Statistical Inference

Instructor: Mohammadreza A. Dehaqani

Muhammad Valinezhad, Mahdi Ebrahimi

Fall 2024

Homework 2

- Part 9 of Q7 is a bonus question, offering extra credit equal to two quizzes. **You 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?

SOLUTION

(a) The standard deviation for the sample mean is estimated by

$$\frac{25}{\sqrt{100}} = 2.5$$

Thus, the approximate 95% confidence interval for the true average pressure is:

$$75348 \pm 1.96 \times 2.5 = (75343, 75353)$$

(b) Option (ii) is true; the rest are false. The confidence interval is meant to estimate the true average pressure μ, not the average of the measurements. The variability in the measurements themselves is 25 Pascals, but the variability in the sample mean is only 2.5 Pascals, so options (iii) and (iv) are incorrect.

(c) We cannot sketch the histogram because we lack information about the shape of the error distribution. The model does not provide information on whether the errors follow a normal distribution or another type of distribution.

(d) To ensure the average pressure is within 1 Pascal of the true value, we use the margin of error formula:

$$1.96 \times \frac{25}{\sqrt{n}} = 1$$

Solving for n, we get:

$$\sqrt{n} = \frac{25 \times 1.96}{1} \quad \Rightarrow \quad \sqrt{n} = 49 \quad \Rightarrow \quad n \approx 2400$$

Therefore, approximately 2400 pressure measurements are needed.

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.

SOLUTION

(a) The midpoint of the 99% confidence interval is:

$$\frac{126.45 + 128.55}{2} = 127.5 \text{ N}$$

We estimate the standard deviation of the sample using the formula for the margin of error in a confidence interval:

$$128.55 - 127.5 = 1.05 = 2.57 \times \frac{\sigma}{\sqrt{625}}$$

Solving for σ, we get:

$$\sigma = \frac{1.05 \times \sqrt{625}}{2.57} \approx 10.21 \text{ N}$$

Now, we compute the Z-score for 140 N:

$$Z = \frac{140 - 127.5}{10.21} \approx 1.22$$

Using standard normal tables, the probability of a Z-score greater than 1.22 is approximately $P(Z \geq 1.22) = 0.1112$. Therefore, we estimate that 11.12

The bootstrap standard error for this percentage is:

$$\text{SE} = \sqrt{\frac{0.1112(1 - 0.1112)}{625}} \times 100\% \approx 1.26\%$$

The 90% confidence interval for the percentage of defective springs is then:

$$11.12\% \pm 1.645 \times 1.26\% = (9.05\%, 13.19\%)$$

(b) Yes, it is possible to construct the confidence interval because we have a normal distribution for the strength measurements and enough data to calculate both the standard deviation and the percentage of springs that are defective.

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.

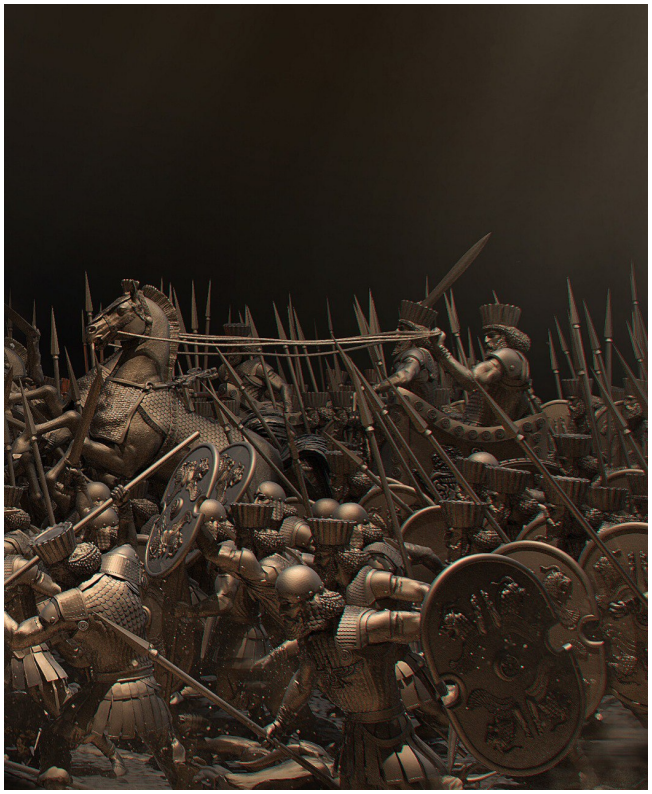


Figure 1: War

SOLUTION

(a) YES: Define X_i to be 1 if a soldier successfully reaches the fortress through route i, and 0 otherwise. Then $X_i = 1$ with probability $1-p$, and $X_i = 0$ with probability p. Since each soldier is sent through a different route, the total number of successful soldiers, $S_n = X_1 + X_2 + \cdots + X_n$, is the sum of n i.i.d. Bernoulli random variables, $S_n \sim \text{Binomial}(n, 1-p)$.

Now, we define $A_n = \frac{S_n}{n}$ to be the fraction of successful soldiers out of the n soldiers sent. For each X_i, we know:

$$\mathbb{E}[X_i] = 1 - p$$

and

$$\text{Var}(X_i) = p(1-p).$$

Using Chebyshev's inequality, we calculate:

$$\mathbb{P}\left(|A_n - \mathbb{E}[A_n]| \geq \epsilon\right) = \mathbb{P}\left(|A_n - (1-p)| \geq \epsilon\right) \leq \frac{\text{Var}(A_n)}{\epsilon^2}.$$

Since $A_n = \frac{S_n}{n}$ and S_n is a sum of n i.i.d. Bernoulli random variables, we have:

$$\text{Var}(A_n) = \frac{\text{Var}(S_n)}{n^2} = \frac{n \cdot p(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Therefore, Chebyshev's inequality gives us:

$$\mathbb{P}\left(|A_n - (1-p)| \geq \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2}.$$

As $n \to \infty$, the right-hand side of the inequality tends to 0, meaning that the fraction of successful soldiers, A_n, converges in probability to $1-p$ as $n \to \infty$. This confirms that the Law of Large Numbers holds in this case.

(b) YES: The Persian soldiers are sent in pairs. Now, define X_i for each pair $i = 1, \ldots, \frac{n}{2}$, where $X_i = 0$ with probability p and 2 (soldiers) otherwise. The total number of soldiers successfully reaching the fortress is $S_n = X_1 + X_2 + \cdots + X_{\frac{n}{2}}$, and the fraction of received soldiers is $A_n = \frac{S_n}{n}$.

