Lisa Yan and Jerry Cain
CS109

Lecture Notes #18
October 23, 2020

Central Limit Theorem

Based on a chapter by Chris Piech

Pre-recorded lecture: Sections 1 and 2
In-lecture: Section 3

1 Independent and Identically Distributed Random Variables

The variables $X_1, X_2, \dots, X_n$ are *independent and identically distributed* (often written i.i.d., iid, or IID) if $X_1, X_2, \dots, X_n$ are independent and each have the same distribution—meaning they have the same PMF (if X_i is discrete) or PDF (if X_i is continuous).

1.1 Examples of IID Random Variables

- For $i = 1, \dots, n$, let $X_i \sim \text{Exp}(\lambda)$, where the X_i are independent. $X_1, X_2, \dots, X_n$ are IID.
- For $i = 1, \dots, n$, let $X_i \sim \text{Exp}(\lambda_i)$, where the X_i are independent. $X_1, X_2, \dots, X_n$ are not IID (unless $\lambda_i = \lambda$ for some constant λ and $i = 1, \dots, n$).
- For $i = 1, \dots, n$, let $X_i \sim \text{Exp}(\lambda)$, where $X_1 = X_2 = X_n$. $X_1, X_2 \dots, X_n$ are not IID because the X_i are dependent.
- For $i = 1, \dots, n$, let $X_i \sim \text{Bin}(n_i, p)$, where the X_i are independent. $X_1, X_2, \dots, X_n$ are not IID (unless $n_i = n$ for some constant n and $i = 1, \dots, n$).

2 The Theory

The Central Limit Theorem (CLT) proves that the averages of samples from *any* distribution themselves must be normally distributed. Consider IID random variables $X_1, X_2 \dots$ such that $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

The Central Limit Theorem states:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \qquad \text{as } n \to \infty$$

It is sometimes expressed in terms of the standard normal, Z:

$$Z = \frac{\left(\sum_{i=1}^{n} X_i\right) - n\mu}{\sigma\sqrt{n}} \qquad \text{as } n \to \infty$$

At this point you probably think that the Central Limit Theorem is awesome. But it gets even better. With some algebraic manipulation we can show that if the sample mean of IID random variables is normal, it follows that the sum of equally weighted IID random variables must also be normal. Let's call the sum of IID random variables $\bar{Y}$:

$$\begin{aligned}
\bar{Y} &= \sum_{i=1}^{n} X_i = n \cdot \bar{X} && \text{If we define } \bar{Y} \text{ to be the sum of our variables} \\
&\sim N(n\mu, n^2 \frac{\sigma^2}{n}) && \text{Since } \bar{X} \text{ is a normal and } n \text{ is a constant.} \\
&\sim N(n\mu, n\sigma^2) && \text{By simplifying.}
\end{aligned}$$

In summary, the Central Limit Theorem explains that both the sample mean of IID variables is normal (regardless of what distribution the IID variables came from) and that the sum of equally weighted IID random variables is normal (again, regardless of the underlying distribution).

Most textbooks will tell you that the CLT holds if $n \geq 30$ (where n is the number of IID random variables you are summing together), but the CLT can hold for smaller n depending on the distribution of your IID random variables.

There are several proofs of the Central Limit Theorem, one of which is in Section 8.3 of the Ross textbook (10th edition). We encourage you to find one that resonates with you.

Normal approximation of the Binomial random variable

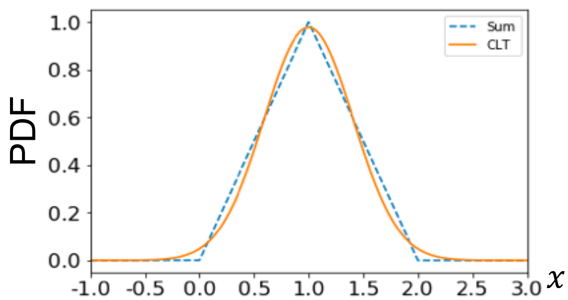
Back in Lecture 10, we discussed that the Binomial random variable could be approximated with a Normal random variable (with continuity correction). The justification for this approximation actually comes from the Central Limit Theorem.

