Lab 3 - The Central Limit Gauntlet

Overview

In this assignment, you and an AI agent (ChatGPT, Gemini, Claude, etc.) will join forces to explore the Central Limit Theorem (CLT) through creativity, mischief, and a dash of statistical chaos.

Throughout this assignment, you remain the chief statistician. You will complete three challenges, each revealing a different face of the CLT: first testing the AI's intuitions against simulation, then reasoning backwards from evidence like a detective, and finally pushing the theorem to its limits.

AI Use Policy

Think of your AI agent as a brainstorming partner and data generator—a collaborator willing to hand you hundreds of observations without complaint and speculate freely about statistical outcomes. You are encouraged to invent scenarios that are whimsical or wildly imaginative, provided they remain appropriate and non-offensive. Your written responses, however, should remain clear, concise, and academically appropriate, even when the contexts themselves are playful.

Keep in mind that the AI's intuitions are not always correct. Part of your job is to evaluate its reasoning critically, not simply accept it.

Deadline: May 18th; 4:00 pm (1 pdf file; Dropbox on Lea)

Note: It is your responsibility to ensure that the pdf is readable. Any empty or corrupted files will result in your lab receiving a grade of zero.

Challenge #1: Beat the AI

Can out-predict a model trained on half the internet?

In this challenge, you will collaborate with your AI agent to invent a creative context, generate data, and test whether the AI's intuition about the CLT matches what actually happens in simulation.

Step 1: Create your scenario

Ask your AI agent to help you invent a creative real-world context involving a strongly non-normal variable (think skewed, lopsided, erratic, or otherwise far from textbook). The scenario should be academically plausible, even if it is absurd.

Step 2: Ask the AI to make predictions

Ask your AI agent to describe what it expects the sampling distribution of the sample mean to look like for each of the following sample sizes: $n = 5$, $n = 20$, and $n = 50$

Record its predictions before running any simulations. This is important: you will be evaluating its reasoning, not just its conclusions.

Step 3: Generate data and run simulations

Ask your AI agent to generate a dataset of approximately 300–500 observations consistent with your scenario. Using software of your choice (Excel, Google Sheets, Python, R, etc.), simulate the sampling distribution of the mean for each of the three sample sizes and produce a histogram for each.

What to Submit

- A brief description of your creative scenario.
- Three histograms of sample means for $n = 5$, $n = 20$, and $n = 50$.
- A paragraph (4–6 sentences) comparing your results to the AI's predictions. Discuss what aligned, what did not, and whether the AI demonstrated genuine statistical reasoning or simply pattern-matched.

Challenge # 2: CLT Detective Mystery

The sampling distributions are your clues. What is the population hiding?

In this challenge, you will step into the role of a statistical detective. Your task is to reconstruct a hidden population using only the sampling distributions of its sample means. Treat it like a mystery: follow the evidence, piece together the details, and uncover the story behind the data.

Step 1: Create your mystery

Ask your AI agent to help you develop a mysterious or intriguing context in which the underlying population is unknown and must be inferred. The scenario should lend itself to genuine ambiguity; something that cannot be solved by guessing alone.

Step 2: Generate the clues

Ask your AI agent to generate a dataset of 400–600 observations from a hidden population of its choosing (it should not reveal the population to you yet). Using that data, simulate and produce histograms of the sampling distribution of the sample mean for: $n = 5$, $n = 30$, and $n = 100$

Step 3: Reconstruct the hidden population

Using the three histograms as your only evidence, apply the following detective toolkit to reason about the original population:

- *Shape at small n:* If the $n = 5$ histogram is strongly skewed or irregular, the population likely shares that shape. Skewness in small samples is a direct echo of the population.
- *Rate of convergence:*How quickly does the distribution approach normality as n grows? Rapid convergence suggests a well-behaved population; slow convergence hints at heavy tails or extreme values.
  - *Persistence of features:* If multimodality or asymmetry persists even at $n = 30$ or $n = 100$, the underlying population may have multiple subgroups or an extremely heavy tail.
  - *Spread and scale:* The standard deviation of the sampling distribution equals $\sigma/\sqrt{n}$. Use the observed spread at different sample sizes to estimate the population's variability.

Once you have formed your hypothesis, ask the AI to reveal the true population. Compare your inference to the actual answer.

What to Submit

- A short paragraph describing your creative mystery scenario.
- Three histograms of the sampling distribution for $n = 5$, $n = 30$, and $n = 100$.
- A case analysis (6–8 sentences) stating the likely shape of the hidden population, whether it is discrete or continuous, and at least two plausible real-world explanations consistent with your reasoning.
- One or two sentences reflecting on how close your inference was once the AI revealed the true population.

Challenge 3: The CLT Stress Test

Every theorem has a breaking point. Find this one's.

The first two challenges demonstrated that the CLT works. This one asks a harder question: when does it struggle? Your goal is to deliberately construct a scenario where convergence to normality is unusually slow or problematic, and then explain why mathematically.

Step 1: Find a resistant distribution

Ask your AI agent to suggest a distribution where CLT convergence is known to be slow or theoretically problematic. Ask it to explain why the distribution is resistant. Good candidates include distributions with heavy tails, extremely rare but enormous values, or properties that technically violate CLT assumptions. Record the AI's explanation before proceeding.

Step 2: Generate your stress-test dataset

Ask the AI to generate 400–600 observations from this resistant distribution. Produce a histogram of the raw population data so the unusual shape is visible.

Step 3: Push it to the limit

Simulate the sampling distribution of the mean at larger sample sizes than the previous challenges required: $n = 10$, $n = 50$, and $n = 200$.

Produce histograms for each and pay close attention to how slowly (or quickly) the distribution approaches normality compared to what you saw in Challenge #1.

Step 4: Explain the resistance

Research the distributional property that makes your chosen distribution resistant to the CLT. You should be able to identify a specific mathematical reason (not just an observation that it "looks non-normal").

What to Submit

- A brief description of your stress-test scenario and the distribution you chose.
- A histogram of the raw population data.
- Three histograms of the sampling distribution for $n = 10$, $n = 50$, and $n = 200$.
- A paragraph (4–6 sentences) connecting the distributional property you identified to the theoretical reason the CLT converges slowly. Compare the rate of convergence to what you observed in Challenge # 1.

Tip: The most interesting submissions are the ones that surprise you. If everything goes exactly as expected, push harder.