CHAPTER 1

The Central Limit Theorem

Central Limit Theorem states that the sum of a large number of independent random variables has a distribution that is approximately normal. Hence, it not only provides a simple method for computing approximate probabilities for sums of independent random variables, but also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit bell-shaped (that is, normal) curves.

THEOREM 1. ***(Central Limit Theorem)*** *Let $X_1, X_2, ..., X_i, ...$ be a sequence of independent random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$, and the common distribution function F and moment-generating function M defined in a neighborhood of zero. Let $Z_n = \frac{\sum\limits_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$, then the limiting distribution of Z_n is the standard normal*

PROOF. The limiting distribution of Z_n is the standard normal, written $Z_n \to_d Z \sim N(0,1)$ as $n \to \infty$, if

$$\lim_{n\to\infty} M_{Z_n}(t) = M_Z(t) = e^{\frac{t^2}{2}}$$

Moment-generating function of Z_n is,

$$\begin{aligned} M_{Z_n}(t) &= E\left(e^{Z_n t}\right) \\ &= E\left(e^{\left(\frac{\sum\limits_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}\right) t} \right) \\ &= \left(E\left(e^{\frac{t}{\sqrt{n}\sigma}(X-\mu)}\right)\right)^n \\ &= \left(M_{X-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right)\right)^n \end{aligned}$$

$m(t) = M_{X-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right)$ has a Taylor series expansion about zero:

$$m(t) = m(0) + m'(0)\,t + \frac{m''(0)\,t^2}{2} + \varepsilon_s$$

where $\frac{\varepsilon_s}{s^2} \to 0$ as $s \to 0$. Since $m(0) = 1$, $m'(0) = 0$, and $m''(0) = \sigma^2$. As $n \to \infty$ and $\frac{t}{\sqrt{n}\sigma} \to 0$ and

$$M_{X-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right) = 1 + \frac{1}{2}\sigma^2\left(\frac{t}{\sqrt{n}\sigma}\right)^2 + \varepsilon_n$$

where $\frac{\varepsilon_n}{\frac{t^2}{n\sigma^2}} \to 0$ as $n \to \infty$. Thus,

$$M_{Z_n}(t) = \left(M_{X-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right)\right)^n = \left(1 + \frac{t^2}{2n} + \varepsilon_n\right)^n$$

It can be shown that if $a_n \to a$, then $\lim\limits_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = e^a$. Therefore,

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty}\left(1 + \frac{t^2}{2n} + \varepsilon_n\right)^n = e^{\frac{t^2}{2}}$$

□

The above Theorem said that

$$\frac{\sum\limits_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \to_d Z \sim N(0,1) \Longleftrightarrow \frac{\frac{\sum\limits_{i=1}^{n} X_i}{n} - \frac{n\mu}{n}}{\frac{\sqrt{n}\sigma}{n}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \to_d Z \sim N(0,1) \Longleftrightarrow \sum_{i=1}^{n} X_i \to_d Y \sim N\left(n\mu, n\sigma^2\right)$$

And for $n = 1$, then $\frac{X-\mu}{\sigma} \to_d Z \sim N(0,1)$.

Consider here some applications of the central limit theorem.

Example 2. An astronomer is interested in measuring the distance, in light-years, from his observatory to a distant star. Although the astronomer has a measuring technique, he knows that, because of changing atmospheric conditions and normal error, each time a measurement is made it will not yield the exact distance, but merely an estimate. As a result, the astronomer plans to make a series of measurements and then use the average value of these measurements as his estimated value of the actual distance. If the astronomer believes that the values of the measurements are independent and identically distributed random variables having a common mean d (the actual distance) and a common variance of 4 (light-years), how many measurements need he make to be reasonably sure that his estimated distance is accurate to within ±0.5 light- year?

