Section 8.1: Confidence Intervals

At the end of 7.3, we defined a *t*-statistic, and I made a video in which we worked out a problem using this definition. In particular, when $\bar{X}$ is the sample mean of 25 observations of a normal random variable with mean μ, we wanted to find the value of *c* such that $P\left(\left|\frac{\bar{X}-\mu}{S}\right| \leq c\right)$ = .90.

We ended up getting the answer *c* = .3422.

So there is a probability of 90% that $-.3422 \leq \frac{\bar{X}-\mu}{S} \leq .3422$. What happens if we solve this inequality for μ?

Multiplying by -S, we get $-.3422S \leq \bar{X} - \mu \leq .3422S$.

Finally, adding $\bar{X}$ to all three sides gives
$\bar{X} - .3422S \leq \mu \leq \bar{X} + .3422S$.

This means given a sample mean and standard deviation, the sample size n, and the critical value *c*, we can find an interval that has a probability of 90% of containing the population mean. This is a confidence interval.

> Confidence Intervals
>
> A **confidence interval** for an unknown parameter θ is an interval that contains a set of plausible values of the parameter. It is associated with a **confidence level** 1 – α, which measures the probability that the confidence interval actually contains the unknown parameter value.

Note: Recall that we got *c* from the t-table using $\alpha = .05$ and 24 degrees of freedom. But the value of α came from using symmetry and computing $\frac{1-.90}{2}$. To avoid confusion (or perhaps more accurately, to create unnecessary confusion), the book now uses α in a different way than it does on the t-table. That is, by this definition, in order to be 90% confident, we set 0.90 = 1 – α, so here, α = 0.10. Long story short, be careful about what α means.

But in our definition of a *t*-statistic, we required the sample mean to be the mean of observations from a normal distribution. However, the Central Limit Theorem tells us that if our sample size is large enough, any sample mean is close to a normal distribution! That is, even if our observations aren't normally distributed, we may assume this to be true to get a confidence interval for the sample mean when the sample size n ≥ 30. If n < 30, we must check the data to see if it can reasonably be taken to be normally distributed.

In the above example, we got a two-sided confidence interval. Numerically, suppose that from our 25 observations, we got a sample mean of 12.5 and a standard deviation of 2.36. Then we found that we can be 90% confident that the actual mean μ of the population lies in the interval
$\left(12.5 - (.3422)(2.36), 12.5 + (.3422)(2.36)\right) = (11.6924, 13.3076)$.

Indeed, the most commonly used confidence interval for a population mean μ based on a sample of n continuous data observations with sample mean $\bar{X}$ and a sample standard deviation S is a two-sided t-interval.

<u>Two-Sided *t*-Interval</u>

A confidence interval with confidence level 1 – α for a population mean μ based upon a sample of n continuous data observations with a sample mean $\bar{x}$ and a sample standard deviation s is

$$\mu \in \left(\bar{x} - t_{\alpha/2,\ n-1} \bullet \frac{s}{\sqrt{n}} \ , \ \bar{x} + t_{\alpha/2,\ n-1} \bullet \frac{s}{\sqrt{n}} \right)$$

The interval is known as a **two-sided *t*-interval.**

The length of the confidence interval is $L = 2 \bullet t_{\alpha/2,\ n-1} \bullet \frac{s}{\sqrt{n}}$ = 2 ·(critical point)·(s.e.($\hat{\mu}$))

Notes about confidence intervals:

1) As the confidence level increases, the length of the confidence interval increases. (To be 99% sure you have included the parameter you must include more possible values than you would to be 95% confident).
2) As the sample size increases, the length decreases. However, due to the square root of the sample size in the formula, it takes a fourfold increase in the sample size to cut the length in half.
3) Even using high confidence levels, the confidence interval still might not actually contain the population mean! In general, there's no way to get 100% confidence and still say something useful.