Suppose we have a Binomial random variable, X where $X \sim \text{Bin}(n, p)$. We can rewrite $X = \sum_{i=1}^{n} X_i$, where $X_i \sim \text{Ber}(p)$ for $i = 1, \dots, n$ and all X_i are independent. By definition, $X_1, X_2, \dots, X_n$ are IID and therefore X is the sum of IID random variables. Note that each X_i has mean $\mu = p$ and variance $p(1-p)$. Therefore as n grows large, $X \sim \mathcal{N}(n\mu = np, n\sigma^2 = np(1-p))$.

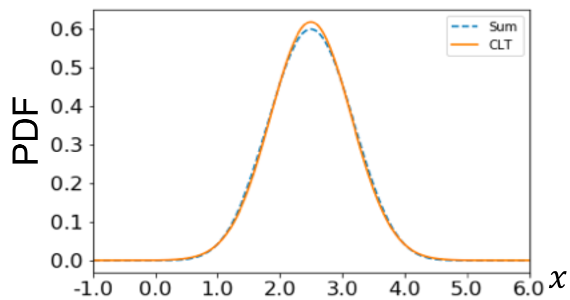
Example: Sum of Uniform random variables

Let $X = \sum_{i=1}^{n} X_i$ be the sum of IID random variables, where $X_i \sim \text{Uni}(0, 1)$. Note that $\mu = E[X_i] = 1/2$ and $\sigma^2 = \text{Var}(X_i) = 1/12$, for $i = 1, \dots, n$.

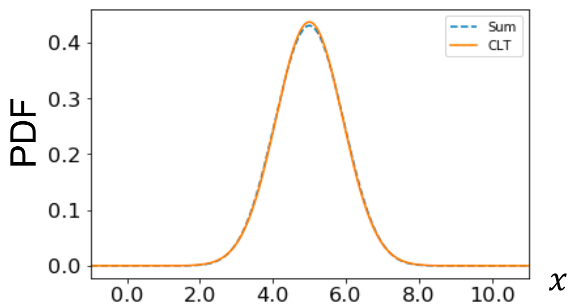
Below, we plot the distribution of X and its normal approximation $Y \sim \mathcal{N}(n\mu, n\sigma^2)$ for different values of n. Note that even when $n = 10 < 30$, the CLT is already a pretty good approximation to the true sum.



(a) $n = 2$



(b) $n = 5$



(c) $n = 10$

3 Exercises

Example: Dice

You will roll a 6 sided dice 10 times. Let X be the total value of all 10 dice = $X_1 + X_2 + \cdots + X_{10}$. You win the game if $X \le 25$ or $X \ge 45$. Use the Central Limit Theorem to calculate the probability that you win.

Recall that $E[X_i] = 3.5$ and $\text{Var}(X_i) = \frac{35}{12}$.

$$\begin{aligned}
P(X \le 25 \text{ or } X \ge 45) &= 1 - P(25.5 \le X \le 44.5) \\
&= 1 - P\left(\frac{25.5 - 10(3.5)}{\sqrt{35/12}\sqrt{10}} \le \frac{X - 10(3.5)}{\sqrt{35/12}\sqrt{10}} \le \frac{44.5 - 10(3.5)}{\sqrt{35/12}\sqrt{10}}\right) \\
&\approx 1 - (2\Phi(1.76) - 1) \approx 2(1 - 0.9608) = 0.0784
\end{aligned}$$

Website crashes

Let X be the number of visitors to a website, where $X \sim \text{Poi}(100)$. The server crashes if there are more than 120 requests in a minute. The probability that the server crashes in the next minute can be computed exactly as $P(X \ge 120) = \sum_{k=120}^{\infty} \frac{100^k e^{-100}}{k!} \approx 0.0282$.

