The Central Limit Theorem

Mike Pence

The Central Limit Theorem

The Central Limit Theorem is one of the most powerful theorems in all of statistics. Let's recap what the theorem says:

Let's say that we have a collection of n independent random variables, $X_1, X_2, ..., X_n$ all with mean $= \mu$ and finite variance σ^2 Define

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

or the average of the random variables.

The Central Limit Theorem states that as $n \to \infty$, not only does $A_n \to \mu$, but also the distribution of A_n converges to a normal distribution that has the same mean and variance as A_n.

What does it mean to converge in distribution? Well, when you are using the CLT, chances are that each individual X_i is not normally distributed, and A_n for finite n is not perfectly normally distributed either. What the CLT is saying is that if we have another random variable,

$$T \sim \mathcal{N}(\mathbf{E}[A_n], Var(A_n))$$

then

$$P(A_n \leq c) \approx P(T \leq c)$$

Remember that we can transform any Normal Distribution into the Standard Normal Distribution (The Normal Distribution with mean 0 and variance 1) subtracting from the mean, and dividing by the square root of the variance. Since we know that $A_n \sim \mathcal{N}(\mathbf{E}[A_n], Var(A_n))$ as $n \to \infty$, we know then that

$$A_n' = \frac{A_n - \mathbf{E}[A_n]}{\sqrt{Var(A_n)}} \sim \mathcal{N}(0, 1)$$

This is useful for us because there is no closed-form cdf for the general Normal Distribution. Using computers, we have calculated emperical values for the CDF of the Standard Normal, which we can use to make calculations.

Before we continue, let's make sure we know what the mean and variance is of A_n. Remember that each of the individual X_i had mean μ and variance σ^2. Using this, and the fact that each of the variables are independent from each other:

$$\mathbf{E}[A_n] = \mathbf{E}[\frac{1}{n}\sum_{i=1}^{n} X_i]$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[X_i], \text{ by linearity of expectation} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\cdot n\mu \\
&= \mu \\
Var(A_n) &= Var(\frac{1}{n}\sum_{i=1}^{n}X_i) \\
&= \frac{1}{n^2}Var(\sum_{i=1}^{n}X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}Var(X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}\cdot n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

Random Variable Transformations

Consider a random variable X, and say that another random variable $Y = cX + k$. For any m, if we wanted to calculate $P(Y = m)$, we can calculate $P(X = \frac{m-k}{2})$ (In the continuous case, we obviously can't calculate $P(Y = m)$, rather, we say that $P(Y \leq m) = P(X \leq \frac{m-k}{2})$). For example, say X is uniformly distributed over the discrete values $\{1,\ 2,\ 3,\ 4\}$, and $Y = 2X + 3$. Clearly, Y is uniformly distributed over $\{5,\ 8,\ 9,\ 12\}$. If we wanted to calculate $P(Y = 12)$, we calculate $P(X = \frac{12-3}{4}) = \frac{1}{4}$. We can verify that we get the same answer if we compute the probability directly.

How is this information useful to us? Well, let's say that we know how some random variable Z is distributed: i.e, we know the cdf $F_z(z)$ of Z. However, we are interested in the cdf of another random variable Y that we don't know: i.e, we want to know $P(Y \leq y), \forall y$. We do know that $Z = cY + k$. How can we calculate $P(Y \leq c)$? We simply calculate $P(Z \leq cy + k) = F_z(cy + k)$.

Using the CLT

Let's put everything toegether and actually use the CLT. Let's say that we have a coin that is biased with probability $p = 0.6$, and we flip the coin 1000 times. Let X be a random variable that is equal to the number heads that we see. We want to calculate the probability that $X \geq 700$, or that we see more than 700 heads. You should recognize that X is Binomially distributed with $p = 0.6$ and $n = 1000$. We can see that

$$
P(X \geq 600) = \sum_{i=600}^{1000} p(X = i) = \sum_{i=600}^{1000}\binom{1000}{i}0.6^i 0.4^{n-i}
$$

However, computing this exact value is computationally impossible. Let's use the CLT to approximate! Define indicator variables X_i as

$$X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ coinflip is a heads} \\ 0 & \text{otherwise} \end{cases}$$

Clearly, each of the X_i are bernoulli random variables with $p = 0.6$. We know that the the mean $\mathbf{E}[X_i] = p = 0.6$, and the variance $Var(X_i) = p(1-p) = 0.24$.

We see that $X = \sum_{i=1}^{n} X_i$. Since X is the sum of independently, identically distributed random variables, by applying the CLT we know that

$$A_n = \frac{X}{n} \sim \mathcal{N}(0.6, \frac{0.24}{1000})$$

We can see that

$$P(X \geq 700) = P(A_n \geq 0.7)$$

How can we calculate $P(A_n \geq 0.7)$? Remember that

$$A_n' = \frac{A_n - 0.6}{\sqrt{\frac{0.24}{1000}}} \sim \mathcal{N}(0, 1)$$

So know we see that

$$P(A_n \geq 0.7) = P(A_n' \geq \frac{0.7 - 0.6}{\sqrt{\frac{0.24}{1000}}}) = P(A_n' \geq 0.645) = 1 - P(A_n' \leq 0.645)$$

Since A_n' is standard normal, we know the cdf $\phi(c)$. So, we simply have to look up the value of $\phi(0.645)$ in our table, which turns out to be ≈ 0.740. So our final answer would be $P(X \geq 700) \approx 1 - 0.740 \approx 0.24$

What if the question was changed slightly: now, we want to find some d such that the probability that we see more than d heads is less than 0.025. Using the same logic as last time, now we're interested in a d such that

$$P(A_n \geq \frac{d}{1000}) \leq 0.025$$

This is equivalent to finding d such that

$$P(A_n \leq \frac{d}{1000}) = P(A_n' \leq \frac{\frac{d}{1000} - 0.6}{\sqrt{\frac{0.24}{1000}}}) \geq 0.975$$

Essentially here, we are looking for some value c such that $P(A_n' \leq c) = 0.975$ - to find this, we need the inverse cdf $\phi^{-1}(p)$ of the normal distribution. It turns out that people have also calculated the tables for the inverse cdf as well, so after looking it up we find that if we want $P(A_n' \leq c) = 0.975$, then $c = \phi^{-1}(0.975) = 1.96$. This means that

$$\frac{\frac{d}{1000} - 0.6}{\sqrt{\frac{0.24}{1000}}} = 1.96$$

Solving for d, we find $d = 904$.

Notice here there is a weakness of CLT - because the CLT approximates the variable as a continuous random variable, we wouldn't be able to approximate the probability of say, there being exactly 700 heads.