Math 316

Section 6.1 Statistical Estimation

Reminder: one of our big goals in statistics is to learn more about an entire population by looking at a sample from that population. Recall the notation for sample and population, on the **middle of page 177**. (We also saw this in Table 5.2.2 on page 156.)

Section 6.2 Standard Error of the Mean

Recall that each sample from the same population will generally be different. See **Examples 5.2.4/5**. Suppose you didn't know the actual population mean, but all you had was a single sample. How much would you trust the mean of that one sample as an estimate for the population mean? Of course we don't expect the sample mean to be exactly the population mean, so the best we could say is:

- The population mean is approximately the sample mean; that is,
- The population mean is the sample mean plus or minus some amount of error.

The standard deviation s of a sample tells us how dispersed the data are (how much variation there is in the data, how "spread out" the data are). The <u>standard error</u> $SE_{\bar{Y}} = s/\sqrt{n}$ of a sample tells us how reliable the sample statistics $\bar{y}$ and s are as estimates for the population parameters μ and σ. A smaller standard error ☺ means that we are more confident about our estimate, that is, the sample statistics are a more reliable estimate for the population parameters. Two things make the SE *smaller*: ☺

Smaller SD s

Larger sample size n

See **Example 6.2.2**.

Every sample will have some variation. The SD s (and mean $\bar{y}$) of the sample will be approximately equal to the SD σ (and mean μ) of the population. See **Example 5.2.5**. This makes sense: each sample is sort of "miniature" version of the population. Notice the statistics (the numbers) of the samples and the distribution (the shapes) of the samples and how they are about the same as the population itself. So $s \approx \sigma$, no matter how larger the sample is. So larger samples will have approximately the same SD as smaller samples. But the mean of a *larger* sample should be a more reliable estimate for the population mean. This is reflected in a smaller standard error.

Section 6.3 Confidence Interval for μ

The formula we'll see over and over today is

$$\bar{y} - t\frac{s}{\sqrt{n}} < \ \mu \ < \bar{y} + t\frac{s}{\sqrt{n}} \qquad i.e. \qquad \mu = \bar{y} \pm \ t\frac{s}{\sqrt{n}}$$

where t comes from **Table 4** (back of the book). For example, $t_{0.025}$ corresponds to a confidence interval of 95%: an area of .95 within the interval means there is area of 0.025 in each of the two tails outside the interval.

Consider the example below, in which we have a sample with $\bar{y} = 70, s = 4, n = 25$ and confidence level of 95%. This is the first line in the table below. Next, let's look at some variations of this example. I'll round my numbers to make things a little simpler.

	Sample mean $\bar{y}$	Sample SD s	Sample size n	Degs. of freedom $df = n - 1$	Conf. level	t-value from Table 4	$t\frac{s}{\sqrt{n}}$	Interval for our estimate for the population mean μ	Narrower/wider interval than $68.3 < \mu < 71.7$?
A	70	4	25	24	95%	2.064	1.7	$68.3 < \mu < 71.7$	–
B	70	**2**	25	24	95%	2.064	**.8**	$69.2 < \mu < 70.8$	Narrower ☺
C	70	**8**	25	24	95%	2.064	**3.3**	$66.7 < \mu < 73.3$	Wider ☹
D	70	4	**12**	11	95%	**2.201**	**2.5**	$67.5 < \mu < 72.5$	Wider ☹
E	70	4	**50**	49	95%	**2.009**	**1.1**	$68.9 < \mu < 71.1$	Narrower ☺
F	70	4	25	24	**90%**	**1.711**	**1.4**	$68.6 < \mu < 71.4$	Narrower ☺
G	70	4	25	24	**99%**	**2.797**	**2.2**	$67.8 < \mu < 72.7$	Wider ☹
H	**60**	4	25	24	95%	2.064	1.7	$58.3 < \mu < 61.7$	Same
I	**80**	4	25	24	95%	2.064	1.7	$78.3 < \mu < 81.7$	Same

Three things that make $t\frac{s}{\sqrt{n}}$ *smaller*, and thus make $\bar{y} - t\frac{s}{\sqrt{n}} < \mu < \bar{y} + t\frac{s}{\sqrt{n}}$ *narrower* (a narrower interval is better ☺ because it means a more precise estimate for μ):

