Statistical Inference

HW Author: Amirhossein Roshandel
Instructor: Mohammadreza A. Dehaqani

Fall 2025

Homework 2

- **Important:** This homework (Homework 2) was uploaded on **Azar 1, 1404**, and you have **two weeks** to complete it. The deadline is **End of Day Azar 15, 1404**. Plan your time wisely and don't leave everything until the last minute!
- If something about the homework looks mysterious, don't silently suffer! Send an email to the HW authors or the Chief TA (with the subject line starting with "ISI Fall 2025"). Asking early usually saves you a lot of stress later.
- If your question could help others too, please post it in the class group. Good questions are part of your contribution to the course, and the TA team will do their best to jump in and clarify.
- Before diving into the problems, take a minute to review the course guidelines on submission format, grace days, and delay penalties. There are no extensions in this course, but you *do* have a limited number of grace days—treat them like a small emergency budget and spend them wisely.
- For computational questions, your grade is not just about getting numbers out of your code; it's about what those numbers *mean*. Write your report like a short academic piece: no code in the PDF, clear and well-labeled figures, precise captions, and focused discussion of your findings.
- You are welcome to use AI tools as a study partner to clarify concepts, explore ideas, or sanity-check your approach. However, the final report must be written by *you* and reflect your own understanding. If you use external resources of any kind (papers, books, websites, discussions), remember to cite them properly.
- For each assignment, there will be a mandatory in-person or virtual hand-in session for the computational parts and an in-person quiz for the theoretical ones. You may be asked to briefly explain your approach and justify your choices. Skipping these evaluation steps unfortunately means receiving a zero for that assignment.
- Above all, think of the assignments as a safe playground for learning statistical inference, not just as a set of hurdles to clear. This is where you can experiment, make mistakes, correct them, and slowly build the skills you'll need in real data and decision-making problems.

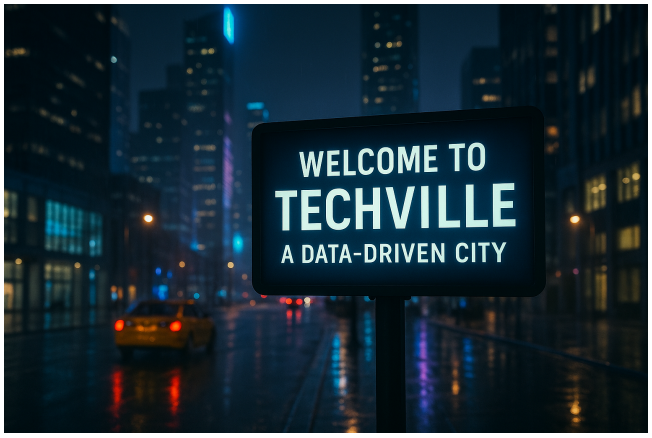

Figure 1: Welcome to TechVille: A Data-Driven City

Problem 1: Statistical Analysis of a Smart City Infrastructure Project

The city of TechVille is implementing a comprehensive smart city infrastructure project. As part of this initiative, the city has hired you as a data analyst to evaluate multiple aspects of the project using statistical methods. Your analysis will help city officials make informed decisions about resource allocation and system improvements.

A) Pipeline Pressure Monitoring System

TechVille has installed IoT sensors throughout its oil pipeline network to monitor pressure levels in real-time. The pressure monitoring system records measurements $X_1, X_2, \ldots, X_n$ that follow the model:

$$X_i = \mu + \epsilon_i$$

where μ is the unknown true average pressure, and ϵ_i represents random measurement error from sensor noise and environmental fluctuations (i.i.d. with mean 0 and unknown standard deviation σ).

Over the first week of operation, the system collected 100 pressure measurements. The sample mean pressure is 75,348 Pascals, with a sample standard deviation of 25 Pascals.

a) Construct an approximate 95% confidence interval for the true average pressure μ of the pipeline.

b) The city's chief engineer asks you to clarify what this confidence interval represents. The interval you constructed in part (a) serves one of the following purposes. Indicate which is correct and briefly explain your reasoning:

(i) To estimate the average of the 100 pressure measurements that were collected and give ourselves some margin for error in the estimate

(ii) To estimate the true average pressure of the pipeline and give ourselves some margin 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 future pressure measurements are likely to fall

c) The engineering team wants to improve the monitoring system's precision for safety compliance. They need to ensure that their estimate of the average pressure is within 1 Pascal of the true pressure μ with 95% confidence. How many pressure measurements should the system collect to achieve this precision?

