Math 316

Section 6.4 Planning a Study to Estimate μ

In 6.3 we found sample mean $\bar{y}$ and standard deviation s (and standard error $\frac{s}{\sqrt{n}}$) and use them to estimate the population mean $\mu = \bar{y} \pm t\frac{s}{\sqrt{n}}$, where t is determined by the desired confidence level and sample size n (Table 4). What if you want to have a certain interval width? That is, how can we control the size of $t\frac{s}{\sqrt{n}}$? **Example 6.4.1**.

So we can see that we can estimate (usually based on past experience) how large a sample to take, as a function of the estimated sample (or population) standard deviation and t (which corresponds to what level of confidence we want).

Section 6.5 Conditions for Validity of Estimation Methods

Look at the **Summary of Conditions** on **Page 202**. The issue of a truly random sample (which means that the sample truly accurately reflects the entire population) is important. Suppose you were conduct a study about what percentage of the American population pays their bills online, and suppose you were to conduct your study by randomly selecting 500 Internet users. Does anyone see a problem with this?

Section 6.6 Comparing Two Means

Recall that in Section 6.3 we found sample mean $\bar{y}$ and standard deviation s (and standard error $\frac{s}{\sqrt{n}}$) to estimate the population mean

$$\mu = \bar{y} \pm t \cdot SE_{\bar{Y}} \quad \text{where} \quad SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}.$$

Later in 6.7 we will do the same thing, but for the *difference* between two populations:

$$\mu_1 - \mu_2 = (\bar{y}_1 - \bar{y}_2) \pm t \cdot SE_{\bar{Y}_1 - \bar{Y}_2}$$

For now, in 6.6 we learn about standard error if we are dealing with samples from *two* populations. We don't actually do anything with this standard error until Section 6.7.

The <u>standard error</u> of the *difference between the sample means*, as given on Page 201, is:

$$SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Compare to $SE_{\bar{Y}}$ above. Note that $SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{SE_1^2 + SE_2^2}$. Since $SE^2_{\bar{Y}_1 - \bar{Y}_2} = SE_1^2 + SE_2^2$, then $SE_1,\ SE_2 < SE_{\bar{Y}_1 - \bar{Y}_2} < SE_1 + SE_2$, as illustrated in **Figure 6.6.2**.

Remember that the standard error tells us how good/useful the information from our sample—such as the sample mean—is as an estimate for the population. As we've seen before, there are two things that will make SE *smaller*: ☺

1. Less variation in the samples; that is, smaller s_1 and/or s_2.
2. Larger sample sizes n_1 and/or n_2.

And SE *smaller* results in a narrower confidence interval (a more precise estimate)

$$\mu_1 - \mu_2 = (\bar{y}_1 - \bar{y}_2) \pm t \cdot SE_{\bar{Y}_1 - \bar{Y}_2} \text{ . } ☺$$

Example 6.6.2. All of the problems in Section 6.6 are pretty much like this, except that some of them are more dressed up with different words.

A second way to compute SE is the Pooled Standard Error, given on **Page 209**. The SE given on **Page 207** is sometimes called the unpooled SE. (In the Pooled Standard Error on Page 209, there is a typo in their s^2 definition: there should *not* be a bar over the y_i term—see **Page 60** for the typo-free definition.) Since $s^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$, then $(n-1)s^2 = \sum(y_i - \bar{y})^2$. So in finding s^2_{pooled}, we add up *all* of the squares of the differences $(y_i - \bar{y})^2$ from *both* samples *together*, as if they were one giant sample, then divide them by their combined sample sizes -1 (twice): $n_1 - 1 + n_2 - 1$. As seen in the s^2_{pooled} formula, we can view s^2_{pooled} as a weighted average of the two sample standard variances. (If time, I'll show this in class.) We see this in **Example 6.6.4**. Notice that 0.1628 is a (weighted) average of .1892 and .1232, but it closer to .1892 since sample 1 is larger than sample 2, so its variance is weighted more in computing the pooled variance.

Notice that we can write

$$SE_{pooled} = \sqrt{s^2_{pooled}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{s^2_{pooled}}{n_1} + \frac{s^2_{pooled}}{n_2}}.$$

Compare this to the (non-pooled)

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

In either case, *larger samples* ☺ and/or *less variation in the samples* ☺ result in smaller SE, ☺ which gives us a more precise estimate for the population, ☺ that is, a narrower confidence interval. ☺

At the bottom of page 209 they point out that if $n_1 = n_2$ or $s_1 = s_2$ (or both), then the pooled and non-pooled SE are equal. (If time, I'll show you in class.)

See the short discussion starting at the top of **Page 210** on pooled vs. non-pooled SE.

For now we'll use the non-pooled $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. Later we'll use SE_{pooled}.