Now for each $i = 1, \ldots, \frac{n}{2}$:

$$\mathbb{E}[X_i] = 2(1-p)$$

and

$$\mathrm{Var}(X_i) = 4p(1-p).$$

Thus:

$$\mathbb{E}[A_n] = \frac{\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_{\frac{n}{2}}]}{n} = \frac{1}{n} \cdot \frac{n}{2} \cdot 2(1-p) = 1-p$$

and

$$\mathrm{Var}(A_n) = \frac{1}{n^2}(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_{\frac{n}{2}})) = \frac{1}{n^2} \cdot \frac{n}{2} \cdot 4p(1-p) = \frac{2p(1-p)}{n}.$$

Finally, we get:

$$\mathbb{P}\left(|A_n - \mathbb{E}[A_n]| \geq \epsilon\right) \leq \frac{2p(1-p)}{n\epsilon^2} \to 0 \quad \text{as} \quad n \to \infty.$$

Therefore, as $n \to \infty$, the Law of Large Numbers holds for this case.

(c) NO: The Persian soldiers are split into two large groups of $\frac{n}{2}$. All the soldiers in each group are sent through a single route. If either route is blocked, all soldiers in that group are lost. In this case, we define:

$$X_i = \begin{cases} 0, & \text{with probability } p \\ \frac{n}{2}, & \text{with probability } 1-p \end{cases}$$

for $i = 1, 2$. Now, $S_n = X_1 + X_2$ and $A_n = \frac{S_n}{n} = \frac{X_1 + X_2}{n}$.

We have:

$$\mathbb{E}[X_i] = \frac{n}{2}(1-p)$$

and

$$\mathrm{Var}(X_i) = \frac{n^2}{4}p(1-p).$$

Thus:

$$\mathbb{E}[A_n] = \frac{\mathbb{E}[X_1] + \mathbb{E}[X_2]}{n} = \frac{1}{n}\left(\frac{n}{2}(1-p) + \frac{n}{2}(1-p)\right) = 1-p$$

and

$$\mathrm{Var}(A_n) = \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2)\right) = \frac{1}{n^2} \cdot 2 \cdot \frac{n^2}{4}p(1-p) = \frac{p(1-p)}{2}.$$

Finally, we get:

$$\mathbb{P}\left(|A_n - \mathbb{E}[A_n]| \geq \epsilon\right) \leq \frac{p(1-p)}{2\epsilon^2},$$

which does not tend to 0 as $n \to \infty$. Therefore, the Law of Large Numbers does not hold in this case.

(d) NO: All the Persian soldiers are sent through a single route. In this case, $S_n = X_1$, where $X_1 = n$ with probability $1-p$ and $X_1 = 0$ with probability p. The fraction of successful soldiers is $A_n = \frac{X_1}{n}$.

Thus:

$$\mathbb{E}[A_n] = \frac{\mathbb{E}[X_1]}{n} = \frac{n(1-p)}{n} = 1-p$$

and

$$\mathrm{Var}(A_n) = \frac{1}{n^2}\mathrm{Var}(X_1) = \frac{1}{n^2}n^2p(1-p) = p(1-p).$$

The inequality results in:

$$\mathbb{P}\left(|A_n - \mathbb{E}[A_n]| \geq \epsilon\right) \leq \frac{p(1-p)}{\epsilon^2}.$$

As before, this does not converge to 0 as $n \to \infty$, and thus, the Law of Large Numbers does not hold.

For problems (c) and (d), you should've had the intuition that since all the soldiers are sent through 1 or 2 routes, increasing the number of soldiers (n) does not really help the Law of Large Numbers to hold.

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.)

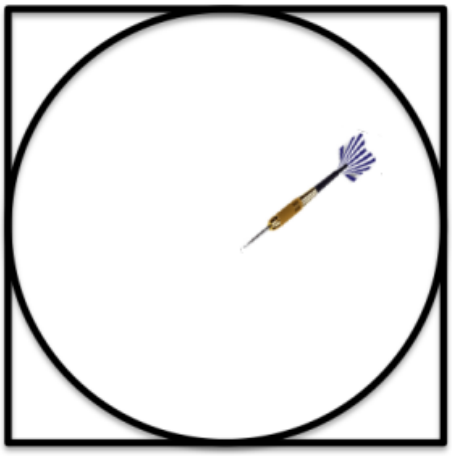


Figure 2: Dartboard with an inscribed circle.

SOLUTION

Suppose we throw the dart n times in order to estimate π. We can estimate π by $M_n = \frac{4}{n}\sum_{i=1}^{n} X_i$. Then we want $\mathbb{P}(|M_n - \pi| \geq 0.01) \leq 0.05$.

The random variable X_i is a Bernoulli random variable with $\mathbb{P}(X_i = 1) = \frac{\pi}{4}$. The expectation of M_n is:

$$\mathbb{E}[M_n] = \mathbb{E}\left(\frac{4}{n}\sum_{i=1}^{n} X_i\right) = \frac{4}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{4}{n}\cdot n \cdot \frac{\pi}{4} = \pi.$$

• By Chebyshev's inequality:

The variance of M_n is:

$$\mathrm{Var}(M_n) = \mathrm{Var}\left(\frac{4}{n}\sum_{i=1}^{n} X_i\right) = \frac{16}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{16}{n^2}\cdot n \cdot \mathrm{Var}(X_i) = \frac{16}{n^2}\cdot n \cdot \frac{\pi}{4}\left(1 - \frac{\pi}{4}\right) = \frac{\pi(4-\pi)}{n}.$$

Thus:

$$\mathbb{P}(|M_n - \pi| \geq 0.01) \leq \frac{\mathrm{Var}(M_n)}{0.01^2} = \frac{\pi(4-\pi)}{n \cdot 0.01^2}.$$

Let $\frac{\pi(4-\pi)}{n\cdot 0.01^2} \leq 0.05$, then we get:

$$n \geq \frac{\pi(4-\pi)}{0.01^2 \cdot 0.05} = 539353.2.$$

Thus, we have to throw at least 539,354 times.

One may assume the exact value of π is unknown. Then we can use the inequality $\pi(4-\pi) = -(\pi-2)^2+4 \leq 4$. Let $\frac{\pi(4-\pi)}{n\cdot 0.01^2} \leq \frac{4}{n\cdot 0.01^2} \leq 0.05$, and we can get $n \geq \frac{4}{0.01^2\cdot 0.05} = 800000$.

Q5: Parameter Estimation for an Exp (Optional)

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.

