Project 1: Breaking the Central Limit Theorem

BYU STAT 250

Introduction

The **Central Limit Theorem (CLT)** is one of the most fundamental results in probability and statistics. It states that, given a sufficiently large sample size, the distribution of the sample mean will be approximately normal, regardless of the shape of the original population distribution. In many cases, people treat $\mathbf{n = 30}$ as a magical threshold where normality is assured.

However, is this always true? In this project, you will investigate whether there exist distributions for which, even when using $\mathbf{n = 30}$, the sample mean does not appear normally distributed according to the **Shapiro-Wilk test**.

Objective

Your goal is to:

1. Select a non-normal distribution.
2. Simulate sampling distributions of the mean for $\mathbf{n = 30}$.
3. Apply the **Shapiro-Wilk test** to check for normality.
4. Identify a distribution where the Shapiro-Wilk test more often than not rejects normality at $\mathbf{n = 30}$.
5. Write a report discussing your findings, including well-documented code and interpretation of results.

Sections of you report

Introduction

- Clearly explain the Central Limit Theorem (CLT).
- Discuss the assumption that $\mathbf{n = 30}$ is typically sufficient for normality.

Simulation Study

- Select a non-normal distribution that may violate the CLT at $\mathbf{n = 30}$.
- Write well-commented R code that:
  - Generates many samples of size $\mathbf{n = 30}$ from your chosen distribution.
  - Computes the sample means.
  - Applies the **Shapiro-Wilk test** to these sample means.
- Summarize how often the Shapiro-Wilk test rejects normality.

Conclusion

- Summarize your results
- Try to provide some explanation or discussion as to why your chosen distribution breaks the CLT.

Submission Details

- **Format:** Submit a **PDF report** including your write-up and code.
- **Collaboration:** You may work **individually or in a group of up to 3 people**.
- **Need a Group?** If you want to work in a group but don't have one, **let me know and I will match you with others**.
- **Class Time:** We will discuss good presentation details in class and provide time to work on this project.
- **Due Date: February 26th**.

Grading Rubric

Criterion	Points
Successfully finding a distribution that breaks CLT at $\mathbf{n = 30}$	40%
Well-commented, logical code	30%
Clear and well-organized presentation of results	30%

Total: **100 points**

Getting Started

Here's a basic framework to start:

```r
set.seed(123)

# Define parameters
n <- 30    # Sample size
simulations <- 10000  # Number of simulations

# Generate samples from a chosen non-normal distribution
sample_means <- replicate(simulations, {
  sample_data <- YOUR_DISTRIBUTION_FUNCTION(n)  # Replace with your chosen distribution
  mean(sample_data)
})

# Perform the Shapiro-Wilk test on sample means
shapiro_result <- shapiro.test(sample_means)
print(shapiro_result)
```

You will need to modify YOUR_DISTRIBUTION_FUNCTION to test different distributions and analyze results.