Central Limit Theorem

Michael Filletti

March 2019

The origins of the Central Limit Theorem (CLT) could be traced back to 1733, when the French mathematician Abraham de Moivre proposed a very early and rough version of the theorem. Pierre-Simon Laplace began proving this theorem in 1810, however, the necessary and sufficient conditions were formulated in the late 19th century in Russia, by the mathematicians Chebyshev, Markov and Liapounov. The theorem was titled the Central Limit Theorem by Georg Pólya in 1920, when the theorem was in its more final state. The Belgian social scientist Adolphe Quetelet provided one of the earliest observations of the vastness of the applicability of the normal distribution, and its usefulness. In one of his experiments he measured the chests of 5738 Scottish soldiers, and plotted the results on a histogram, finding that it was more normal. In a later study, he recorded the heights of 100,000 recruits, plotting them on a histogram at intervals of 1 inch. Once again, he found that the data followed the normal distribution, except for three intervals around the 62 inch mark. While attempting to understand why this was the case, he was able to find out that various recruits were bending their knees so as to not be eligible for the minimum height of 62 inches required to join the army.

Theorem. *The Central Limit Theorem states that given a sequence $(X_i)_{i\in\mathbb{N}}$ of independent identically distributed random variables with mean μ and variance σ^2, then the sequence of distributions corresponding to the sequence of random variables $\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)_{n\in\mathbb{N}}$ converges to the standard normal distribution $\mathcal{N}(0,1)$.*

The CLT essentially claims that as $n \to \infty$ the random variable $\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)$ approximates a standard normal distribution (i.e. a normal distribution with mean 0 and standard deviation 1). There are many versions of this theorem, and the random variables can be transformed from $\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)_{n\in\mathbb{N}}$ to $(\bar{X})_{n\in\mathbb{N}}$, which represents the sample means. We can say that the sample means $(\bar{X})_{n\in\mathbb{N}}$ approach the normal distribution $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ as $n \to \infty$. It is one of the most fundamental theorems in all statistics, and is used to prove an endless number of theorems, not just within the field of statistics. The elegance of this theorem also comes from the fact that the normal distribution is one of the more easy distributions to deal with, and statisticians understand it very well. Similar to Quetelet, it can allow us to observe any strange goings on within our data.

In practice, a large enough sample size for the CLT to take effect occurs when $n > 29$. This theory can be used to find confidence intervals for various characteristics of the population. Consider the example below:

Example. *A sample of size 100 taken from a population of endemic bandicoots gives a mean weight of 879.4g. Suppose (unlikely as it is) we know that the variance of the weights is 9.2. From this information we can find the confidence interval, with probability 0.95:*

$$\begin{aligned}
-1.96 \leq & \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96 \\
-1.96 \leq & \frac{879.4-\mu}{\frac{9.2}{\sqrt{100}}} \leq 1.96 \\
-1.96 \cdot \frac{9.2}{\sqrt{100}} - 879.4 \leq & -\mu \leq 1.96 \cdot \frac{9.2}{\sqrt{100}} - 879.4 \\
879.4 - 1.96 \cdot \frac{9.2}{\sqrt{100}} \leq & \mu \leq 1.96 \cdot \frac{9.2}{\sqrt{100}} + 879.4 \\
877.6 \leq & \mu \leq 881.2
\end{aligned}$$