Solution. Suppose that the astronomer decides to make n observations. If $X_1, X_2, ..., X_n$ are the n measurements, then, from the central limit theorem, it follows that $Z_n = \frac{\sum\limits_{i=1}^{n} X_i - nd}{2\sqrt{n}}$ has approximately a standard normal distribution. Hence,

$$P\left\{-0.5 \le \frac{\sum\limits_{i=1}^{n} X_i}{n} - d \le 0.5\right\} = P\left\{-0.5\frac{\sqrt{n}}{2} \le Z_n \le 0.5\frac{\sqrt{n}}{2}\right\} \approx \Phi\left(\frac{\sqrt{n}}{4}\right) - \Phi\left(-\frac{\sqrt{n}}{4}\right) = 1 - 2\Phi\left(\frac{\sqrt{n}}{4}\right)$$

Therefore, if the astronomer wants, for instance, to be 95 percent certain that his estimated value is accurate to within .5 light year, $1 - 2\Phi\left(\frac{\sqrt{n}}{2}\right) = 0.95$. Thus, from table normal distribution $\frac{\sqrt{n}}{4} = 1.96$. As n is not integral valued, he should make 62 observations.

Example 3. (Normal Approximation to the Binomial Distribution) The probability that a basketball player hits a shot is p = 0.5. If he takes 20 shots, what is the probability that he hits at least nine)?

Solution. The exact probability is

$$\begin{aligned} P\left(Y_{20} \geq 9\right) &= 1 - P\left(Y_{20} \leq 8\right) \\ &= 1 - \sum_{y=0}^{8} \left(\begin{array}{c} 20 \\ y \end{array}\right) 0.5^y 0.5^{20-y} \\ &= 0.7483 \end{aligned}$$

Since a binomial random variable is the sum of independent Bernoulli random variables, its distribution can be approximated by a normal distribution (Example **??**). The approximation is best when the binomial distribution is symmetric, that is, when $p = \frac{1}{2}$. A frequently used rule of thumb is that the approximation is reasonable when $np > 5$ and $n\left(1-p\right) > 5$. The approximation is especially useful for large values of n, for which tables are not readily available. A normal approximation is

$$\begin{aligned} P\left(Y_{20} \geq 9\right) &= 1 - P\left(Y_{20} \leq 8\right) \\ &= 1 - \Phi\left(\tfrac{8-20(0.5)}{\sqrt{20(0.5)(0.5)}}\right) \\ &= 1 - \Phi\left(\tfrac{8-10}{\sqrt{5}}\right) \\ &= 1 - \Phi\left(-0.89\right) \\ &= 0.8133 \end{aligned}$$

Because the binomial distribution is discrete and the normal distribution is continuous, the approximation can be improved by making a continuity correction. In particular, each binomial probability, $b\left(y; n, p\right)$ has the same value as the area of a rectangle of height $b\left(y; n, p\right)$ and with the interval $\left[y-0.5, y+0.5\right]$ as its base, because the length of the base is one unit. The area of this rectangle can be approximated by the area under the pdf of $Y \sim N\left(np, npq\right)$. In general, if $Y_n \sim BIN\left(n, p\right)$ and $a \leq b$ are integer, then $P\left(a \leq Y_n \leq b\right) = \Phi\left(\frac{b+0.5-np}{\sqrt{npq}}\right) - \Phi\left(\frac{a-0.5-np}{\sqrt{npq}}\right)$.

Example 4. Let $X_1, X_2, ..., X_n$ be a random sample from a uniform distribution, $X_i \sim UNIF\left(0,1\right)$ and let $Y_n = \sum_{i=1}^{n} X_i$. Find the limiting distribution of Y_n.

Solution. Since $E\left(X_i\right) = \frac{1}{2}$ and $Var\left(X_i\right) = \frac{1}{12}$ and from Theorem 1, then $Y_n = \sum_{i=1}^{n} X_i \to_d Z \sim N\left(n\mu, n\sigma^2\right) = N\left(\frac{n}{2}, \frac{n}{12}\right)$. And for example $n = 12$, then $Y_{12} - 6 \to_d Z \sim N\left(0,1\right)$.

Example 5. Suppose that the number of insurance claims, N, filed in a year is Poisson distributed with $E(N) = 10,000$. Use the normal approximation to the Poisson to approximate $P(N > 10,200)$.

Solution.