Project Proposal

Research Question

When the underlying population distribution is skewed, heavy-tailed, or discrete (e.g., Bernoulli, Exponential, Lognormal, t-distribution), how large must the sample size n be before the standardized sample mean is "close enough" to normal? In other words, how valid is the heuristic rule of thumb that "$n \geq 30$ is sufficient" for the Central Limit Theorem to apply?

Analysis Plan

I will conduct a Monte Carlo simulation study to evaluate how quickly the distribution of the standardized sample mean approaches the standard normal under different underlying population distributions. For each candidate distribution (Bernoulli(p), Exponential(1), Lognormal, and Student's t with small degrees of freedom), I will generate a large pseudo-population and repeatedly draw random samples of varying sizes ($n = 10, 20, 30, 50, 100, 200, 500$).

For each sample size, I will compute standardized sample means across many replications (e.g., 2,000 resamples) and assess their closeness to normality using several criteria:

- Kolmogorov–Smirnov (KS) statistic comparing the empirical distribution of standardized means with $N(0, 1)$.
- QQ-plot linearity measures (e.g., R^2 from a regression of quantiles).
- Coverage rate of nominal 95% z-confidence intervals for the true mean.

The results will provide empirical guidance on how large n must be for the CLT approximation to perform adequately under different non-normal conditions. I will summarize findings in both tables and plots (KS distance vs. n, coverage vs. n).