Stat 218 - Day 10
Confidence Intervals

Recall:

- A numerical characteristic of a population is a **parameter**; its value is fixed but typically not known.
- A numerical characteristic of a sample is a **statistic**; it can vary from sample to sample and is often used to estimate the value of an unknown parameter.

Example: Foreign birthweights

We have said that birthweights in the U.S. follow a normal distribution with mean $\mu = 3250$ grams and standard deviation $\sigma = 550$ grams. Suppose that we record birthweights of newborn babies in a foreign country, where we are still willing to believe that the distribution is normal with standard deviation $\sigma = 550$ grams, but we want to estimate the value of the population mean μ. Suppose that we take a random sample of n=10 babies and record their birthweights (in grams) as: `2476 2533 2576 2643 2735 2889 3146 3345 3537 4152`

(a) How would you suggest estimating the population mean μ with a single number?

(b) Do you believe that the population mean (if we could record the birthweight of *every* newborn baby in the country this year) equals this number exactly?

The natural choice to estimate a population mean μ is the sample mean $\bar{y}$, but we know (sampling variability) that this number is unlikely to equal the population mean exactly. We would like to estimate the population mean μ with an *interval* of values that we believe to contain the value of the population mean μ with a certain *level of confidence*.

(c) What does our key result from last week say about how the sample mean birthweight would vary from sample to sample? Draw a sketch to support your answer.

(d) Determine the probability that the sample mean birthweight would fall within $\pm$341 grams of the population mean μ.

(e) Is this probability correct regardless of the actual value of the population mean μ?

(f) Determine the probability that the interval from $\bar{y}$ -341 to $\bar{y}$ +341 would contain the actual value of the population mean μ.

This suggests that we can be about 95% confident that if we take the sample mean $\bar{y}$ and construct an interval by subtracting 341 and then adding 341, we can be about 95% confident that our interval contains the actual value of the population mean μ.

(g) Determine this interval for the sample data above. (The sample mean birthweight is 3003.2.)

This interval is called a 95% **confidence interval**. The value 95% is called the **confidence level**, and the number 341 is called the **margin-of-error**. This margin-of-error depends on three factors:

- the confidence level C
- the sample size *n*
- the population standard deviation σ (tomorrow we will learn what to do when σ is unknown, which is almost always the case)

Where did this number 341 come from? To understand this we first need to determine the *z*-score (let's call it *z**) with the property that 95% of the area under a standard normal curve falls between –*z** and *z**.

(h) Determine this *z** value. [Again start with a sketch.]

(i) Multiply this *z** value by σ/sqrt(10). Does the result look familiar?

Our CI formula for estimating a population mean μ when the population standard deviation σ is known is: $\bar{y} \pm z^* \dfrac{\sigma}{\sqrt{n}}$.

(j) To construct a 90% confidence interval, first find the z^* value such that the area under a standard normal curve falls between $-z^*$ and z^* is .90.

(k) Multiply this value by σ/sqrt(10). Then construct an interval by adding and subtracting this value from the sample mean $\bar{y}$. How does this interval compare to the 95% CI? Does this make sense?

(l) Now suppose that the sample size had been n=40. Determine the margin-of-error of a 95% confidence interval for the population mean μ in this case. How does it compare to the M-of-E in the n=10 case?

Applet Demonstration:
Interpreting confidence intervals correctly requires us to think about what would happen if we took random samples from the population over and over again, constructing a CI for the population mean μ from each sample. We will turn to a Java applet called "Simulating Confidence Intervals" to do this.

First we will make sure that the method is set to "Means" and "z with sigma." We'll also set the population mean to be 3000, the population standard deviation to be 550, the sample size to be 10, and the confidence level to be 95%.

(a) As we take new samples, what do you notice about the intervals? Are they all the same?

(b) Does the value of the population mean change as we take new samples?

(c) As we take hundreds and then thousands of samples and construct their intervals, about what percentage seem to be successful at capturing the population mean?

(d) Sort the intervals, and comment on what the intervals that fail to capture the population mean have in common.

(e) Now change the confidence level to 90%. What two things change about the intervals?

(f) Now change the sample size to 40. Does this produce a higher percentage of successful intervals? What does change about the intervals?

(g) Now change the population standard deviation to 1000 and comment on how the intervals change.

(h) Now change the population mean to 4000 and comment on how the intervals change.

Preview: The problem with this CI analysis is that it is based on the ridiculous assumption that we know the population standard deviation σ. But if we knew that, then we would also know the population mean μ, and so we would not have to estimate it!

(a) What's a reasonable substitution to make for the population standard deviation σ based on the sample data?

To see what's wrong with this, go back to the applet and set the sample size back to 10 and the confidence level back to 95%. Change the method to "z with s" and generate several hundred samples and intervals.

(b) Is the success rate of these intervals very close to 95%?

To fix this problem, we need to use a different multiplier of the s/sqrt(n) term. This multiplier will come from the t-distribution.