B) Smart Vending Machine Calorie Monitoring

As part of the smart city's public health initiative, TechVille has installed smart vending machines that automatically verify nutritional content of products. The city is monitoring snack packs advertised as containing "100 Calories" to ensure truth in advertising. The calorie measurement system has a known stable standard deviation of $\sigma = 5$ calories.

The smart vending machines collected data from $n = 36$ snack packs, yielding a sample mean of $\bar{x} = 102.5$ calories.

a) Based on the Central Limit Theorem, what is the sampling distribution of the sample mean $\bar{X}_n$? Specify both its mean and standard error.

b) Construct a 95% confidence interval for the true mean calorie content μ of these snack packs. Does this interval suggest the manufacturer's claim of 100 calories might be inaccurate?

c) The city's health department director wants to be more certain before taking regulatory action against the manufacturer. Construct a 98% confidence interval for μ. (Hint: For a 98% confidence interval, the critical value is $z^* = 2.33$.)

d) In one or two sentences, explain to the health department director why the 98% confidence interval is wider than the 95% confidence interval.

C) Smart Health Kiosk Temperature Screening

TechVille has installed smart health kiosks at key public locations as part of its health monitoring infrastructure. To calibrate the system and establish appropriate alert thresholds, you need to understand the distribution of normal body temperatures. Medical research has established that body temperatures of healthy adults are distributed approximately normally with mean $\mu = 98.2°\text{F}$ and standard deviation $\sigma = 0.73°\text{F}$.

a) Using the empirical rule (68-95-99.7 Rule), determine what percentage of healthy adults have a body temperature between 97.47°F and 98.93°F. The city will use this to set the "normal range" displayed at the kiosks.

b) A kiosk user registers a body temperature of 96.8°F. Calculate the Z-score for this temperature. Should the kiosk flag this as requiring medical attention?

c) The city wants to program the kiosks to display a warning when someone's temperature falls in the lowest 3% of the healthy population distribution. What is this cutoff temperature threshold? (Hint: Use $z_{0.03} \approx -1.88$.)

D) Smart City Investment Fund Analysis

To fund the ongoing smart city initiatives, TechVille has created a technology sector mutual fund. The city's financial office needs to communicate the fund's performance to potential investors (city employees and residents). Based on historical performance data, a 95% confidence interval for the mean 1-year return is $[2.5\%, 8.3\%]$.

a) One resident states: "This is great! This means that 95% of us who invest in this fund will see returns between 2.5% and 8.3% over the next year."

Is this statement correct? If not, explain what is wrong with this interpretation and help the resident understand what the interval actually means.

b) Another resident responds: "No, I think it means that if the city's financial office repeated this analysis with 100 different years of data, exactly 95 of those years would produce this exact interval $[2.5\%, 8.3\%]$."

Is this statement correct? If not, explain what is wrong with this interpretation.

c) Provide a correct interpretation of this 95% confidence interval that would be appropriate for the town hall audience. Your interpretation should help residents make informed investment decisions.

E) Smart Waste Management and Recycling Program

The final component of TechVille's smart city project involves smart recycling bins that track participation. Six months after launching the program, the city government wants to evaluate its effectiveness. They analyze data from $n = 500$ households, finding that 310 are actively using the recycling system ($\hat{p} = 0.62$).

a) For a proportion, the population variance is given by $\sigma^2 = p(1-p)$, where p is the true population proportion. Since p is unknown, estimate this variance using the sample proportion $\hat{p} = 0.62$.

b) Calculate the standard error of the sample proportion using the formula:

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

c) Using the Central Limit Theorem, construct an approximate 95% confidence interval for the true proportion p of all households in TechVille that are participating in the recycling program. Based on this interval, write a brief statement for the city council about the program's success. (The city had set a goal of at least 60% participation.)

