Math 316

Reminder: the distribution consists of the outcomes and their frequency or relative frequency or probability. **Figure 3.4.1**.

Section 5.1 Basic Ideas

We typically take a sample and compute its mean and standard variation in order to learn about the population, but without having to get data from the entire population, which would usually be difficult or often impossible.

Let's consider an example we'll see in Section 5.2. Suppose that we want to know (1) the average weight of all Princess Bean seeds and (2) the variability (variance or standard deviation) of those weights. We will use the *sample* mean $\bar{y}$ and standard deviation s as estimates for the *population* mean μ and standard deviation σ. (We're usually more interested in the mean.) Consider the sample shown in **Figure 5.2.6b**. If this were the only information we had about the population (that is, if we didn't actually know the population's mean or standard deviation), then our best guess is that the population mean (the average weight of Princess Bean seeds) is $\mu = \bar{y} = 526.1$. Of course we wouldn't necessarily expect that $\mu = 526.1$ exactly, but we would expect that μ is close to that value. Eventually, based on information from the sample, we will end up making a statement like "we are 95% confident that $\mu = 526.1 \pm 46.9$." So where does this margin of error 46.9 come from? (This 95% level of confidence is arbitrary, but pretty standard.) The margin of error 46.9 depends on three things:

1. How confident we want to be: more confidence $\Rightarrow$ larger margin of error.
2. How many seeds in our original sample: larger sample $\Rightarrow$ smaller margin of error.
3. How much variability was in the sample: less variability (smaller standard deviation) within the sample $\Rightarrow$ smaller margin of error.

Section 5.2 The Sample Mean

Given a sample, we want to make an estimate about the population.** To move toward that goal, for now we'll do the *opposite*: given a population (in particular, its mean and standard deviation), we will estimate what a sample from that population would look like.

So what would a single sample look like? Answer: like a miniature version of the population itself—more or less—since there is always some randomness in each sample. See the **nine samples on pages 156 – 157**. Each has essentially the same distribution (shape) as the population with approximately the same mean and standard deviation as the population, but some samples are more different than the population than others.

This is a good time to remind you about the issue of "do I divide by n or by $n-1$?" when computing standard deviation. If your data are from the *entire population*, then you divide by n, while if your data are from a *sample* which you are using to estimate the population parameters, then you divide by $n-1$. Since we are usually just working with samples (rather than the entire population), the only formula the book gives for standard deviation is the one dividing by $n-1$. See the **box on page 60, and the **note about "Why $n-1$" on page 62**.

Related to this is a second issue: what would the sampling distribution be? The sampling distribution is the distribution of the sample means. That is: we take a bunch of samples, we compute the mean of each sample, and these sample means are now our data. See **Figure 5.2.1**. Consider the **nine samples on p. 156 - 157**. The nine sample means are

526 481 538 445 502 461 488 518 514

What would you guess the mean of the nine sample means is? It is 497, which is pretty close to what you probably guessed: the population mean 500. The standard deviation of these sample means is 41. (In a minute, we'll discover how this *sample means standard deviation* $\sigma_{\bar{Y}} = 29$ relates to the *population standard deviation* $\sigma = 120$.) Some important intuition:

Larger samples $\Rightarrow$ Each sample mean is closer to population mean
$\Rightarrow$ Less variation in samples means
$\Rightarrow$ That is, smaller $\sigma_{\bar{Y}}$.

Suppose we could take more than just these nine samples, say hundreds or thousands of (or ideally *all* possible) samples from this population. This is referred to as a meta study.** See **Figure 5.1.1**. As **Figure 5.2.1** illustrates, we are interested in the means of these samples. So we're interested in the mean and the standard deviation of the sampling distribution (that is, the mean and standard deviation of all of these sample means). Let's experiment a bit with the **Central Limit Theorem simulator online** (via the class homepage) a bit to help us answer these two questions.

We are seeing (and it can be proven) that where μ and σ are the mean and standard deviation of the population (the original collection of data), then the mean $\mu_{\bar{Y}}$ and standard deviation $\sigma_{\bar{Y}}$ of the sampling distribution (the $\bar{Y}$ denotes sample mean) are

$$\mu_{\bar{Y}} = \mu \quad \text{and} \quad \sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}.$$

See **Theorem 5.2.1** and **Table 5.2.2**. (Why were the mean and standard deviation of each of the nine sample means *not* exactly $\mu_{\bar{Y}} = 500$ and $\sigma_{\bar{Y}} = 120/\sqrt{25} = 24$?)

Finally, and very importantly, *regardless of the original distribution, the sampling distribution will be approximately normal.* This is really significant, and the main reason we care about the normal distribution so much (it's not because most real life data are normally distributed—most data are not): the larger the sample size n is, the more normally distributed the sample means will be. See **Examples 5.3.1** and **5.3.2**. Let's experiment more with the **Central Limit Theorem simulator** with data that is not normally distributed. Let's work some class examples based on **Example 5.2.2/3**.

**For most populations we cannot actually take every possible sample, since there are simply too many, essentially infinitely many. Book Example 5.1.3 gives an example in which every possible sample is actually listed, thus it is possible to find the complete sampling distribution, which is given in Table 5.1.2 and Figure 5.1.2. In that example, there are 3 possibilities for each woman, so for three women there are $3^3 = 27$ possible outcomes, which are listed in Tables 5.1.1 and 5.1.2 and Figure 5.1.2. Also note: I've seen the word "meta" or phrase "meta study" refer to analyzing a bunch of different studies on the same thing.

Math 316

Section 5.4 Normal Approximation to the Binomial Distribution

It turns out that when n, the number of binomial trials (for example, the number of free throws being shot) is large, the binomial distribution is nicely approximated by the normal distribution. See **Online Probability distribution of coin flips** (Sections 3.4 – 3.5). Recall that a normal distribution is characterized by its mean μ and its standard deviation σ. Also recall the formulae for those given in Section 3.6:

$$\mu = np \qquad\qquad \sigma = \sqrt{np(1-p)}$$

So let's work **Class Example 1** using the normal distribution approximation. Notice the other way of thinking of the problem: with proportions rather than amounts. This way of thinking is what they are talking about **in (b) in the box** on page 163. **Remark 2** on page 164 is a little confusing. Another way to see things is to simply divide both by the sample size n, as I just did in working Example 1 in class. **See Figure 5.4.1a,b**.

Finally, what about the beginning point of 60 shots and the end point of 80 shots? We should have done what is called a continuity correction. One last look at **Class Example 1** with that in mind.

Note that the larger n is (that is, the more binomial *trials*), the better the normal distribution approximates the binomial distribution. See **How Large Must n be?** at end of Section 5.4 on page 167**.**