Lesson 2.2

The Central Limit Theorem

Introduction

If $Y_1, Y_2, \cdots, Y_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution, then we know that $\overline{Y} \sim N(\mu, \sigma^2/n)$. What would be the distribution of $\overline{Y}$ if the data do not come from a normal distribution?

Central Limit Theorem (CLT)

Suppose that $Y_1, Y_2, \cdots, Y_n$ is a random sample from a (population) distribution with $E(Y) = \mu$ and $V(Y) = \sigma^2 < \infty$. Let $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ denote the sample mean and define

$$U_n = \sqrt{n}\left(\frac{\overline{Y} - \mu}{\sigma}\right) = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$$

Then, as $n \to \infty$, the cumulative distribution function of U_n converges pointwise to the cumulative distribution function of a $N(0, 1)$ random variable.

Remarks:

1. We write $U_n \xrightarrow{d} N(0, 1)$. The symbol "$\xrightarrow{d}$" is read as "converges in distribution to." The mathematical statement that

$$U_n = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1)$$

implies that, for large n, $\overline{Y}$ has an **approximate** normal sampling distribution with mean μ and variance σ^2/n. Thus, it is common to write

$$\overline{Y} \sim AN(\mu, \sigma^2/n)$$

2. The CLT states that averages will be **approximately** normally distributed even if the underlying population distribution is not.

3. The approximation depends on

   a) *sample size*: The larger the sample size, the better the approximation

   b) *symmetry* of the population distribution: The more symmetric the population distribution, the better the approximation.

4. If the population distibution is highly skewed we need a larger sample size for the CLT to kick in.

Before we propve the CLT, let us recall the following results from calculus.

Lemma:

For all $a \in \mathbb{R}$,

$$\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = e^a.$$

A slight variant of the above result states that if $a_n \to a$, as $n \to \infty$, then

$$\lim_{n\to\infty}\left(1+\frac{a_n}{n}\right)^n = e^a.$$

PROOF:

Let $Y_1, Y_2, \cdots, Y_n$ is a random sample from a (population) distribution with $E(Y) = \mu$ and $V(Y) = \sigma^2 < \infty$.

Define

$$Z_i = \frac{Y_i - \mu}{\sigma}$$

and let $m_Z(t)$ denote the common MGF of each $Z_1, Z_2, \cdots, Z_n$. This MGF $m_Z(t)$ exists for all $t \in (-\sigma h, \sigma h)$. Now,

$$U_n = \sqrt{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i.$$

Thus, the MGF of U_n is given by

$$
\begin{aligned}
m_{U_n}(t) &= E(e^{tU_n}) \\
&= E\left[e^{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i} \right] \\
&= E\left[e^{(t/\sqrt{n})Z_1} \times e^{(t/\sqrt{n})Z_2} \times \cdots \times e^{(t/\sqrt{n})Z_n} \right] \\
&= E\left[e^{(t/\sqrt{n})Z_1} \right] \times E\left[e^{(t/\sqrt{n})Z_2} \right] \times \cdots \times E\left[e^{(t/\sqrt{n})Z_n} \right] \\
&= \left[m_Z(t/\sqrt{n}) \right]^n
\end{aligned}
$$

Now, consider the McLaurin series expansion (the Taylor series expansion about 0) of $m_Z(t/\sqrt{n})$, we have

$$
m_Z(t/\sqrt{n}) = \sum_{k=0}^{\infty} m_Z^{(k)}(0) \frac{(t/\sqrt{n})^k}{k!}
$$

where: $m_Z^{(k)}(0)$ is the k^{th} derivative of $m_Z(t)$ evaluated at zero. Since each Z_i has mean 0 and variance 1, then

$$
m_Z^{(0)}(0) = 1 m_Z^{(1)}(0) = 0 m_Z^{(2)}(0) = 1
$$

Thus,

$$
\begin{aligned}
m_Z(t/\sqrt{n}) &= \sum_{k=0}^{\infty} m_Z^{(k)}(0) \frac{(t/\sqrt{n})^k}{k!} \\
&= 1 + \frac{(t/\sqrt{n})^2}{2!} + R_Z(t/\sqrt{n})
\end{aligned}
$$

where $R_Z(t/\sqrt{n}$ is the remainder term in hte expansion, that is,

$$
R_Z(t/\sqrt{n}) = \sum_{k=3}^{\infty} m_Z^{(k)}(0) \frac{(t/\sqrt{n})^k}{k!}
$$

The key to finishing the proof is recognizing that

$$
\lim_{n \to \infty} n R_Z(t/\sqrt{n}) = 0
$$

This is not difficult to see since the $k = 3$ term in $R_Z(t/\sqrt{n})$ contains an $n\sqrt{n}$ in the denominator; the $k = 4$ term contains an n^2 in its denominator, and so on.

Now, for any fixed t, we can write

$$\begin{aligned}
\lim_{n\to\infty} m_{U_n}(t) &= \lim_{n\to\infty} \left[m_Z(t/\sqrt{n})\right]^n \\
&= \lim_{n\to\infty} \left[1 + \frac{(t/\sqrt{n})^2}{2!} + R_Z(t/\sqrt{n})\right]^n \\
&= \lim_{n\to\infty} \left[1 + \frac{1}{n}\left(\frac{t^2}{2} + nR_Z(t/\sqrt{n}\right)\right]^n \\
&= \lim_{n\to\infty} \left[1 + \frac{1}{n}(t^2/2)\right] \\
&= e^{t^2/2}, \text{ based on the Lemma}
\end{aligned}$$

Thus, as $n \to \infty$ we have shown that $\lim_{n\to\infty} m_{U_n}(t) = e^{t^2/2}$, which is the MGF of the $N(0,1)$ distribution. **QED**

Example 2.2.1

A chemist is studying the degradation behavior of vitamin B_6 in a multiviatmin. The chemist selects a random sample of 36 multivitamin tablets and for each tablet counts the number of days until the B_6 content falls below FDA requirement. Let $Y_1, Y_2, \cdots, Y_{36}$ denote the measurements for the 36 tablets, and assume that $Y_1, Y_2, \cdots, Y_{36}$ is an iid sample from a Poisson distribution with mean 50.

a. What is the approximate probability that teh average number of days $\overline{Y}$ will exceed 52?

b. How many tablets does the chemist need to observe so that $P(\overline{Y} < 49.5) \approx 0.01$?

SOLUTION: Left as a classroom exercise!