Problem 2: Predicting the 2016 USA Presidential Election

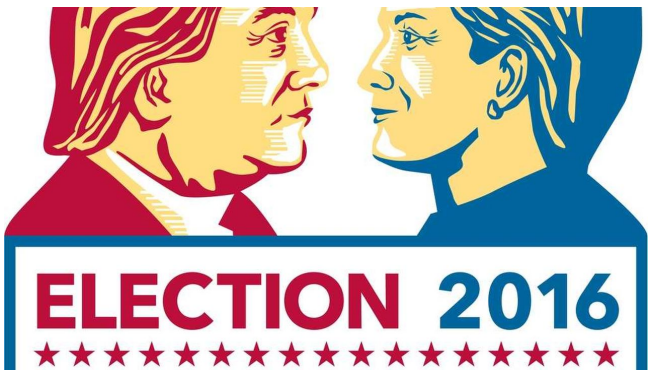

Figure 2: 2016 Presidential Election Polling Data

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.

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.

A) Theoretical Foundation

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.

B) Monte Carlo Verification

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 Part A captures the true proportion p approximately 95% of the time.

C) Data Preparation

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.

D) Time Series Visualization

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.

E) Total Sample Size

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

F) Aggregate Proportions

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

G) Confidence Intervals for Both Candidates

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

Problem 3: Ancient War Strategy and the Law of Large Numbers

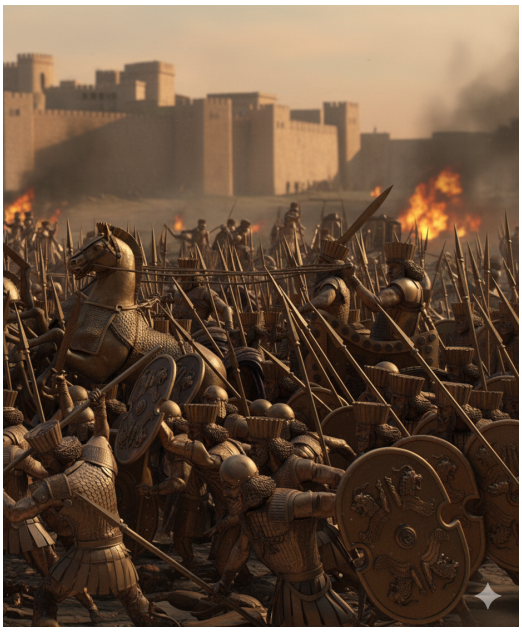

Figure 3: Ancient Persian-Greek warfare

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.

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

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

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

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.)

[label=(c)]

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

2. 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).

3. 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.

4. All the soldiers are sent through one route.

Problem 4: Roulette Simulation and Profit Analysis

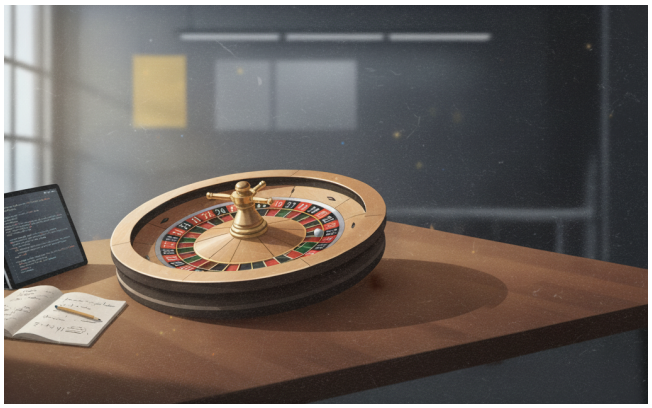

Figure 4: American Roulette

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.

[label=(d)]

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.

Problem 5: Convergence Analysis in Stochastic Processes (Bonus)

Bonus Problem: This problem explores advanced theoretical concepts in convergence theory—going beyond what's strictly required for the course. It's designed for students who want to deepen their understanding of *why* and *how* limiting theorems work at a mathematical level. Successfully completing this problem demonstrates mastery of convergence concepts and will earn you bonus points. *Feel free to skip this if you want to focus on core material first!*

This question explores different types of convergence through both theoretical analysis and computational simulation, integrating concepts from the Law of Large Numbers and Central Limit Theorem.

A) Theoretical Convergence Properties

Consider a sequence of random variables $\{Y_n\}_{n=1}^{\infty}$ where each Y_n is defined as:

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{Z_n}{\sqrt{n}}$$

where $X_1, X_2, \ldots$ are i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$, and $Z_n \sim N(0, \tau^2)$ are independent normal random variables (also independent of the X_i's).