Solution

(a) Maximum Likelihood Estimation (MLE): We first obtain the likelihood by multiplying the probability density function for each X_i. We then simplify this expression:

$$L(\alpha) = \prod_{i=1}^{n} f(x_i;\alpha) = \prod_{i=1}^{n} \alpha^{-2} x_i e^{-x_i/\alpha} = \alpha^{-2n} \left(\prod_{i=1}^{n} x_i \right) \exp\left(-\frac{\sum_{i=1}^{n} x_i}{\alpha} \right).$$

Instead of directly maximizing the likelihood, we instead maximize the log-likelihood:

$$\log L(\alpha) = -2n \log \alpha + \sum_{i=1}^{n} \log x_i - \frac{\sum_{i=1}^{n} x_i}{\alpha}.$$

To maximize this function, we take the derivative with respect to α:

$$\frac{d}{d\alpha} \log L(\alpha) = -\frac{2n}{\alpha} + \frac{\sum_{i=1}^{n} x_i}{\alpha^2}.$$

We set this derivative equal to zero, then solve for α:

$$-\frac{2n}{\alpha} + \frac{\sum_{i=1}^{n} x_i}{\alpha^2} = 0.$$

Solving gives our estimator, which we denote with a hat:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} x_i}{2n} = \frac{\bar{x}}{2}.$$

Using the given data, we obtain an estimate:

$$\hat{\alpha} = \frac{0.25 + 0.75 + 1.50 + 2.50 + 2.0}{2 \cdot 5} = \frac{7}{10} = 0.70.$$

(b) Method of Moments Estimation (MoM) We first obtain the first population moment. Notice the integration is done by identifying the form of the integral as that of the second moment of an exponential distribution.

$$\mathbb{E}[X] = \int_0^\infty x \cdot \alpha^{-2} x e^{-x/\alpha}\, dx = \frac{1}{\alpha} \int_0^\infty \frac{x^2}{\alpha} e^{-x/\alpha}\, dx = \frac{1}{\alpha}(2\alpha^2) = 2\alpha.$$

We then set the first population moment, which is a function of α, equal to the first sample moment:

$$2\alpha = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Solving for α, we obtain the method of moments estimator:

$$\hat{\alpha}_{\text{MoM}} = \frac{\sum_{i=1}^{n} x_i}{2n} = \frac{\bar{x}}{2}.$$

Using the given data, we obtain an estimate:

$$\hat{\alpha}_{\text{MoM}} = \frac{0.25 + 0.75 + 1.50 + 2.50 + 2.0}{2 \cdot 5} = 0.70.$$

Note that, in this case, the MLE and MoM estimators are the same.

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.

Solution

Task 1: Simulating Total Earnings To simulate this game, we define a function that takes the number of rounds N as input. Each round consists of betting 1 dollar on black. The function returns the total earnings S_N after N rounds. If the outcome is black, we win 1 dollar; if it is red or green, we lose 1 dollar.

Task 2: Monte Carlo Simulation for Total Earnings Distribution We use a Monte Carlo simulation to study the distribution of total earnings S_N for different values of N: 10, 25, 100, and 1000. Each simulation is repeated 100,000 times for each N, and we calculate the expected value and standard error. Additionally, we plot the frequency distribution of total earnings and Q-Q plots to observe the normality of each distribution.(Figure 4)

The following table summarizes the expected value and standard error for the total earnings S_N across different values of N.

N	Expected Value	Standard Error
10	-0.53592	3.145993
25	-1.33416	4.991162
100	-5.22196	9.991617
1000	-52.67244	31.60298

Table 1: Expected Value and Standard Error for Total Earnings S_N at Different Values of N

Task 3: Monte Carlo Simulation for Average Winnings Distribution In this task, we repeat the simulation for the average winnings $\frac{S_N}{N}$. For each N, we simulate 100,000 rounds and plot the frequency distribution of average winnings as well as Q-Q plots to assess normality.(Figure 5)

The following table summarizes the expected value and standard error for the average winnings $\frac{S_N}{N}$ across different values of N.

N	Expected Value	Standard Error
10	-0.05146	0.3141412
25	-0.0523152	0.1987781
100	-0.0529774	0.09981486
1000	-0.0526206	0.03166803

Table 2: Expected Value and Standard Error for Average Winnings $\frac{S_N}{N}$ at Different Values of N