A common problem we'll have will be to find a confidence interval with a given maximum length. For example, we might be required to be 95% confident of the population mean ±5 units. This means we need the length to be less than or equal to 10. We can rig this by solving the length formula for n. The book likes to call the required maximum length L_0, so we have:

$$L_0 \geq 2 \bullet t_{\alpha/2,\ n-1} \bullet \frac{s}{\sqrt{n}} \quad \Rightarrow \quad \sqrt{n} \geq 2 \bullet t_{\alpha/2,\ n-1} \bullet \frac{s}{L_0} \quad \Rightarrow \quad n \geq 4 \bullet \left(\frac{s \bullet t_{\alpha/2,\ n-1}}{L_0} \right)^2$$

But wait: $t_{\alpha/2, n-1}$ also depends on n, and we can't know the sample standard deviation without knowing what the observations will be! To deal with this, we can use an underestimate for n in choosing a value for $t_{\alpha/2, n-1}$, since we know the value of this critical number decreases as n increases. The way the book phrases a lot of these problems is that they will tell you a preliminary value of n and some assumed knowledge of s that will let you plug in s and $t_{\alpha/2, n-1}$ to find a better value of n.

Generally, n is rounded up to the next highest integer, since we can't have a fraction of an observation. Note, however, that it is always rounded UP, because rounding down gives an integer that is not sufficient large!

The following box gives the rule for one-sided confidence intervals. These are used when only an upper (or lower) bound on the mean is needed. Because of this we do not have to split α between two tails, so we will be using $t_{\alpha,\, n-1}$ not $t_{\alpha/2,\, n-1}$. One-sided *t*-intervals give a bound that is closer to the sample mean than the limits of the two-sided *t*-interval at the same level of confidence.

<u>One-sided *t*-interval</u>

One-sided confidence intervals with confidence levels 1 – α for a population mean μ based on a sample of n continuous data observations with a sample mean $\bar{x}$ and a sample standard deviation s are

$\mu \in \left(-\infty \,,\, \bar{x} + t_{\alpha,\, n-1} \bullet \frac{s}{\sqrt{n}} \right)$ for an upper bound on the population mean μ, and

$\mu \in \left(\bar{x} - t_{\alpha,\, n-1} \bullet \frac{s}{\sqrt{n}} \,,\, \infty \right)$ for a lower bound on the population mean μ.

These confidence intervals are known as **one-sided *t*-intervals**.

Finally, if the population standard deviation σ is known then it is possible to use the standard normal distribution instead of the *t*-distribution. This is not typically going to be used, since it may be unrealistic to already know exactly what the actual standard deviation is, and yet still require a confidence interval for the population mean. However, the formulas for the two-sided and one-sided versions of this situation are given here:

<u>Two-Sided z-Interval (σ known)</u>

A confidence interval with confidence level 1 – α for a population mean μ based upon a sample of size n with a sample mean $\bar{x}$ and a known population standard deviation σ is

$$\mu \in \left(\bar{x} - z_{\alpha/2} \bullet \frac{\sigma}{\sqrt{n}} \,,\, \bar{x} + z_{\alpha/2} \bullet \frac{\sigma}{\sqrt{n}} \right),$$

where z_α is the critical number of the standard normal distribution (from Table I on page 787).
The interval is known as a **two-sided z-interval** or variance known confidence interval.
A confidence interval might not contain the population mean. 95% confidence is not 100% confidence.

One-sided z-interval (σ known)

One-sided confidence intervals with confidence levels 1 – α for a population mean μ based on a sample of n observations with a sample mean $\bar{x}$ and a "known" population standard deviation σ are

$\mu \in \left(-\infty \,,\; \bar{x} + z_\alpha \bullet \frac{\sigma}{\sqrt{n}} \right)$ for an upper bound on the population mean μ, and

$\mu \in \left(\bar{x} - z_\alpha \bullet \frac{\sigma}{\sqrt{n}} \,,\; \infty \right)$ for a lower bound on the population mean μ.

These confidence intervals are known as **one-sided *z*-intervals**.

As a last note, in the brief Section 5.4 notes, I mentioned that the t-distribution looked more and more like a standard normal as the degrees of freedom approached infinity. This means that for large samples, there is little difference between the z-intervals and the t-intervals.