1. Prove that Y_n converges in probability to μ. Specifically, show that for any $\epsilon > 0$:

$$\lim_{n\to\infty} P(|Y_n - \mu| > \epsilon) = 0$$

2. Determine whether Y_n converges to μ in mean-square. That is, evaluate:

$$\lim_{n\to\infty} \mathbb{E}[(Y_n - \mu)^2]$$

3. Find the limiting distribution of $\sqrt{n}(Y_n - \mu)$ as $n \to \infty$. What is the asymptotic variance?

4. Now consider the sequence $W_n = -Y_n$. Prove that:

   - W_n converges in distribution to a constant $-\mu$
   - W_n does NOT converge in probability to μ (assuming $\mu \neq 0$)

   Explain what this reveals about the relationship between convergence in distribution and convergence in probability. Under what conditions would convergence in distribution imply convergence in probability?

B) Dice Simulation and Empirical Convergence

A researcher is studying the convergence properties of sample means using a fair six-sided die. Recall that for a single die roll, $\mu = 3.5$ and $\sigma \approx 1.708$.

1. State precisely what the Law of Large Numbers (LLN) tells us about $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ as $n \to \infty$. Does the LLN tell us anything about the distribution of $\bar{X}_n$ for finite n?

2. State precisely what the Central Limit Theorem (CLT) tells us about the distribution of $\bar{X}_n$ for large n. Write the complete statement including the limiting distribution.

3. Explain the fundamental difference between these two theorems:
    - What question does the LLN answer about $\bar{X}_n$?
    - What question does the CLT answer about $\bar{X}_n$?
    - If you wanted to construct a confidence interval for μ, which theorem would you use and why?

C) Computational Verification

Write a comprehensive Python simulation to demonstrate and compare the LLN and CLT.

1. Implement a simulation with the following specifications:
    - Simulate rolling a fair die n times and compute $\bar{X}_n$
    - Repeat this experiment $K = 10000$ times for each value of $n \in \{5, 10, 20, 50, 100, 500, 1000\}$
    - For each n, store all K values of $\bar{X}_n$
2. To demonstrate the Law of Large Numbers:
    - For each n, calculate the proportion of sample means $\bar{X}_n$ that fall within ϵ of $\mu = 3.5$ for $\epsilon \in \{0.5, 0.3, 0.1, 0.05\}$
    - Create a plot showing these proportions as a function of n (one curve for each ϵ value)
    - Explain how your results demonstrate the LLN
3. To demonstrate the Central Limit Theorem:
    - For each n, compute the standardized sample means: $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$
    - Create a grid of histograms (one for each n value) showing the distribution of Z_n
    - Overlay the standard normal density $N(0, 1)$ on each histogram
    - Calculate and report the empirical mean and variance of Z_n for each n
    - Discuss how the distribution approaches normality as n increases

Problem 6: Margin of Error and Sample Size Determination

The 95% margin of error is defined as $ME = 1.96 \times SE = 1.96 \times \left(\frac{\sigma}{\sqrt{n}}\right)$.

A) Multi-Stage Polling Design

A polling agency is conducting a nationwide survey across three regions with different population characteristics:

Region	**Known σ**	**Target ME**	**Population Weight**
Urban	0.18	0.02	45%
Suburban	0.15	0.015	35%
Rural	0.22	0.025	20%

1. For each region, calculate the minimum sample size required to achieve the target margin of error at 95% confidence level.

2. The agency has a total budget for $n_{total} = 2000$ respondents. Instead of using the individual targets, they want to allocate samples proportionally to population weights. Calculate:
   - The sample size for each region under proportional allocation
   - The actual margin of error achieved in each region
   - Whether any region fails to meet its target ME
3. Prove mathematically that if you want to reduce the margin of error by a factor of k (i.e., $ME_{new} = ME_{old}/k$), the required sample size must increase by a factor of k^2. Use this to calculate:
   - If the urban region wants to cut its ME from 0.02 to 0.01, by what factor must they increase their sample?
   - If the total budget is fixed, what trade-off must be made?

Problem 7: Stratified Sampling and Variance Reduction

A research team is studying average household income in a metropolitan area with 10 districts.