We can also approximate this probability using the Central Limit Theorem. Recall that the sum of independent Poisson random variables is also Poisson. We can therefore arbitrarily define X to be a sum of independent Poisson random variables $X_1, \ldots, X_n$, each of which covers exactly $1/n$ of the minute, for some value of n. Then $X_i \sim \text{Poi}(100/n)$, and therefore $X_1, \ldots, X_n$ are IID.

Define $\mu = E[X_i] = 100/n$ and $\sigma^2 = \text{Var}(X_i) = 100/n$. Then we can approximate $\sum_{i=1}^{n} X_i \approx Y \sim \mathcal{N}(n\mu = 100, n\sigma^2 = 100)$. With continuity correction, $P(X \geq 120) \approx P(Y \geq 119.5) \approx 0.0256$.

Example: Clock running time

Say you have a new algorithm and you want to test its running time. You have an idea of the variance of the algorithm's run time: $\sigma^2 = 4\text{sec}^2$ but you want to estimate the mean: $\mu = t$ sec. You can run the algorithm repeatedly (IID trials). How many trials do you have to run so that your estimated runtime $= t \pm 0.5$ with 95% certainty? Let X_i be the run time of the i-th run (for $1 \leq i \leq n$).

$$0.95 = P(-0.5 \leq \frac{\sum_{i=1}^{n} X_i}{n} - t \leq 0.5)$$

By the central limit theorem, the standard normal Z must be equal to:

$$\begin{aligned} Z &= \frac{\left(\sum_{i=1}^{n} X_i\right) - n\mu}{\sigma\sqrt{n}} \\ &= \frac{\left(\sum_{i=1}^{n} X_i\right) - nt}{2\sqrt{n}} \end{aligned}$$

Now we rewrite our probability inequality so that the central term is Z:

$$\begin{aligned} 0.95 = P\left(-0.5 \leq \frac{\sum_{i=1}^{n} X_i}{n} - t \leq 0.5\right) &= P\left(\frac{-0.5\sqrt{n}}{2} \leq \frac{\sum_{i=1}^{n} X_i}{n} - t \leq \frac{0.5\sqrt{n}}{2}\right) \\ &= P\left(\frac{-0.5\sqrt{n}}{2} \leq \frac{\sqrt{n}}{2}\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sqrt{n}}{2}t \leq \frac{0.5\sqrt{n}}{2}\right) = P\left(\frac{-0.5\sqrt{n}}{2} \leq \frac{\sum_{i=1}^{n} X_i}{2\sqrt{n}} - \frac{\sqrt{n}}{\sqrt{n}}\frac{\sqrt{n}t}{2} \leq \frac{0.5\sqrt{n}}{2}\right) \\ &= P\left(\frac{-0.5\sqrt{n}}{2} \leq \frac{\sum_{i=1}^{n} X_i - nt}{2\sqrt{n}} \leq \frac{0.5\sqrt{n}}{2}\right) \\ &= P\left(\frac{-0.5\sqrt{n}}{2} \leq Z \leq \frac{0.5\sqrt{n}}{2}\right) \end{aligned}$$

And now we can find the value of n that makes this equation hold.

$$\begin{aligned} 0.95 &= \Phi\left(\frac{\sqrt{n}}{4}\right) - \Phi\left(-\frac{\sqrt{n}}{4}\right) = \Phi\left(\frac{\sqrt{n}}{4}\right) - \left(1 - \Phi\left(\frac{\sqrt{n}}{4}\right)\right) \\ &= 2\Phi\left(\frac{\sqrt{n}}{4}\right) - 1 \\ 0.975 &= \Phi(\frac{\sqrt{n}}{4}) \\ \Phi^{-1}(0.975) &= \frac{\sqrt{n}}{4} \\ 1.96 &= \frac{\sqrt{n}}{4} \\ n &= 61.4 \end{aligned}$$

Thus it takes 62 runs. If you are interested in how this extends to cases where the variance is unknown, look into variations of the students' t-test.