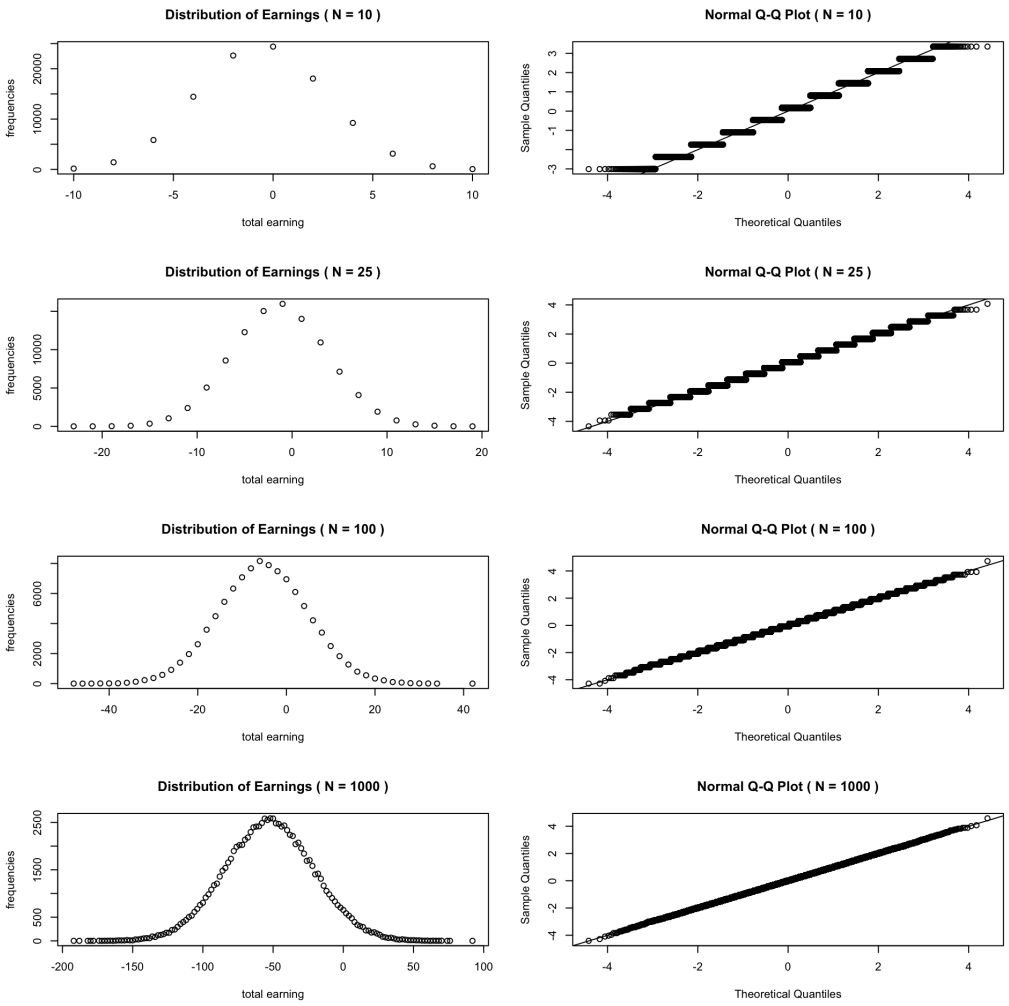


Figure 4: The frequency distribution of total earnings and Q-Q plots

Task 4: Comparing Theoretical and Simulated Values The expected earning per bet can be calculated based on the probabilities of winning or losing in a bet on black. In roulette, there are 18 black slots, 18 red slots, and 2 green slots. Thus, the probability p of landing on black is:

$$p = \frac{18}{38} = \frac{9}{19}$$

The probability of losing the bet (landing on red or green) is therefore:

$$1 - p = \frac{10}{19}$$

The expected value (or mean) μ of a single bet is calculated as the weighted average of the possible outcomes:

$$\mu = (1 \times p) + (-1 \times (1 - p))$$

Substituting the probabilities:

$$\mu = 1 \times \frac{9}{19} + (-1) \times \frac{10}{19} = -\frac{1}{19}$$

This negative expected value indicates that, on average, the player loses a small amount on each bet due to the house edge created by the green slots.

If we place N bets, the expected total earnings $E[S_N]$ is simply N times the expected earnings per bet:

$$E[S_N] = N \times \mu = N \times \left(-\frac{1}{19}\right) = -\frac{N}{19}$$

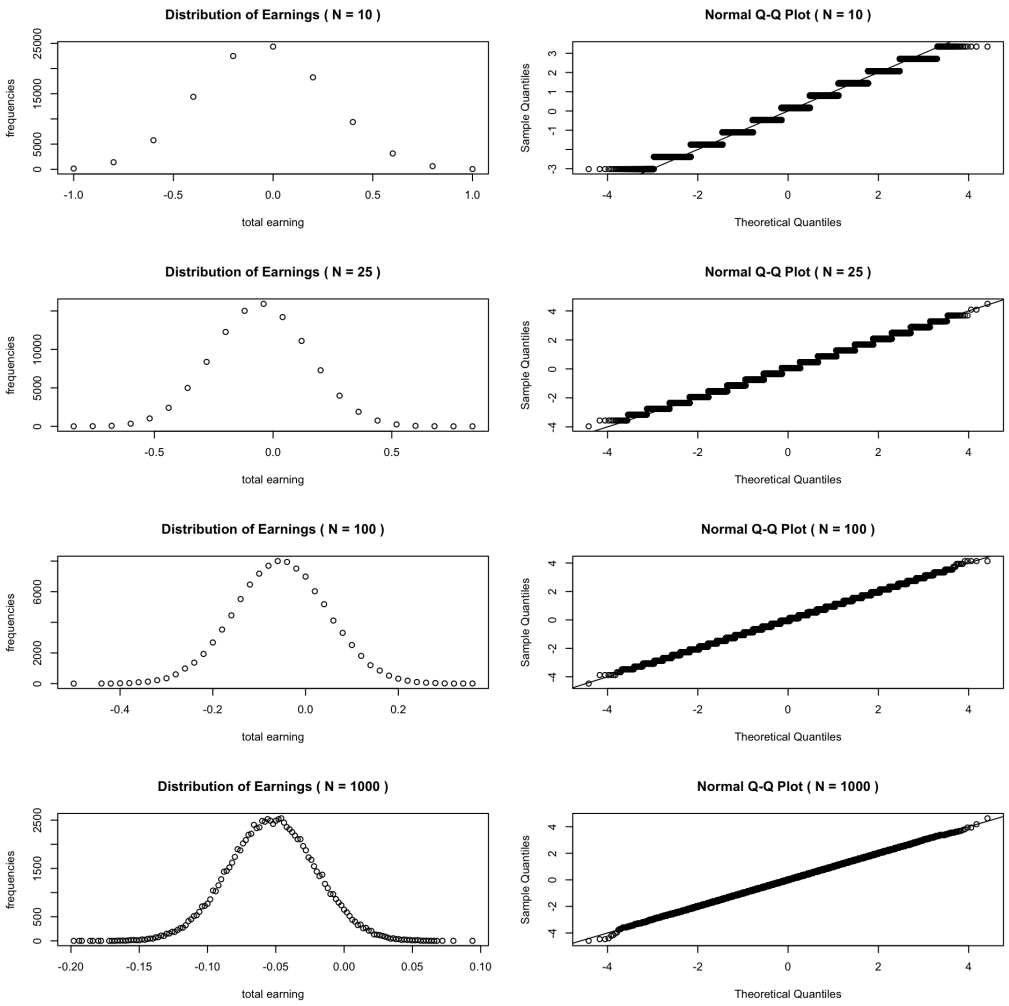


Figure 5: The frequency distribution of average winnings as well as Q-Q plots

Thus, the theoretical expected earnings become increasingly negative as N increases, which aligns with the house advantage in the game.

The standard error σ for the earnings per bet can be calculated using the variance formula. Since the outcomes are $+1$ (win) or -1 (lose), the variance per bet is:

$$\sigma^2 = (1-(-1))^2 \times p \times (1-p) = 4 \times \frac{9}{19} \times \frac{10}{19} = \frac{90}{361}$$

Therefore, the standard deviation (or standard error) per bet is:

$$\sigma = 2 \times \sqrt{\frac{90}{361}}$$

For N bets, the total standard error σ_{S_N} scales by $\sqrt{N}$:

$$\sigma_{S_N} = \sqrt{N} \times \sigma = \sqrt{N} \times 2 \times \sqrt{\frac{90}{361}}$$

Using these theoretical values, we compare them with the Monte Carlo simulation results.

As shown in the table above, we can calculate the theoretical values of earnings when betting 10 times or more. When we compare our Monte Carlo simulation results with these theoretical values, we see that the differences between them are no more than 0.06. This indicates that our simulation results are very close to the theoretical values, validating our calculations for the expected values of earnings when betting 10, 25, 100, or 1000 times.

