6.1: The Central Limit Theorem

Taking the average [mean] of a sample of quantitative data is actually a very nice process: the arithmetic is simple, and the average often has the nice property of being closer to the center of the data than the values themselves being combined or averaged. This is because while a random sample may have randomly picked a few particularly large (or particularly small) values from the data, it probably also picked some other small (or large) values, so that the mean will be in the middle. It turns out that these general observations of how nice a sample mean can be explained and formalized in a very important Theorem:

FACT 6.1.1. **The Central Limit Theorem** [**CLT**] Suppose we have a large population on which is defined a quantitative random variable X whose population mean is μ_X and whose population standard deviation is σ_X. Fix a whole number $n \geq 30$. As we take repeated, independent SRSs of size n, the distribution of the sample means $\overline{x}$ of these SRSs is approximately $N(\mu_X, \sigma_X/\sqrt{n})$. That is, the distribution of $\overline{x}$ is approximately Normal with mean μ_X and standard deviation $\sigma_X/\sqrt{n}$.

Furthermore, as n gets bigger, the Normal approximation gets better.

Note that the CLT has several nice pieces. First, it tells us that the middle of the histogram of sample means, as we get repeated independent samples, is the same as the mean of the original population – *the mean of the sample means is the population mean*. We might write this as $\sigma_{\bar{X}} = \mu_X$.

Second, the CLT tells us precisely how much less variation there is in the sample means because of the process noted above whereby averages are closer to the middle of some data than are the data values themselves. The formula is $\sigma_{\bar{X}} = \sigma_x/\sqrt{n}$.

Finally and most amazingly, the CLT actually tells us exactly what is the shape of the distribution for $\overline{x}$ – and it turns out to be that complicated formula we gave Definition 4.3.19. This is completely unexpected, but somehow the universe knows that formula for the Normal distribution density function and makes it appear when we construct the histogram of sample means.

Here is an example of how we use the CLT:

EXAMPLE 6.1.2. We have said elsewhere that adult American males' heights in inches are distributed like $N(69, 2.8)$. Supposing this is true, let us figure out what is the probability that 52 randomly chosen adult American men, lying down in a row with each one's feet touching the next one's head, stretch the length of a football field. [Why 52? Well, an American football team may have up to 53 people on its active roster, and one of them has to remain standing to supervise everyone else's formation lying on the field....]

First of all, notice that a football field is 100 yards long, which is 300 feet or 3600 inches. If every single one of our randomly chosen men was exactly the average height for adult men, that would a total of $52 * 69 = 3588$ inches, so they would not stretch the whole length. But there is variation of the heights, so maybe it will happen sometimes....

So imagine we have chosen 52 random adult American men. Measure each of their heights, and call those numbers $x_1, x_2, \ldots, x_{52}$. What we are trying to figure out is whether $\sum x_i \geq 3600$. More precisely, we want to know

$$P\left(\sum x_i \geq 3600\right) . \tag{6.1.1}$$

Nothing in that looks familiar, but remember that the 52 adult men were chosen randomly. The best way to choose some number, call it $n = 52$, of individuals from a population is to choose an SRS of size n.

Let's also assume that we did that here. Now, having an SRS, we know from the CLT that the sample mean $\overline{x}$ is $N(69, 2.8/\sqrt{52})$ or, doing the arithmetic, $N(69, .38829)$.

But the question we are considering here doesn't mention $\overline{x}$, you cry! Well, it almost does: $\overline{x}$ is the sample mean given by

$$\overline{x} = \frac{\sum x_i}{n} = \frac{\sum x_i}{52} . \tag{6.1.2}$$

What that means is that the inequality

$$\sum x_i \geq 3600 \tag{6.1.3}$$

amounts to exactly the same thing, by dividing both sides by 52, as the inequality

$$\frac{\sum x_i}{52} \geq \frac{3600}{52} \tag{6.1.4}$$

or, in other words,

$$\overline{x} \geq 69.23077\ . \tag{6.1.5}$$

Since these inequalities all amount to the same thing, they have the same probabilities, so

$$P\left(\sum x_i \geq 3600\right) = P\left(\overline{x} \geq 69.23077\right)\ . \tag{6.1.6}$$

But remember $\overline{x}$ was $N(69, .38829)$, so we can calculate this probability with **LibreOffice Calc** or **Microsoft Excel** as

$$\begin{aligned} P\left(\overline{x} \geq 69.23077\right) &= 1 - P\left(\overline{x} < 69.23077\right) \\ &= \text{NORM.DIST}(69.23077, 69, .38829, 1) \\ &= .72385 \end{aligned}$$

where here we first use the probability rule for complements to turn around the inequality into the direction that NORM.DIST calculates.

Thus the chance that 52 randomly chosen adult men, lying in one long column, are as long as a football field, is 72.385%.