A) Understanding Stratification

The city has three types of districts:

- District A (wealthy): $N_A = 5,000$ households, estimated $\sigma_A = \$80,000$
- Districts B-D (middle-income): Total $N_B = 25,000$ households, estimated $\sigma_B = \$30,000$
- Districts E-J (low-income): Total $N_C = 20,000$ households, estimated $\sigma_C = \$15,000$

Total population: $N = 50,000$ households.

1. If the research team takes a Simple Random Sample (SRS) of $n = 1,000$ households from the entire city:
   - How many households do they EXPECT to sample from each district type (in expectation)?
   - Why is this problematic given that District A is small but has high variability?
   - Could the SRS completely miss District A? What is the probability?
2. Calculate the population proportions:
$$W_A = \frac{N_A}{N}, \quad W_B = \frac{N_B}{N}, \quad W_C = \frac{N_C}{N}$$
3. Under proportional allocation with $n = 1,000$ total samples:
   - Calculate n_A, n_B, n_C such that $n_h = n \cdot W_h$
   - Why is this better than SRS for ensuring representation?
   - Is this the OPTIMAL allocation? Why or why not?

B) Optimal Allocation Theory

1. The Neyman optimal allocation formula states:

$$n_h^{opt} = n \times \frac{W_h \sigma_h}{\sum_k W_k \sigma_k}$$

   Explain the intuition: Why should we allocate MORE samples to strata with:

   - Larger population size (W_h large)?
   - Higher within-stratum variance (σ_h large)?

2. Calculate the optimal allocation $(n_A^{opt}, n_B^{opt}, n_C^{opt})$ for $n = 1,000$ total samples.

3. Compare three allocation schemes:

Allocation Method	n_A	n_B	n_C
Proportional	?	?	?
Equal	333	333	334
Optimal (Neyman)	?	?	?

   For each method, calculate the standard error:

$$SE = \sqrt{\sum_h W_h^2 \frac{\sigma_h^2}{n_h}}$$

   Which method gives the smallest SE? By how much does optimal allocation improve upon proportional?

C) Practical Implications (Optional)

1. A city official argues: "We should sample equally from each district type (equal allocation) because it's fair." Write a response explaining:

   - Why equal allocation is statistically inefficient
   - How it leads to larger standard errors
   - Why fairness in sampling doesn't require equal sample sizes

2. Another official suggests: "Just sample 900 households from District A since it has the highest variance, and split the remaining 100 between B and C." Calculate the SE under this allocation and show that it's worse than optimal allocation despite focusing on the high-variance stratum. Explain why we need to balance variance AND population size.

3. If the researchers could only afford $n = 500$ total samples instead of 1,000:

   - How does the SE change under optimal allocation?
   - By what factor does the SE increase when cutting the sample size in half?
   - Connect this to the margin of error formula: why does halving n multiply SE by $\sqrt{2}$?

Problem 8: Confidence Intervals with Known vs Unknown Variance

A quality control engineer is testing a bottling machine that fills containers.

A) The Known Variance Case

The machine is designed with a precision specification: $\sigma = 0.05$ ounces (this is built into the machine and verified by the manufacturer). An inspector takes $n = 64$ bottles and finds $\bar{x} = 12.08$ ounces.

1. Construct a 95% confidence interval for the true mean fill volume μ using the known $\sigma = 0.05$.
2. Construct a 99% confidence interval using the same data.
3. Compare the widths:
   - Calculate: (Width of 99% CI) / (Width of 95% CI)
   - Explain why higher confidence requires a wider interval
   - By what factor does the width increase when going from 95% to 99% confidence?
4. The inspector also calculates the sample standard deviation and finds $s = 0.07$ ounces. Should they:
   - Use $s = 0.07$ because it's based on actual data?
   - Use $\sigma = 0.05$ because it's the known true value?
   - Average them: $(0.05 + 0.07)/2 = 0.06$?

   Construct CIs using each approach and explain which is correct and why.

B) Unknown Variance - Comparing Approaches

Now consider a NEW machine where σ is unknown. The inspector takes $n = 25$ samples: $\bar{x} = 11.95$ oz, $s = 0.08$ oz

1. A naive approach uses the z-critical value with s:

$$\text{Naive CI} = \bar{x} \pm z_{0.025} \times \frac{s}{\sqrt{25}} = 11.95 \pm 1.96 \times \frac{0.08}{\sqrt{25}}$$

   Calculate this interval.