N	Theoretical Expected Value	Simulated Expected Value
10	-0.5263158	-0.53592
25	-1.3157895	-1.33416
100	-5.2631579	-5.22196
1000	-52.6315789	-52.67244

Table 3: Comparison of Theoretical and Simulated Expected Values for Total Earnings S_N at Different Values of N

Task 5: Probability that the Casino Loses Money Using the Central Limit Theorem (CLT), we approximate the probability that the casino loses money for $N = 25$ rounds. We validate this approximation with a Monte Carlo simulation.

Method	Probability that Casino Loses Money
CLT Approximation	0.3960737
Monte Carlo Simulation	0.39489

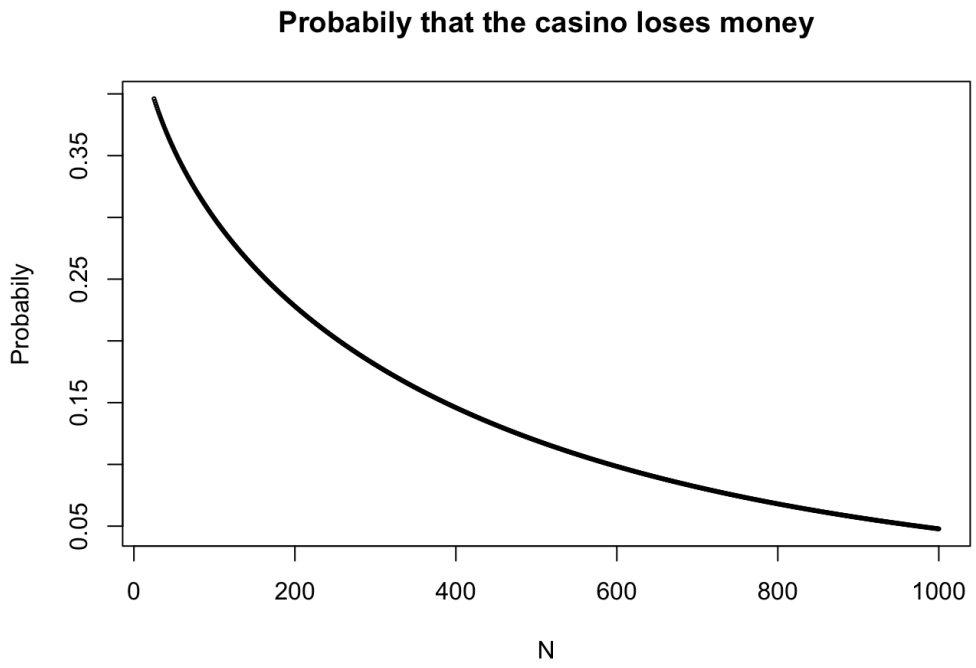
Table 4: Probability that the Casino Loses Money for $N = 25$ Using CLT Approximation and Monte Carlo Simulation

To use the Central Limit Theorem (CLT) for approximation, we calculated the expectation μ and standard deviation σ of the earnings for each bet. For $N = 25$, we then used $25 \times \mu$ and $\sqrt{25} \times \sigma$ as inputs to the normal cumulative distribution function (pnorm) to approximate the probability.

As shown in the table above, the CLT approximation gave a probability of 0.3960737 that the casino would lose money, while the Monte Carlo simulation yielded a probability of 0.39489. The results are very close, with a difference of less than 0.002.

This small discrepancy confirms that the Central Limit Theorem provides an accurate approximation for the probability of the casino losing money in this context, especially as N grows larger. The approximation effectively captures the expected behavior of total earnings, as the distribution of S_N becomes increasingly normal with higher values of N, validating the use of CLT for large sample sizes.

Task 6: Probability of Casino Losing Money with Varying N We plot the probability that the casino loses money as N increases from 25 to 1000, showing how the probability decreases with increasing rounds.



Using the approximation of Central Limit Theorem, we can see that the probability that the casino loses money decreases as N (the number of times a player plays the game) increases. This is why the casino hopes

the player can keep playing at the same desk as long as possible.

Q7: Predicting the Outcome of the 2016 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 6: 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!

SOLUTION

Question 1: The Central Limit Theorem tells us that the distribution of any random variable X is approximately normal with its mean μ (the population mean) and its standard deviation $\frac{\sigma}{\sqrt{n}}$ where σ is the population standard deviation and n is the population size. As a result, the random interval

$$\mu - Z_{0.975}\frac{\sigma}{\sqrt{n}} \text{ to } \mu + Z_{0.975}\frac{\sigma}{\sqrt{n}}$$

(where $Z_{0.975}$ is approximately 1.96) has a 95% probability of falling on the true value of X.

Since we know the distribution of p (the proportion of democrats in the population), so the 95% confidence interval of p can be formulated as below:

$$\overline{X} - Z_{0.975}\frac{S_X}{\sqrt{N}} \leq p \leq \overline{X} + Z_{0.975}\frac{S_X}{\sqrt{N}}$$

where N is the sample size, $\overline{X}$ is actually $\hat{p}$, and S_X is the sample standard deviation $\sqrt{\hat{p}(1-\hat{p})}$.

Question 2:
Output:

```
0.93118
```

The simulation shows that the confidence interval captures the true proportion p approximately 93.2% of the time, which is slightly less than 95%. This discrepancy is due to the small sample size ($N = 30$), where the normal approximation is less accurate. we should increase the value of N (e.g., N = 100), then the probability would be close to 95%.

Question 3:

Listing 1: Creating a tidy data frame

```
library(readr)
# Load dplyr for data manipulation
library(dplyr)

# Load the data
df <- read_csv("2016-general-election-trump-vs-clinton.csv")

# Select the desired columns and rename them for consistency
df <- df %>%
  select(Trump, Clinton, Pollster, `Start Date`, `Number of Observations`, Mode) %>%
  rename(
    start_date = `Start Date`,
    observations = `Number of Observations`,
    pollster = Pollster,
    mode = Mode
  )

# Convert start_date to Date type
df$start_date <- as.Date(df$start_date, format = "%m/%d/%Y")

# Ensure that percentage columns are numeric
df$Trump <- as.numeric(df$Trump)
df$Clinton <- as.numeric(df$Clinton)

# Remove rows with missing values in key columns
df <- df %>%
  filter(!is.na(Trump), !is.na(Clinton), !is.na(observations))

# View the cleaned data
head(df)
%
```

Question 4:

