Introduction to Central Limit Theorem(CLT)

Today I want to introduce you another important theorem in statistical science: Central Limit Theorem.

The central limit theorem (CLT) is one of the most important results in probability theory. It states that: the sampling distribution of the sample means approaches to normal distribution as the sample size gets larger — no matter what the shape of the population distribution.

The CLT is additionally very useful within the sense that it can simplify our computations significantly. If you've got an issue during which you're interested in a sum of 1 thousand i.i.d. random variables, it will be extremely difficult, if not impossible, to search out the distribution of the sum by direct calculation. Using the CLT we are able to immediately write the distribution, if we all know mean and variance of the $\mathrm{X_i}$'s.

Under certain conditions, the sum of a large number of random variables is approximately normal. Here, I am explaining a version of CLT that applies to i.i.d. random variables.
Let $\mathrm{X_1}, \mathrm{X_2}, \ldots, \mathrm{X_n}$ be i.i.d. random variables with expected value $E\mathrm{X_i} = \mu < \infty$ and variance $0 < Var(\mathrm{X_i}) = \sigma^2 < \infty$. The sample mean $\bar{X}$= $\frac{X_1 + X_1 + \ldots + X_n}{n}$ has mean $E\bar{X}$=μ and variance Var($\bar{X}$) =$\frac{\sigma^2}{n}$. Then, define the normalized random variable to be Z_n

$$\mathrm{Z}_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

The Z_n has mean $E\mathrm{Z}_n$=0 and variance Var(Z_n)=1. Then, the random variable Z_n converges in distribution to the standard normal random variable as n goes to infinity, that is

$$\lim_{n\to\infty} P(\mathrm{Z}_n \le x) = \Phi(x), for\ all\ x \in R$$

where $\Phi(x)$ is the standard normal CDF.

The CLT can be applied no matter what the distribution of the $\mathrm{X_i}$'s is. The $\mathrm{X_i}$'s can be discrete, continuous, or mixed random variables.

In many real applications, a certain random variable of interest is a sum of a large number of independent random variables. In these situations, we are often able to use the CLT to justify using the normal distribution. Another question is that how large n should be so that we can use the normal approximation. The answer generally depends on the distribution of the $\mathrm{X_i}$'s. Usually, if n is larger than or equal to 30, then the normal approximation is very good.

So how we use the CLT to solve problems? General steps are:

1. Write the random variable of interest, Y, as the sum of n i.i.d. random variable $\mathrm{X_i}$'s
2. Find EY and Var(Y) that

$$EY = n\mu, \mathrm{Var}(Y) = n\sigma^2$$

3. $\frac{Y-EY}{\sqrt{VAR(Y)}} = \frac{Y-n\mu}{\sigma\sqrt{n}}$ is approximately standard normal. Thus, to find $P(y_1 \leq Y \leq y_2)$

We have:

$$P(y_1 \leq Y \leq y_2) \approx \phi(\frac{y_2-n\mu}{\sigma\sqrt{n}}) - \phi(\frac{y_1-n\mu}{\sigma\sqrt{n}})$$

Above all are my introduction to Central Limit Theorem, hope you learn from it and find it useful when solving problems.

Reference:

https://www.probabilitycourse.com/chapter7/7_1_2_central_limit_theorem.php