Smaller t-value

Smaller SD s

Larger sample size n

Two observations:

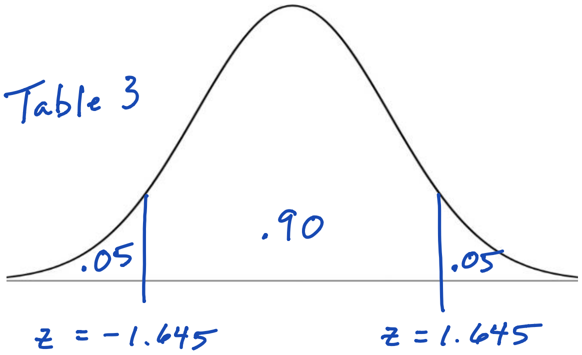
- A larger sample size n makes $t\frac{s}{\sqrt{n}}$ smaller in two ways: (1) due to n itself, and (2) larger n ⇒ smaller t (see Table 4). Compare cases D and E to case A.
- A higher confidence level ☺ ⇒ larger t-value ⇒ wider confidence interval ☹. It's a trade-off. ☺☹ Compare cases F and G to case A.

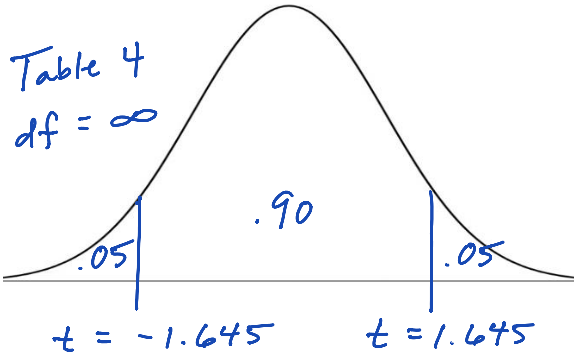
So it is beneficial to have a smaller t-value. Two things that make t smaller (see Table 4):

Larger sample size n (thus larger $df = n - 1$)

Lower level of confidence ☹

If the sample size is really large, then we can work with the sample means (that is, the sampling distribution) as if they are normally distributed, no matter the distribution of the original data. (See Item 3(b) from Page 152.) Notice in Table 4 that the t-values for $df = \infty$ (which means we have perfect normal distribution) correspond to the z-values from Table 3. As an example:





So Table 3 and Table 4 are kind of opposites of each other. In Table 3, we have a z-value and want to find a corresponding area/probability. In Table 4, we have an area/probability and want to find the corresponding t-value. If we could assume perfectly normal distribution, then the t-value (for approximately normal distribution) would actually simply be a z-value (for perfectly normal distribution). The more normal the distribution (which results from having a larger sample size), the closer our t-value is to the corresponding z-value. Again, you can see the t-values in Table 4 with $df = \infty$ are exactly the same as what you see in Table 3, just in reverse. (This is what the example drawn just above this paragraph illustrates.)

You can use Excel to find the t-value, also known as the $\underline{t\text{ multiplier}}$. The command **= TINV(0.05,24)** returns the value of approximately 2.064, which is the t multiplier which corresponds to a two tailed confidence interval of 95%, which means an area/probability of .025 in each tail. This is the same value we see in Table 4.

We can also find one-tailed probabilities. **Examples 6.3.6 and 7**.

Here's another way of describing today's ideas. Suppose we have a sample mean and SD (and the resulting $SE = s/\sqrt{n}$) that leads us to say that we are 95% confident that the population mean is between 68.3 and 71.7 (that is, a mean of 70 ± 1.7), as seen in the example on page 2 of today's handout. Another way to think of this is that if the population really had a mean of 70 with a SD of 4, then 95% of all sample means (of samples of size $n = 25$) would be expected to be in the range of 68.3 to 71.7. Recall the **eight samples on Page 157**, and that each had a mean that was a little different from the true population mean of 500. See **Example 6.3.3 and Figure 6.3.5** and **Example 6.3.4 and Example 6.3.5**, and ***Relationship to Sample…* on Page 190**. Read on your own (starting at the bottom of book **Page 187: "A confidence level…"**) about why the statement $\Pr\{31.4\ cm^2 < \mu < 34.2\ cm^2\} = .95$ doesn't make sense. This is subtle, so *do* try to understand it, but *don't* stress yourself out too much in your efforts.