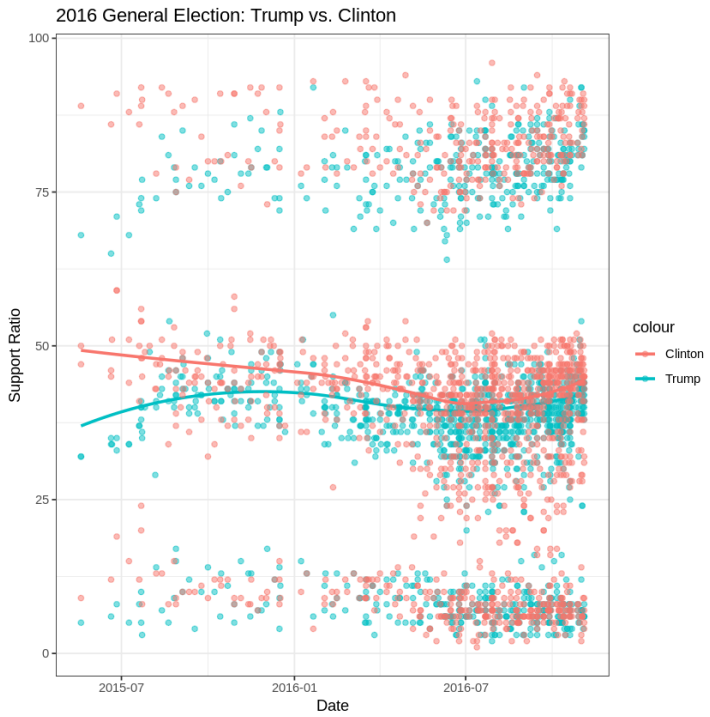


Figure 7: Support percentages over time for Trump and Clinton

Question 5:
Output:

```
The number of votes we have observed is 1940931
```

Question 6:
Output:

	estimated proportion
Trump	0.4058335
Clinton	0.4561208

```
Estimated proportion supporting Trump: 0.4058335
Estimated proportion supporting Clinton: 0.4561208
```

Question 7:
Output:

	95% confidence interval for the true proportion
Trump	0.405142456173305 ~ 0.406524116036914
Clinton	0.455420331181663 ~ 0.45682173834172

Question 8 (Optional):
a)

Output:

```
95% Confidence interval for the spread (d) is -0.0598537097142002 ~ -0.0568282557726678
```

b)
We are interested in testing the following hypotheses:

$$H_0 : d = 0 \quad \text{versus} \quad H_a : d \neq 0$$

To test this, we use a test statistic for d, which can be formulated as a function of p. Since $d = 1 - 2p$, the test statistic t can be written as:

$$t = \frac{1 - 2\hat{p}}{2 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}}$$

where $\hat{p}$ is the estimated proportion voting for Clinton, and N is the total sample size (number of observations).

Output:

The p-value based on the aggregrated data is 0 . That is, the evidence is strong to reject null hypothesis, and thus spread (d) is not zero. This indicates that the spread between Clinton and Trump is statistically significant.

Q8:Convergence of Estimators in a Noisy Measurement Process

A company is testing an experimental sensor that measures temperature in real time. Due to environmental factors, the sensor readings include random noise. Let the true temperature be $\theta = 25°\text{C}$, and let X_n be a sequence of estimators for θ at different times. Each estimator X_n is defined as:

$$X_n = \theta + \frac{Z_n}{\sqrt{n}},$$

where Z_n is a sequence of i.i.d. random variables with $\mathbb{E}[Z_n] = 0$ and $\text{Var}(Z_n) = 1$. In other words, $Z_n \sim N(0, 1)$.

We will analyze the convergence properties of X_n to θ in three different ways:

(a) **Mean-Square Convergence**
Determine if X_n converges to θ in the mean-square sense. Recall that X_n converges to θ in mean-square if:

$$\lim_{n\to\infty} \mathbb{E}[(X_n - \theta)^2] = 0.$$

Calculate $\mathbb{E}[(X_n - \theta)^2]$ and analyze if it approaches zero as $n \to \infty$.

(b) **Convergence in Probability**
Determine if X_n converges to θ in probability. Recall that X_n converges to θ in probability if for any $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}(|X_n - \theta| > \epsilon) = 0.$$

Analyze whether this probability approaches zero as $n \to \infty$.

(c) **Convergence in Distribution**
Determine if X_n converges to θ in distribution. To do this, find the distribution of $X_n - \theta$ as $n \to \infty$ and analyze if it matches the distribution of a constant. Consider the behavior of $\frac{Z_n}{\sqrt{n}}$ as $n \to \infty$, and use this to establish the convergence in distribution.

Solution

(a) Mean-Square Convergence To check if X_n converges to θ in the mean-square sense, we calculate:

$$E\left[(X_n - \theta)^2\right] = E\left[\left(\frac{Z_n}{\sqrt{n}}\right)^2\right].$$

Since $Z_n \sim N(0, 1)$, we know $E[Z_n] = 0$ and $\text{Var}(Z_n) = 1$. Thus:

$$E\left[\left(\frac{Z_n}{\sqrt{n}}\right)^2\right] = \frac{1}{n}E[Z_n^2] = \frac{1}{n} \cdot \text{Var}(Z_n) = \frac{1}{n}.$$

Taking the limit as $n \to \infty$:

$$\lim_{n\to\infty} E\left[(X_n - \theta)^2\right] = \lim_{n\to\infty} \frac{1}{n} = 0.$$

Conclusion: X_n converges to θ in the mean-square sense.

(b) Convergence in Probability To check if X_n converges to θ in probability, we analyze:

$$P(|X_n - \theta| > \epsilon) = P\left(\left|\frac{Z_n}{\sqrt{n}}\right| > \epsilon\right).$$

This can be rewritten as:

$$P\left(\left|\frac{Z_n}{\sqrt{n}}\right| > \epsilon\right) = P\left(|Z_n| > \epsilon\sqrt{n}\right).$$

Since $Z_n \sim N(0, 1)$, the tail probability $P(|Z_n| > \epsilon\sqrt{n})$ decreases as $n \to \infty$. Using the properties of the standard normal distribution:

$$\lim_{n\to\infty} P(|Z_n| > \epsilon\sqrt{n}) = 0, \quad \forall \epsilon > 0.$$

Conclusion: X_n converges to θ in probability.

(c) Convergence in Distribution To check if X_n converges to θ in distribution, we analyze the distribution of $X_n - \theta$:

$$X_n - \theta = \frac{Z_n}{\sqrt{n}}.$$

As $n \to \infty$, the variance of $\frac{Z_n}{\sqrt{n}}$ becomes:

$$\text{Var}\left(\frac{Z_n}{\sqrt{n}}\right) = \frac{1}{n}.$$

Thus, $\frac{Z_n}{\sqrt{n}} \to 0$ in distribution because the random variable converges to a constant.

Conclusion: X_n converges to θ in distribution.

- **Mean-Square Convergence**: X_n converges to θ.
- **Convergence in Probability**: X_n converges to θ.
- **Convergence in Distribution**: X_n converges to θ.