2. The correct approach for unknown σ uses the t-distribution with $n - 1 = 24$ degrees of freedom. Given $t_{24,0.025} = 2.064$:

$$\text{Correct CI} = \bar{x} \pm t_{n-1,\alpha/2} \times \frac{s}{\sqrt{n}}$$

   Calculate this interval.

3. Compare the two intervals:
   - Which is wider?
   - By what percentage is the correct interval wider than the naive interval?
   - Explain why we need a wider interval when σ is unknown

4. Repeat the calculation for $n = 100$ samples with the same $\bar{x}$ and s. Use $t_{99,0.025} = 1.984$.
    - How do the naive and correct CIs compare now?
    - Why does the difference decrease as n increases?
    - At what sample size are z and t critical values approximately equal?

C) Interpretation and Misconceptions

The inspector reports a 95% confidence interval of $[11.92, 11.98]$ ounces.

Evaluate the following statements (TRUE or FALSE, with explanation):

1. "There is a 95% probability that the true mean μ is between 11.92 and 11.98."
2. "If we repeated this sampling procedure many times, approximately 95% of the constructed intervals would contain the true μ."
3. "95% of all bottles have fill volumes between 11.92 and 11.98 ounces."
4. "We are 95% confident that the sample mean $\bar{x}$ is between 11.92 and 11.98."
5. "If we took a larger sample, the confidence interval would be narrower (assuming the same confidence level)."

D) Sample Size for Desired Precision (Optional)

1. The quality control manager wants the 95% CI to have a total width of at most 0.04 ounces (i.e., ME ≤ 0.02). Based on historical data, they know $\sigma \approx 0.08$ ounces.

    How many bottles must they sample to achieve this precision?
2. After collecting the samples from part (i), they find that $s = 0.10$ instead of the assumed $\sigma = 0.08$.
    - Does their confidence interval still meet the desired width?
    - How many ADDITIONAL samples would they need?
3. If they want to be conservative and ensure ME ≤ 0.02 even if σ could be as large as 0.12, what sample size should they use?

Problem 9: From MGFs to the Central Limit Theorem: Theory and Counterexamples

This problem bridges the gap between theoretical Moment Generating Functions and the limiting theorems (CLT/LLN). It explores *why* the Central Limit Theorem works for most distributions and examines a "pathological" case where standard statistical intuition breaks down.

A) Proving the Central Limit Theorem via MGFs (Optional)

The most common proof of the CLT relies on Moment Generating Functions and Taylor series expansions. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean $\mu = 0$ and variance $\sigma^2 = 1$. Let $M_X(t)$ be the MGF of a single X_i.

1. Recall the Taylor series expansion of $M_X(t)$ around $t = 0$:

$$M_X(t) = M_X(0) + tM'_X(0) + \frac{t^2}{2!}M''_X(0) + O(t^3)$$

   Substitute the values for $M_X(0)$, $M'_X(0)$ (mean), and $M''_X(0)$ (second moment) to show that for small t:

$$M_X(t) \approx 1 + \frac{t^2}{2}$$

2. Now consider the standardized sample mean $Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\overline{X}_n$ (since $\mu = 0, \sigma = 1$). Show that the MGF of Z_n, denoted as $M_{Z_n}(t)$, can be written as:

$$M_{Z_n}(t) = \left[M_X\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

3. Using the approximation from part (1), show that:

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{t^2/2}$$

   Hint: Use the standard limit identity $\lim_{n\to\infty}(1 + \frac{x}{n})^n = e^x$.

4. **Interpretation:** Recognize the function $e^{t^2/2}$. What distribution corresponds to this MGF? Explain how this result mathematically proves the Central Limit Theorem.

B) The Cauchy Anomaly (When CLT Breaks)

The CLT and LLN rely on the assumption that the population has a finite variance ($\sigma^2 < \infty$). Let's see what happens when this assumption is violated.

Introduction to the Cauchy Distribution: The **Standard Cauchy Distribution** is a continuous probability distribution that, despite having a well-defined probability density function, has neither a defined mean nor a defined variance. This is because the integrals $\int_{-\infty}^{\infty} xf(x)dx$ and $\int_{-\infty}^{\infty} x^2 f(x)dx$ do not converge. The Cauchy distribution appears in physics (especially in resonance phenomena and quantum mechanics) and serves as a classic counterexample in probability theory where our usual statistical intuition completely breaks down. You do not need any prior knowledge of this distribution—all necessary properties are provided below.

