Central Limit Theorem

Gregory M. Shinault

Big Idea

Recall the normal approximation to the binomial distribution. A binomial random variable can be expressed as a sum of IID Bernoulli RVs.

It turns out this can be restated as a sum of (almost) any IID RVs and the normal approximation is still valid.

This material corresponds to section 9.3 of the textbook.

Formal Statement

Suppose $X_1, X_2, \ldots$ is an IID sequence of RVs with $\mathbb{E}X_1 = \mu < \infty$ and $\mathrm{Var}(X_1) = \sigma^2 < \infty$. Then for all real t we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le t\right) = \Phi(t)$$

where $\Phi(t)$ is the standard normal CDF.

Intuitive Interpretation

1. Sums of RVs are approximately normal:

$$X_1 + \cdots + X_n \approx Y \sim N(n\mu, n\sigma^2).$$

2. Sample means are approximately normal:

$$\bar{X_n} = \frac{X_1 + \cdots + X_n}{n} \approx Y \sim N(\mu, \sigma^2/n).$$

Key Idea in Proof of CLT | Continuity Theorem for MGFs

Assume the MGFs of $S_1, S_2, \ldots$ and Z satisfy

$$\lim_{n\to\infty} M_{S_n}(t) = M_Z(t)$$

for all $-\varepsilon < t < \varepsilon$ for some $\varepsilon > 0$. Then

$$\lim_{n\to\infty} \mathbb{P}(S_n \le s) = \mathbb{P}(Z \le s)$$

for all real s.

Example

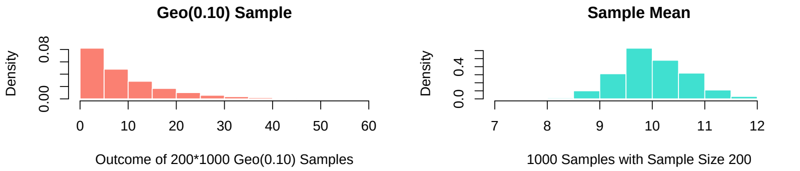
Suppose X and $\mathbb{E}X = 50$.

1. Give a bound on $\mathbb{P}(X \geq 60)$.
2. Assume further $\sigma^2 = 25$. Give a bound on $\mathbb{P}(X \geq 60)$.
3. Assume further X is binomial. Give a bound on $\mathbb{P}(X \geq 60)$.

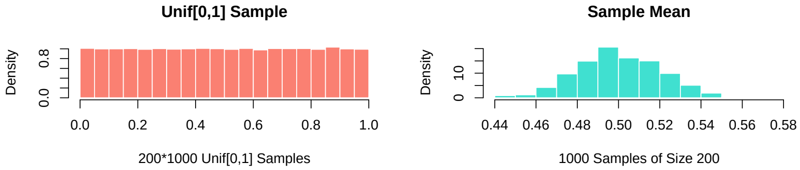
Example

We roll 1000 dice and add up the values. Estimate the probability that the sum is at least 3600.

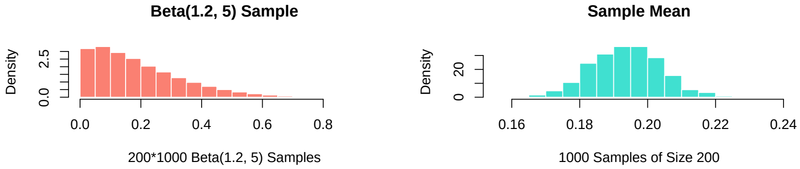
Simulation of Sample Mean



Simulation of Sample Mean



Simulation of Sample Mean



The Key Lesson

The data itself does not become normally distributed because the sample size is large. The distribution for the sample mean does, because it is a sum of random variables.

The CLT does not say large samples are normally distributed. It says that large sums are normally distributed.

Example

A brewer makes 25 independent measurements of the specific gravity of a certain wort. We will denote the true specific gravity of the wort by s. She knows that the limitations of her equipment are such that the standard deviation of each measurement is σ. The good news is that her equipment is unbiased, so the expected value of each measurement is s.

Estimate the probability that the average of her measurements will differ from the actual specific gravity of the wort by less than $\sigma/4$ units. Note that your final answer should be a number that does not depend on any parameters.

Summary

1. The MGF is a critical mathematical tool in proving the CLT.
2. The CLT roughly says that large sums are approximately normally distributed, and thus large sample means are approximately normally distributed.
3. The CLT does NOT say that a lot of data will be normally distributed, only that data based on sum of random variables will be approximately normally distributed.