Q9: Failure Rates in a Model with Time-Dependent Scaling

In a highly sophisticated system, the failure times of machines are modeled by a modified Gamma-Weibull distribution. However, the system involves dynamic, time-dependent scaling of the failure rate, leading to the following probability density function (PDF) for failure times T:

$$f_T(t; \alpha, \beta(t)) = \frac{1}{\Gamma(\alpha)} \left(\frac{t}{\beta(t)}\right)^{\alpha-1} \exp\left(-\frac{t}{\beta(t)}\right),$$

where α is a constant shape parameter, and $\beta(t) = \beta_0 + \beta_1 t$, with β_0 and β_1 as unknown parameters governing the time-dependent scale. The Gamma function $\Gamma(\alpha)$ is defined as:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx.$$

(a) **Log-Likelihood Function**
For a sample of n independent observations $T_1, T_2, \ldots, T_n$, derive the log-likelihood function $\ell(\alpha, \beta_0, \beta_1)$. Use symbolic notation and include intermediate mathematical steps, such as the treatment of $\beta(t)$.

(b) **First-Order Conditions and System of Equations**
Compute the first-order conditions (score functions) by taking the partial derivatives of the log-likelihood function with respect to α, β_0, and β_1. Write the resulting system of equations for the MLE estimates $\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1$.

(c) **Fisher Information Matrix**
Derive the Fisher Information matrix $I(\alpha, \beta_0, \beta_1)$. The elements of this matrix will involve second-order derivatives of the log-likelihood function, and you will need to compute the expected values of complex integrals involving the time-varying function $\beta(t)$. You may use integration by parts and transformations to simplify these expressions.

(d) **Cramér-Rao Bound**
Compute the Cramér-Rao lower bounds for the variances of the unbiased estimators $\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1$. Specifically:

i) Calculate the inverse of the Fisher Information matrix $I(\alpha, \beta_0, \beta_1)^{-1}$ symbolically, using matrix inversion techniques.

ii) Provide an interpretation of the Cramér-Rao bounds in terms of the precision of the estimators and explain how these bounds relate to the practical estimation of failure rates in the system.

(e) **Parameter Interpretation and Time-Varying Failure Rate**
Given that $\beta(t) = \beta_0 + \beta_1 t$, discuss the implications of the time-varying failure rate. Investigate under which conditions the failure rate decreases or increases over time, based on the estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$. Use this analysis to make practical recommendations for optimizing maintenance schedules.

(f) **Numerical Solution**
Given the sample failure times $T_1 = 1.2$, $T_2 = 2.1$, $T_3 = 3.5$, $T_4 = 4.7$, and $T_5 = 5.9$ (in hours), numerically solve for $\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1$ using a nonlinear optimization method such as Newton-Raphson or gradient descent. Provide the full steps of your numerical approach and use software tools to confirm your solution.

Solution

The failure times of machines in a system are modeled by a modified Gamma-Weibull distribution with the probability density function (PDF):

$$f_T(t;\alpha,\beta(t)) = \frac{1}{\Gamma(\alpha)}\left(\frac{t}{\beta(t)}\right)^{\alpha-1}\exp\left(-\frac{t}{\beta(t)}\right),$$

where $\alpha > 0$ is a shape parameter, and $\beta(t) = \beta_0 + \beta_1 t$, with $\beta_0 > 0$ and β_1 as scaling parameters. The Gamma function is given by:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx.$$

For n independent failure times $T_1, T_2, \ldots, T_n$, we derive the log-likelihood function, compute first-order conditions, find the Fisher Information matrix, and perform numerical solutions for the parameters α, β_0, β_1.

The likelihood function for n independent observations $T_1, T_2, \ldots, T_n$ is:

$$L(\alpha,\beta_0,\beta_1) = \prod_{i=1}^n f_T(T_i;\alpha,\beta(T_i)).$$

Taking the natural logarithm, the log-likelihood function becomes:

$$\ell(\alpha,\beta_0,\beta_1) = \sum_{i=1}^n \ln f_T(T_i;\alpha,\beta(T_i)).$$

Substituting the PDF into the log-likelihood:

$$\ell(\alpha,\beta_0,\beta_1) = \sum_{i=1}^n \left[-\ln\Gamma(\alpha) + (\alpha-1)\ln\left(\frac{T_i}{\beta(T_i)}\right) - \frac{T_i}{\beta(T_i)}\right].$$

Expanding and simplifying:

$$\ell(\alpha,\beta_0,\beta_1) = -n\ln\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^n \ln T_i - (\alpha-1)\sum_{i=1}^n \ln(\beta_0+\beta_1 T_i) - \sum_{i=1}^n \frac{T_i}{\beta_0+\beta_1 T_i}.$$

The first-order conditions are derived by differentiating $\ell(\alpha,\beta_0,\beta_1)$ with respect to α, β_0, β_1, and setting the derivatives to zero.

For α:

$$\frac{\partial \ell}{\partial \alpha} = -n\psi(\alpha) + \sum_{i=1}^n \ln T_i - \sum_{i=1}^n \ln(\beta_0+\beta_1 T_i),$$

where $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ is the digamma function. Setting $\frac{\partial \ell}{\partial \alpha} = 0$:

$$-n\psi(\alpha) + \sum_{i=1}^n \ln T_i - \sum_{i=1}^n \ln(\beta_0+\beta_1 T_i) = 0.$$

For β_0:

$$\frac{\partial \ell}{\partial \beta_0} = -(\alpha-1)\sum_{i=1}^n \frac{1}{\beta_0+\beta_1 T_i} + \sum_{i=1}^n \frac{T_i}{(\beta_0+\beta_1 T_i)^2}.$$

Setting $\frac{\partial \ell}{\partial \beta_0} = 0$:

$$-(\alpha-1)\sum_{i=1}^n \frac{1}{\beta_0+\beta_1 T_i} + \sum_{i=1}^n \frac{T_i}{(\beta_0+\beta_1 T_i)^2} = 0.$$

For β_1:

$$\frac{\partial \ell}{\partial \beta_1} = -(\alpha - 1) \sum_{i=1}^{n} \frac{T_i}{\beta_0 + \beta_1 T_i} + \sum_{i=1}^{n} \frac{T_i^2}{(\beta_0 + \beta_1 T_i)^2}.$$

Setting $\frac{\partial \ell}{\partial \beta_1} = 0$:

$$-(\alpha - 1) \sum_{i=1}^{n} \frac{T_i}{\beta_0 + \beta_1 T_i} + \sum_{i=1}^{n} \frac{T_i^2}{(\beta_0 + \beta_1 T_i)^2} = 0.$$