The **Standard Cauchy Distribution** has the following Probability Density Function (PDF):

$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad -\infty < x < \infty$$

1. The Cauchy distribution has **undefined mean** and **undefined variance** (the integrals do not converge). Consequently, it has **no MGF**. However, it has a remarkable property known as being "closed under convolution" or "closed under averaging." Specifically, if $X_1, \ldots, X_n$ are i.i.d. Standard Cauchy random

variables, their sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ follows the **exact same** Standard Cauchy distribution as a single observation X_1. In other words, averaging does not change the distribution at all!

Based on this property, determine if the Law of Large Numbers (LLN) holds for the Cauchy distribution. Does $\overline{X}_n$ converge to a single number as $n \to \infty$? Explain your reasoning.

2. If you were to construct a "Confidence Interval" for the center of a Cauchy distribution using the standard formula $\overline{x} \pm 1.96(s/\sqrt{n})$, would increasing the sample size n reduce the width of the interval effectively? Explain why this standard approach fails here and what this tells us about applying standard statistical methods blindly.

C) Computational Contrast: Stability vs. Chaos

Write a short Python script to compare the convergence behavior of a Normal distribution versus a Cauchy distribution.

1. **Normal Case:** Generate $n = 1000$ samples from a Standard Normal distribution ($N(0, 1)$). Calculate the sequence of cumulative averages (i.e., calculate the mean of the first k samples for $k = 1$ to 1000).
2. **Cauchy Case:** Generate $n = 1000$ samples from a Standard Cauchy distribution (use `numpy.random.standard_cauchy`). Calculate the sequence of cumulative averages.
3. **Visualization:** Plot both sequences of cumulative averages on the same graph with:
   - X-axis: Sample Size k (from 1 to 1000)
   - Y-axis: Cumulative Average Value
   - Include a horizontal line at $y = 0$ for reference (the true mean of the Standard Normal distribution)
   - Label both curves clearly

   Observation: Describe the behavior of the two lines. Which one stabilizes ("converges")? Which one continues to fluctuate wildly ("jumps") even with large n? Does the Cauchy line ever settle down?
4. **Conclusion:** Explain what this visual difference tells us about the reliability of the "sample mean" as an estimator in the real world if we suspect our data might have extreme outliers or heavy tails. When should we be cautious about using standard methods like confidence intervals based on the sample mean?

Problem 10: Sample Distribution vs Sampling Distribution

A) Conceptual Understanding

1. Define precisely:
   - Sample distribution: The distribution of _____
   - Sampling distribution: The distribution of _____
2. A statistics instructor conducts two experiments:

   Experiment 1: Take ONE sample of $n = 50$ students, record their exam scores, plot a histogram of the 50 scores.

Experiment 2: Repeat the following 1,000 times: take a new sample of $n = 50$ students, calculate the mean score $\bar{x}$. Then plot a histogram of the 1,000 means.

- Which experiment produces a sample distribution?
- Which experiment produces a sampling distribution?
- Which histogram does the Central Limit Theorem describe?

3. True or False (explain):
   - "As n increases, the sample distribution becomes more normal."
   - "As n increases, the sampling distribution of $\bar{X}_n$ becomes more normal."
   - "The spread of the sample distribution decreases as n increases."
   - "The spread of the sampling distribution decreases as n increases."

B) Detailed Example

Consider a population with the following distribution:

$$P(X = 1) = 0.3, \quad P(X = 2) = 0.2, \quad P(X = 3) = 0.5$$

1. Calculate the population mean μ and standard deviation σ.
2. If we take ONE sample of size $n = 2$, list all possible samples (with replacement) and their probabilities:
   - (1,1), probability = ?
   - (1,2), probability = ?
   - etc.
3. For each possible sample from (ii), calculate the sample mean $\bar{x}$.
4. Create the probability distribution of $\bar{X}_2$ (the sampling distribution):

$$P(\bar{X}_2 = 1) =?, \quad P(\bar{X}_2 = 1.5) =?, \quad \text{etc.}$$

5. Verify that $E[\bar{X}_2] = \mu$ and $\text{Var}(\bar{X}_2) = \frac{\sigma^2}{2}$.

C) CLT Connection

1. For the population in Part B, as n increases ($n = 2, 5, 10, 30, 100$), what happens to:
   - The shape of the sampling distribution of $\bar{X}_n$?
   - The mean of the sampling distribution?
   - The standard deviation of the sampling distribution?
2. The CLT states that for large n:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1)$$

Rewrite this in terms of $\bar{X}_n$ directly. What is the approximate distribution of $\bar{X}_n$ for large n?

