David Youngberg

Acct/Busa/Econ 222—Bethany College

**Lecture 05: Understanding Intervals**

1. Let’s review
	1. *The* *law of large numbers* states a large sample is a better approximation of the population than a small sample. As sample size increases, the sample mean will approach the population mean.
	2. *The central limit theorem* states that for sufficiently large number of samples taken from ***any*** distribution, the distribution of the means of those samples will approximate a normal distribution. In other words for large sample sizes, we can assume we have a normal distribution.
	3. *Confidence Interval*—a range of values constructed from sample data so that the population parameter is likely to occur within that range at a specified probability. That probability is the *level of confidence*.
		1. Or, if we were to gather several samples, the proportion of sample means that is in our interval would be equal to our level of confidence.
		2. If a 95% level of confidence for a sample is between 0 and 3, then after gathering 100 samples, 95 of those samples would have an average between zero and three.
		3. Note 95% confidence is *not* “I am 95% confident…” Confidence is a mental state of being. No one says “I am 95% smart.”
2. The z distribution
	1. Recall when we used Excel to calculate the area under the distribution. As noted, this was the probability that, selecting randomly, you’ll get a value in that range.

X

0.025

0.025

-1.96

+1.96

0.95

* 1. The amounts to the same thing as a confidence interval: the probability that the true population mean is between certain values. Indeed, I can use the same technique for the lab to arrive at 95%.



* + 1. Notice I’m using a standard normal for simplicity purposes.
		2. Descriptive statistics in Excel give us the confidence interval. It does this by multiplying 1.96 (for 95% confidence) by the standard error of the mean. The result is added and subtracted from the mean to get our interval.
	1. As we discussed, when you take a sample, you hope/assume/pray that it represents the population. If you know what the standard error of the population (which sometimes happens) you can determined the standard error of the sample with the following equation.

$$σ\_{\overbar{x}}=\frac{σ\_{p}}{\sqrt{n}}$$

* + 1. Where *σx-bar* is the standard error of the sample mean,
		2. And *σp* is the standard deviation of the population,
		3. And *n* is the sample size.
	1. To calculate confidence levels for z levels of confidence, use this equation:

$$\overbar{x}\pm z\frac{σ\_{p}}{\sqrt{n}}$$

* + 1. Where *x-bar* is the sample mean,
		2. And *n* is the sample size,
		3. And *σp* is the standard error of the population,
		4. And *z* is the standard deviations for your desired level of confidence. Common *z* values are 1.96 (for 95%) and 2.58 (for 99%).
1. The t distribution
	1. In the previous section, we knew the standard deviation of the population. For example, there’s a long history of lots of observations recorded under similar circumstances (the average amount of toothpaste in a tube in a factory; the average weight of freshly forged steel in a steel mill; the average American blood pressure, since virtually everyone’s blood pressure is taken and recorded many times over their life).
	2. Thanks to the law of large numbers, if you get enough samples you might as well be looking at the population.
	3. But often you don’t know the population standard deviation because you don’t have access to enough measurements? We have lots of stock information, but circumstances change day-to-day so using that to construct a true population stock price is meaningless. If the factory changes its equipment, that data it collected on toothpaste isn’t relevant anymore. The population standard deviation (and mean) is once more a mystery.
	4. So we use the t distribution:

$$t=\frac{\overbar{x}-μ}{{s}/{\sqrt{n}}}$$

* + 1. Where *s* is the sample’s standard deviation,
		2. And *μ* is the population mean,
		3. And all other variables are the same as for the z distribution.
		4. The resulting t score is compared to a critical value based on desired significance level and degrees of freedom. You can use Excel to determine the significance value you’ve reached with =TDIST.
		5. Recall that the “degrees of freedom” is a measure of the amount of information that goes into the estimation. The larger the sample, the larger the degrees of freedom. As the sample size approaches infinity, the t distribution becomes identical to the z distribution.
	1. For the confidence interval, we do something very similar to what we did earlier:

$$\overbar{x}\pm t\frac{s}{\sqrt{n}}$$

* 1. Note that the t distribution is different from the z distribution; the t distribution has a higher standard deviation, depending on the degrees of freedom.
		1. For example, 95% confidence for a z distribution is 1.96 standard deviations. For that same level of confidence for a t distribution with 4 degrees of freedom, you need 2.776 standard deviations.
1. Proportion
	1. Consider confidence intervals of a proportion (a ratio of a binomial variable)
		1. Examples: the proportion of English majors (you can either be one or not); the proportion of Burger Kings with drive up windows (one has it or it doesn’t); the proportion of hours you are in class (you either are in class or you are not)
	2. By definition the proportion of one thing equals 1 – the proportion of the other. Despite what your coach may have told you, you cannot give 110%: proportions are bounded between 0 and 1.
	3. To find the proportion confidence interval, we will be using the z distribution.

$$p\pm z\sqrt{\frac{p(1-p)}{n}}$$

* + 1. Suppose Craig is running for office and he’s curious how likely it is he’ll win. In a sample of 200 voters, 88 say they will vote for him. With only two candidates, Craig needs 26 (or 50% + 1) voters. This doesn’t look good for Craig, but if he’s in a 95% confidence interval, it’ll be easier for him to raise some money.

$$0.44\pm 1.96\sqrt{\frac{0.44(1-0.44)}{200}}=0.44\pm 1.96\sqrt{\frac{0.44(0.56)}{200}}$$

$$0.44\pm 1.96\sqrt{\frac{0.2464}{200}}=0.44\pm 1.96\left(0.035\right)=0.44\pm 0.0688$$

* + 1. There is a 95% chance that Craig will get somewhere between 0.3712 and 0.5088, or 37.12% and 50.88% of the vote.
		2. Things are not looking good for Craig as 50% is barely in our interval, but this evidence might be enough to raise the money needed to improve his chances.
1. Optimal sample size
	1. Choosing your sample size is a tough decision. On one hand, more increases accuracy. But each additional observation becomes more expensive to get (remember, increasing marginal cost).
	2. One way to get an idea on how big the sample size should be for estimating a population mean is to use a little algebra on the equation from III.d.

$$n=\left(\frac{zσ\_{p}}{E}\right)^{2}$$

* + 1. Where *E* is the size of the error.
		2. Example: Cassandra wants to estimate the number of ships that come in on an average day, with an error no more than 1 ship each day at 95% confidence. If the standard deviation is 3 ships per day, how many days should Cassandra sample?

$$n=\left(\frac{(1.96)3}{1}\right)^{2}=5.88^{2}=35.5 days, or 36 days$$

* + 1. Always round up to make sure you have enough observations.
	1. We can do similar algebra on VII.c. for proportions.

$$n=p(1-p)\left(\frac{z}{E}\right)^{2}$$

* + 1. Where *E* is the size of the error;
		2. And *p* is the population proportion. If we don’t know the population proportion (which is common), p=0.5.
		3. Example: Roxanne wants to know what portion of men use cologne. She wants a range no more than 0.2 (0.1 on either side of the mean) at 95% confidence.

$$n=0.5\left(1-0.5\right)\left(\frac{1.96}{0.1}\right)^{2}=0.25\left(384.16\right)=96.04 men, or 97 men$$

**Lab Section**

1. Chapter 9
2. Homework
	1. Chapter 9: 1-4