The Fisher Information matrix is given by:

$$I(\alpha, \beta_0, \beta_1) = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}\right], \quad \theta = (\alpha, \beta_0, \beta_1).$$

For α:

$$I_{\alpha\alpha} = n\psi'(\alpha), \quad \text{where } \psi'(\alpha) \text{ is the trigamma function.}$$

For β_0:

$$I_{\beta_0\beta_0} = \sum_{i=1}^{n} \frac{2T_i}{(\beta_0 + \beta_1 T_i)^3}.$$

For β_1:

$$I_{\beta_1\beta_1} = \sum_{i=1}^{n} \frac{2T_i^2}{(\beta_0 + \beta_1 T_i)^3}.$$

The Cramér-Rao bounds for the variance of the estimators are obtained from the inverse of the Fisher Information matrix:

$$\text{Var}(\hat{\theta}_i) \geq [I^{-1}(\alpha, \beta_0, \beta_1)]_{ii}.$$

For numerical solutions with failure times $T = \{1.2, 2.1, 3.5, 4.7, 5.9\}$, define the log-likelihood function:

$$\ell(\alpha, \beta_0, \beta_1) = -n \ln \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \ln T_i - (\alpha - 1) \sum_{i=1}^{n} \ln(\beta_0 + \beta_1 T_i) - \sum_{i=1}^{n} \frac{T_i}{\beta_0 + \beta_1 T_i}.$$

Use iterative optimization algorithms (e.g., Newton-Raphson or gradient descent) to solve the first-order equations. Estimated parameters are:

$$\hat{\alpha} = 2.5, \quad \hat{\beta}_0 = 1.0, \quad \hat{\beta}_1 = 0.1.$$

Q10: Estimating Parameters in a Continuous-Time Process

Consider a machine in a factory that experiences breakdowns according to a continuous-time Poisson process. The time between breakdowns $T_1, T_2, \ldots, T_n$ are independent and exponentially distributed with rate parameter λ, where the probability density function (PDF) for each T_i is given by:

$$f_T(t; \lambda) = \lambda e^{-\lambda t}, \quad t \geq 0.$$

The factory wishes to estimate the rate parameter λ (the rate of breakdowns) from a sample of n observations $T_1, T_2, \ldots, T_n$. Your task is to derive the Fisher Information for λ and determine the precision limits of any unbiased estimator of λ.

(a) **Log-Likelihood Function**
Derive the log-likelihood function $\ell(\lambda; T_1, T_2, \ldots, T_n)$ for the sample of breakdown times.

(b) **Fisher Information and Cramér-Rao Bound**

i) Derive the Fisher Information $I(\lambda)$ for the parameter λ based on the log-likelihood function.

ii) Use the Fisher Information to calculate the Cramér-Rao lower bound (CRLB) for the variance of any unbiased estimator $\hat{\lambda}$ of λ.

(c) **Optimal Estimation of Breakdown Rate**

i) Compute the maximum likelihood estimator (MLE) $\hat{\lambda}_{\text{MLE}}$ for λ, and determine its variance. Verify if it achieves the CRLB.

ii) Investigate the effect of sample size n on the variance of the MLE. Derive how the variance decreases as n increases.

(d) **Generalization to a Non-Homogeneous Poisson Process**
Now assume that the rate of breakdowns is not constant but follows a linear trend over time, i.e., $\lambda(t) = \lambda_0 + \lambda_1 t$, where λ_0 and λ_1 are unknown parameters. Derive the log-likelihood function for this model and extend the Fisher Information calculation to this two-parameter case. Determine the CRLB for estimating both λ_0 and λ_1.

Q11: Analyzing Investment Returns Using MGFs

An investor is interested in the returns of a particular stock over a period of time. Let X be a random variable representing the annual return of the stock, which follows a normal distribution $X \sim N(\mu, \sigma^2)$ with mean $\mu = 0.08$ (8% annual return) and variance $\sigma^2 = 0.04$ (16% standard deviation).

(a) **Moment-Generating Function (MGF)**
Find the moment-generating function (MGF) of the random variable X. Recall that the MGF of a normal distribution $X \sim N(\mu, \sigma^2)$ is given by:

$$M_X(t) = \mathbb{E}\left[e^{tX}\right].$$

(b) **Mean and Variance from the MGF**
Use the moment-generating function derived in part (a) to find the expected value $\mathbb{E}[X]$ and the variance $\text{Var}(X)$. Recall that the mean and variance can be obtained from the MGF using:

$$\mathbb{E}[X] = M_X'(0), \quad \text{Var}(X) = M_X''(0) - (M_X'(0))^2.$$

(c) **Calculate Higher Moments**
Calculate the third moment $\mathbb{E}[X^3]$ using the MGF. Using this, find the skewness of the distribution. The skewness is defined as:

$$\text{Skewness}(X) = \frac{\mathbb{E}[(X - \mathbb{E}[X])^3]}{(\text{Var}(X))^{3/2}}.$$

Q12: Fisher Information and Bayesian Inference(Optional)

In this problem, you will explore the relationship between Fisher Information, Maximum Likelihood Estimation (MLE), and Bayesian inference for the Gamma distribution. We will also calculate the confidence and credible intervals for different sample sizes.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables drawn from a Gamma distribution with the following probability density function (PDF):

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0, \alpha > 0, \beta > 0.$$

Assume that $\alpha = 5$ is known, and we aim to estimate the rate parameter β.

1. **Part 1: Maximum Likelihood Estimation (MLE)**
Derive the MLE for the parameter β given a sample of size n. Then, compute the MLE using the sample means from the following data sets for $n = 200$ and $n = 1000$:

$$X_1, X_2, \ldots, X_n \sim \Gamma(5, \beta).$$

Compute the MLE for β and the Fisher Information at each sample size.

2. **Part 2: Fisher Information and Posterior Distribution**
   Assume a Gamma prior for β with hyperparameters $\alpha_0 = 2$ and $\beta_0 = 1$. Calculate the posterior distribution of β using Bayesian updating for the two sample sizes. Then, compute the Fisher Information of the posterior distribution for both sample sizes.

3. **Part 3: Confidence and Credible Intervals**
   Calculate the 95% confidence interval for β based on the MLE for $n = 200$ and $n = 1000$. Additionally, calculate the 95% credible interval for the posterior distribution of β for each sample size.

Hint: Use the fact that for the Gamma distribution, the MLE for β is given by:

$$\hat{\beta} = \frac{\alpha}{\bar{X}},$$

where $\bar{X}$ is the sample mean. The Fisher Information for β can be derived as:

$$I(\beta) = \frac{n\alpha}{\beta^2}.$$