LLN and the Central Limit Theorem

Tyler Zhu

April 29, 2020

The idea of this worksheet is to get you to the point where applying the Central Limit Theorem becomes natural and doesn't require memorizing a formula to apply it. If you actually try solving each problem on its own, this should do its job.

1 Warm-Up: Law of Large Numbers

Problem 1. This problem will walk you through a proof of the Law of Large Numbers. The setup is as follows: we have n i.i.d. random variables $X_1, \ldots, X_n$ where $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Our intuition tells us that in the long run, we'd expect the average of these random variables to approach the true underlying mean.

(a) Create an unbiased estimator M_n for the mean μ.

(b) What are $\mathbb{E}[M_n]$ and $\mathrm{Var}(M_n)$?

(c) Suppose I used Chebyshev's inequality to create a confidence interval $[\mu - \epsilon, \mu + \epsilon]$ for M_n. As $n \to \infty$, how confident am I that M_n will be in this interval? How does this relate to the law of large numbers?

So LLN tells us that their average will eventually be very close to the mean, but it doesn't tell me how confident I can be for a particular n. That's where CLT comes in.

2 Central Limit Theorem

First we'll see how we can derive the Central Limit Theorem to gain intuition for why it works, then I'll present it.

Problem 2. Your friend comes up to you and tells you about the new BLT theorem, which says that the sum of n i.i.d. random variables is approximately Gaussian for large n, but forgot what the parameters of it looks like. It's your job to figure out exactly what they should be. Similar to the last problem, suppose that $X_1, \ldots, X_n$ are i.i.d. where $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$.

(a) Let $S_n = X_1 + \cdots + X_n$. What are $\mathbb{E}[S_n]$ and $\mathrm{Var}(S_n)$?

(b) Define a new r.v. Z_n in terms of S_n which has mean 0 and standard deviation 1.

(c) The *z-score* of a data point is how many standard deviations (+/-) it's away from the mean. Given some observation x in the distribution S_n, what would it's z-score be?

(d) Your friend's BLT theorem tells you that $Z_n \sim \mathcal{N}(0, 1)$, i.e. it's approximately a standard normal Gaussian, in the following sense:

$$\lim_{n\to\infty} \mathbb{P}[Z_n \le z] = \Phi(z)$$

where $\Phi(z)$ is the CDF of the standard normal Gaussian. Given this, what is $\mathbb{P}[S_n \le x]$ for some x?

Here's the general theorem for reference.

Theorem 1 (Central Limit Theorem)**.** Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s where $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, and define $S_n = X_1 + \cdots + X_n$. Then as $n \to \infty$, the CDF of S_n approaches that of the CDF of a $\mathcal{N}(n\mu, n\sigma^2)$ distribution. In terms of z-scores, if $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$, then

$$\lim_{n\to\infty} \mathbb{P}[Z_n \leq z] = \Phi(z)$$

where $\Phi(z)$ is the CDF of a standard normal Gaussian.

3 Past Practice Problems

Problem 3 (Sp19 Final #7)**.** Suppose the number of people that walk into Jonathan's favorite McDonald's in an hour is $\sim$ Poisson(λ), where λ is unknown but is definitely at most 10. How many hours does Jonathan need to be at McDonalds to be able to construct a 95% confidence interval for λ that is of width 2?

Problem 4 (Fa19 Final #8)**.** Let $X_i \sim$ Poisson(1) be independent r.v.'s and $S_n = X_1 + \cdots + X_n$ be their sum, and let c, ϵ be some constants. For $\epsilon < \frac{1}{2}$, what is $\lim_{n\to\infty} \mathbb{P}[S_n < cn^\epsilon + n]$?