1 Central Limit Theorem

Recall that the Law of Last Numbers says that if $X_1, X_2, \ldots$ are IID with finite means μ and variances σ^2, then the averages $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i \mapsto \mu$. In particular, we proved the LLN using Chebychev's Inequality, which gives

$$\mathbb{P}\left(|A_n - \mu| \geq C\frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{C^2}.$$

The **Central Limit Theorem (CLT)** says more. The Central Limit Theorem says that the normalized averages

$$Z_n = \frac{A_n - \mu}{\sigma/\sqrt{n}}$$

converge in distribution (this sequence of RV converges to another RV) to a standard Normal(0, 1) random variable Z. More precisely, this means that the CDFs

$$F_{Z_n}(z) \mapsto F_Z(z)$$

as $n \mapsto \infty$.

<table>
<tr><th>Chebyshev's Inequality</th><th>Central Limit Theorem</th></tr>
<tr><td>

$$\mathbb{P}\left(|A_n - \mu| \geq C\frac{\sigma}{\sqrt{n}}\right) = \boxed{\mathbb{P}(|Z_n| \geq C) \leq \frac{1}{C^2}}.$$

So, this gives us an upper-bound. Note that this works for *any* n.

</td><td>

The CLT says that $\mathbb{P}(|Z_n| \geq C) \mapsto \mathbb{P}(|Z| \geq C)$, and

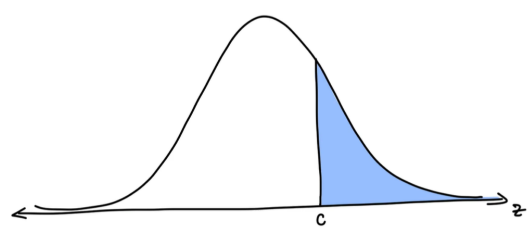
$$\mathbb{P}(|Z| \geq C) = 2\int_C^\infty \frac{e^{-z^2/2}}{\sqrt{2\pi}}dz$$

as $n \mapsto \infty$. Note that this works better for significantly large values of n.

Note that

$$\int_C^\infty \frac{e^{-z^2/2}}{\sqrt{2\pi}}dz$$

is the area under the standard "bell-shaped curve" (i.e., the Normal(0, 1) PDF) to the right of $z = C$.



As a remark, the integral above doesn't have an antiderivative, but we can make use of an online z-score calculator to find (very good approximations to) these values.

</td></tr>
</table>

Remark: For this class, we usually let $\mu = 0$, $\sigma = 1$, and x be the value of interest ($|Z|$, for example).

(Example.) We note that, by using a z-score calculator, we know that

$$\mathbb{P}(|Z| \geq 2) = 2\mathbb{P}(Z \geq 2) \approx 4.55\%.$$

Using Chebyshev's Inequality, we find that the upperbound is $\leq \frac{1}{2^2} = 25\%$.

Theorem 1.1: Central Limit Theorem

Suppose that $X_1, X_2, \ldots$ are IID with common mean $\mu < \infty$ and variance $\sigma^2 < \infty$. Put

$$S_n = \sum_{i=1}^{n} X_i.$$

Then, for any $b \in \mathbb{R}$,

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right) \mapsto \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b} e^{-z^2/2} dz.$$

Note that $\mathbb{E}(S_n) = n\mu$ and $SD(S_n) = \sqrt{\mathrm{Var}(S_n)} = \sigma\sqrt{n}$.

Remarks:

- Note that
$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{A_n - \mu}{\sigma/\sqrt{n}}$$
where $S_n = \sum_{i=1}^{n} X_i$ is the sum and $A_n = \frac{1}{n} S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the average.
- The key is that you do not need to know the actual distribution of the X_i's. We only need to know that they are IIDs (and that their means are variances exist). So, in essense, *the CLT gives useful information about averages of a distribution without needing to know what the distribution really is.*

Note that, when applying the CLT to *discrete* IID sequences $X_1, X_2, \ldots$, it is often useful to make a "discrete adjustment" to get a slightly better approximation.

(Example: Normal Approximation to the Binomial.) Recall that a Binomial(n, p) RV X_n is the sum of n IID Bernoulli(p) RVs, and that its mean is np and variance npq, where $q = 1 - p$. Thus, by the CLT,

$$\mathbb{P}(i \le X_n \le j) \approx \mathbb{P}\left(\frac{i - np}{\sqrt{npq}} \le Z \le \frac{j - np}{\sqrt{npq}}\right)$$

for large n.

We can, however, get a better approximation (unless n is very large, in which case it makes little difference) if we instead approximate

$$\mathbb{P}(i \le X_n \le j) \approx \mathbb{P}\left(\frac{i - 1/2 - np}{\sqrt{npq}} \le Z \le \frac{j + 1/2 - np}{\sqrt{npq}}\right).$$

The reason why is because this makes a correction to get all of the relevant "bars."

(Example.) A fair coin is tossed 100 times. Estimate the probability that "Heads" is tossed between 40 and 60 times.

Let $S_n = \sum_{i=1}^{n} X_i$ where X_i indicates if the ith toss is "Heads." Letting X_i be a Bernoulli, where X_i is 1 if the ith toss is "Heads" and 0 otherwise. Then, applying the Binomial approximation, we have

$$\begin{aligned}\mathbb{P}(40 \le S_n \le 60) &= \mathbb{P}\left(\frac{40 - 0.5 - 100(0.5)}{\sqrt{100(0.5)(1-0.5)}} \le Z \le \frac{60 + 0.5 - 100(0.5)}{\sqrt{100(0.5)(1-0.5)}}\right) \\ &= \mathbb{P}(|Z| \le 2.1).\end{aligned}$$

So, using the online calculator, we want to compute

$$\mathbb{P}(-2.1 \leq Z \leq 2.1).$$

Doing this (letting $\mu = 0$, $\sigma = 1$, $x = 2.1$, and $\mathbb{P}(-|x| < X < |x|)$ in the dropdown menu) gives us 96.42%.

Without the discrete correction, we would have found

$$\mathbb{P}(-2 \leq Z \leq 2) \approx 95.45\%.$$

But, by calculating the true probability, we get

$$\frac{1}{2^{100}} \sum_{i=40}^{60} \binom{100}{i} = 96.479\ldots\%.$$

Theorem 1.2

If the $X_1, X_2, \ldots$ are independent with means $\mu_i < \infty$ and variances $\sigma^2 < \infty$, and such that for some constant $A < \infty$, all $|X_i| \leq A$ (i.e., the X_i are "uniformly bounded"), then the conclusion of the CLT still holds; that is,

$$\mathbb{P}\left(\frac{S_n - \mathbb{E}(S_n)}{SD(S_n) \leq b}\right) \mapsto \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b} e^{-z^2/2} dz.$$

Remark: Note that $\mathbb{E}(S_n) = \sum_{i=1}^n \mu_i$ and $SD(S_n) = \sqrt{\sum_{i=1}^n \sigma_i^2}$.
Course Note: This will not be tested on homework assignments or exams.