican candidate.

With $i = 1, \ldots, N$, the Central Limit Theorem (CLT) states that if N is large:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \hat{p} \approx N\left(p, \frac{\hat{p}(1-\hat{p})}{N}\right)$$

where p is the true proportion of voters supporting the Democratic candidate. Based on the CLT result, derive and compute the 95% confidence interval (CI) for p.

Question 2: Suppose the true population proportion $p = 0.47$. Perform a Monte Carlo simulation with $N = 30$ and 10^5 iterations to show that the CI derived in Question 1 captures the true proportion p approximately 95% of the time.

Question 3: Load the data from `2016-general-election-trump-vs-clinton.csv` into your coding workspace and, using the `dplyr` library, create a tidy data frame that includes only the columns `Trump`, `Clinton`, `Pollster`, `Start Date`, `Number of Observations`, and `Mode`. Exclude any rows where `Number of Observations` is missing.

Question 4: Create a time-series plot of poll results showing support percentages for Trump and Clinton, using different colors for each candidate. Include a smooth trend line to visualize support trends over time.

Question 5: Calculate the total number of voters observed by summing all poll observations in the dataset.

Question 6: Calculate the estimated proportion of voters favoring Trump and Clinton. Display these estimates in a table.

Question 7: Using the aggregated data, compute the 95% confidence intervals for Trump and Clinton support proportions.

Question 8 (Optional): For illustrative purposes, assume there are only two parties, and let p denote the proportion of voters supporting Clinton. Consequently, $1 - p$ represents the proportion supporting Trump. We define the **spread** as the difference in support between Clinton and Trump:

$$d = p - (1 - p) = 2p - 1$$

Using the aggregated poll data, we estimate p as $\hat{p}$. Therefore, the estimated spread d can be approximated as:

$$d \approx 2\hat{p} - 1$$

This also implies that the standard error for the spread is twice as large as the standard error for $\hat{p}$. So, our confidence interval for the spread d is:

$$\text{CI for } d = (2\hat{p} - 1) \pm 1.96 \times (2 \times \text{SE}_{\hat{p}})$$

where $\text{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$ is the standard error of $\hat{p}$.

a) Calculate the 95% confidence interval for the spread d, using the formula provided above.

b) Conduct a hypothesis test to determine if the spread d is significantly different from zero by testing $H_0 : d = 0$ vs. $H_a : d \neq 0$. Provide the test statistic and p-value.

Question 9 (Bonus): Now, let's fast-forward to right now, the 2024 presidential election! Find a similar dataset (it doesn't need identical labels to 2016) and put your skills to the test by working through all 8 questions again. Use your best judgment to fill in any gaps—you're the data scientist here, so don't be afraid to improvise!