M5-04: Central Limit Theorem

Part of the "Polling, Confidence Intervals, and the Normal Distribution" Learning Badge

Video Walkthrough: https://discovery.cs.illinois.edu/m5-04/

The Central Limit Theorem

The normal approximation for random variables amounts to taking advantage of the *Central Limit Theorem*. We replace the true probability histogram for the sum, average, or percentage of draws by the normal curve before computing areas.

Central Limit Theorem:

- Regardless of the distribution shape of the population, the sampling distribution of the sample mean becomes approximately normal as the sample size n increases (conservatively n≥30).
- In other words, if we repeatedly take independent random samples of size *n* from any population, then when *n* is large, the distribution of the sample means will approach a normal distribution.

This is very interesting! So it doesn't matter if the distribution shape was left-skewed, right-skewed, uniform, binomial, anything - the distribution of the sample mean will always become normal as the sample size increases. **What an amazing result! The CLT is one of the most important theorems in statistics!**

We can use the central limit theorem and the normal curve to calculate areas just like we did previously. However, when dealing with random variables, our z-score formula changes slightly:

Convert to standard units using the EV for the average and the SE for the SD.

$$Z = (Value - EV) / SE$$

Puzzle #1: In 100 coin tosses, what is the chance of getting less than 45 heads?

We could figure out the exact probability histogram (pmf), but it would take too long since there are 2^{100} possible ways the coin could fall. Since there are 100 draws we know that the probability histogram will be very close to the normal curve (CLT), so we can use that instead