3. Explain why the CLT is about the sampling distribution, not the sample distribution. If we took one very large sample ($n = 10,000$) from the population in Part B, would the histogram of those 10,000 individual values look normal?

Problem 11: Integrated Computational Analysis

A) CLT Visual Demonstration

Write a Python script to demonstrate the Central Limit Theorem using an exponential distribution.

1. Create a population of 100,000 values from Exponential($\lambda = 0.1$):

   ```
   import numpy as np
   import matplotlib.pyplot as plt

   population = np.random.exponential(scale=10, size=100000)
   ```

   Calculate and display μ_{pop} and σ_{pop} from this population.

2. For each sample size $n \in \{2, 5, 10, 30, 50, 100\}$:
   - Generate 5,000 sample means (each from n draws)
   - Store them in a list/array

3. Create a 2×3 grid of histograms showing the sampling distributions for each n.

   For each histogram:
   - Plot the 5,000 sample means
   - Overlay the theoretical normal curve: $N(\mu, \sigma^2/n)$
   - Add a title showing n and the empirical mean/SD
   - Add vertical lines at $\mu \pm \sigma/\sqrt{n}$

4. In comments, answer:
   - At what n does the distribution start looking normal?
   - How does the spread change as n increases?
   - Does the center of the distribution change with n?
   - Why does this demonstrate the CLT?

B) Confidence Interval Coverage

Verify that a 95% confidence interval actually captures the true mean 95% of the time.

1. Set true parameters: $\mu = 50$, $\sigma = 10$, $n = 40$, number of simulations = 5,000.

2. For each simulation $k = 1, \ldots, 5000$:
   - Generate $n = 40$ samples from $N(50, 100)$
   - Calculate $\bar{x}_k$ and s_k
   - Construct CI: $[\bar{x}_k - 1.96 \times 10/\sqrt{40}, \bar{x}_k + 1.96 \times 10/\sqrt{40}]$
   - Check if $\mu = 50$ is inside this CI
   - Store TRUE/FALSE
3. Calculate and print:

   ```
   Coverage rate: XXXX / 5000 = 0.XXXX
   ```

4. Repeat the entire simulation for:
   - Different sample sizes: $n \in \{10, 20, 40, 80, 160\}$
   - Different confidence levels: 90%, 95%, 99%

   Create a plot showing coverage rate vs. n (separate curve for each confidence level).
5. In comments, explain:
   - Why the coverage rate should be close to 0.95 (not exactly)
   - What happens to coverage rate as n increases?
   - Why does higher confidence level give higher coverage?

C) Sample Size Calculator (Optional)

1. Create a function:

```
def required_sample_size(sigma, margin_of_error, confidence_level):
    """
    Calculate required n for given ME and confidence level

    Parameters:
    sigma: population standard deviation
    margin_of_error: desired ME
    confidence_level: 0.90, 0.95, or 0.99

    Returns:
    n: required sample size (rounded up)
    """
    # Your code here
```

2. Use this function to create a table:

σ	ME	90% CI	95% CI	99% CI
0.10	0.01	?	?	?
0.10	0.02	?	?	?
0.20	0.01	?	?	?
0.20	0.02	?	?	?

3. Create a plot showing required n as a function of desired ME for $\sigma = 0.15$ and 95% confidence. Use $ME \in [0.005, 0.05]$.

   Add annotations explaining:

   - Why the relationship is non-linear (curved)
   - Why very small ME requires very large n
   - The practical implications for survey design

Deliverables

1. Complete Python code with clear comments
2. All requested plots with proper labels, titles, and legends
3. Brief written analysis (1-2 paragraphs) for each part explaining:
   - What the simulation shows
   - How it connects to theoretical concepts (CLT, CI, sample size formulas)
   - Any surprising results or insights