Will Monroe
CS 109

Lecture Notes #19
August 7, 2017

Central Limit Theorem

Based on a chapter by Chris Piech

The **central limit theorem** says that equally-weighted averages of samples from *any* distribution themselves are normally distributed. Consider the sample mean of IID random variables $X_1, X_2 \ldots$ such that $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$:

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

Mathematically, the central limit theorem states:

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

It is often expressed as a way of obtaining the standard normal, Z:

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

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 IID random variables must also be normal. Let's call the sum of IID random variables $\bar{Y}$:

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

In summary, the central limit theorem explains that both the **average** of IID random variables and the **sum** of IID random variables are normal. This is true *regardless of what distribution the IID variables came from.* It could be normal, uniform, exponential, or a distribution shaped like the San Francisco skyline. It could be continuous or discrete, although in the discrete case, a continuity correction (adding/subtracting 0.5) is needed to get the best results for small n—say, $n < 50$.

Example 1

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: $\sigma^2 = 4$ 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 $1 \leq i \leq n$).

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

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:

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

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

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

Thus, it takes at least 62 runs.

Compare this to if we had used Chebyshev's equality. Remember the sample mean has a mean of t and a variance of σ^2/n. So

$$\begin{aligned}
P(|X - E[X]| > k) &\leq \frac{\text{Var}(X)^2}{k^2} \\
P(|\bar{X} - t| > 0.5) &\leq \frac{4/n}{0.5^2} \leq 0.05 \\
\frac{16}{n} &\leq 0.05 \\
n &\geq 320
\end{aligned}$$

The central limit theorem tells us that even 62 is sufficient; it gives a lower value because we know something about the distribution we are bounding, namely that it is a mean of IID random variables.

Example 2

You roll a 6-sided die 10 times. Let X be the total value of all 10 dice = $X_1 + X_2 + \cdots + X_{10}$. You win the game if $X \leq 25$ or $X \geq 45$. Use the central limit theorem to calculate the probability that you win.

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

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

Normal Approximation to the Poisson

You may have noticed that the Poisson distribution looks rather like a bell-curve. For large values of λ, the Poisson is well-approximated by the normal:

$$\text{Poi}(\lambda) \approx N(\lambda, \lambda) \qquad \text{for large } \lambda$$

One can use a central limit theorem argument to show this, by dividing up the unit of time into many smaller units and adding the number of events in each smaller unit (each of which is an independent Poisson random variable). The below example shows this explicitly.

Example 3

Your website receives $W \sim \text{Poi}(5/3)$ requests in each second. The server crashes if it gets 120 or more requests in a single minute.

Problem: What's the probability that the server crashes in the next minute?

Solution: We could adjust λ to lengthen the unit of time we're using: $X \sim \text{Poi}(\lambda = 60 \cdot 5/3 = 100)$. This is given by the Poisson CDF:

$$P(Y \geq 120) = 1 - \sum_{i=120}^{\infty} \frac{e^{-100}(100)^i}{i!} \approx 0.0282$$

However, let's instead try this with the Central Limit Theorem. We can say that Y is the sum of 60 IID random variables with mean and variance $\mu = \sigma^2 = 5/3$, and therefore we can approximate $X \approx Y \sim N(60 \cdot (5/3), 60 \cdot (5/3))$.

$$\begin{aligned} P(X \geq 120) \approx P(Y > 119.5) &= P(\frac{Y - 100}{10} > \frac{119.5 - 100}{10}) \\ &= P(Z > 1.95) \\ &= 1 - \Phi(1.95) \approx 0.0256 \end{aligned}$$