Confidence Intervals

A 95% confidence interval is an interval constructed from a random sample in such a way that approximately 95% of such intervals will contain the true (and unknown) population mean, μ. In other words, if you do an experiment 100 times and generate one hundred $\bar{x}$ means, then about 95 of the intervals constructed, one for using each $\bar{x}$, will contain μ. (It's *not* correct to say that there is a 95% chance that the population mean lies within the interval. Explained later.)

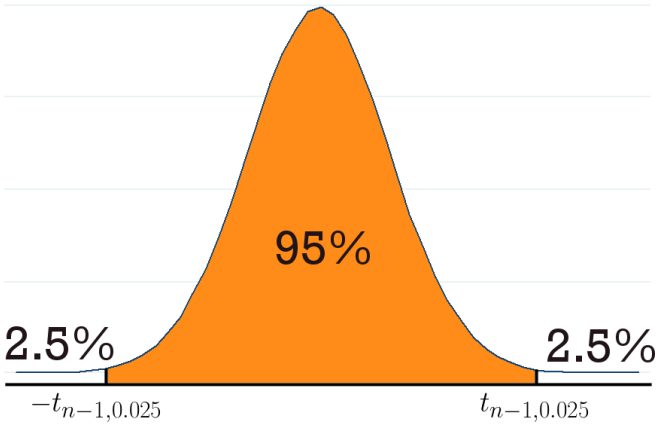
Ultimately we are trying to construct some value A, which depends on our data, such that

$$\Pr\left(-A \leq \mu \leq A\right) = 0.95.$$

One way to approach this is to standardize. We know it is approximately true (and sometimes exactly true) that

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim T(n-1). \tag{1}$$

Because the distribution is symmetric, there must exist some value $t_{n-1,0.025}$ such that there is a 95% probability that anything drawn from this distribution lies within the interval $[-t_{n-1,0.025}, t_{n-1,0.025}]$.



Hence we can write

$$\begin{aligned}
0.95 &= \Pr\left(-t_{n-1,0.025} \leq \frac{\bar{x}-\mu}{s/\sqrt{n}} \leq t_{n-1,0.025}\right) \\
&= \Pr\left(-t_{n-1,0.025} \times \frac{s}{\sqrt{n}} \leq \bar{x}-\mu \leq t_{n-1.0.025} \times \frac{s}{\sqrt{n}}\right) \\
&= \Pr\left(-\bar{x} - t_{n-1,0.025} \times \frac{s}{\sqrt{n}} \leq -\mu \leq -\bar{x} + t_{n-1,0.025} \times \frac{s}{\sqrt{n}}\right) \\
&= \Pr\left(\bar{x} + t_{n-1,0.025} \times \frac{s}{\sqrt{n}} \geq \mu \geq \bar{x} - t_{n-1,0.025} \times \frac{s}{\sqrt{n}}\right).
\end{aligned}$$

The first step multiplied all sides by $s/\sqrt{n}$. The second step subtracted $\bar{x}$ from all sides. The third step multiplied all sides by -1 to get μ instead of $-\mu$.

So we have constructed the 95% confidence interval for μ,

$$\left[\bar{x} - t_{n-1,0.025} \times \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,0.025} \times \frac{s}{\sqrt{n}}\right]. \tag{2}$$

That's the formula to use, and you will be seeing it repeatedly. The Stata command for $t_{n-1,0.025}$ is `invttail(n-1, 0.025)`, or to actually see the number,

```
di invttail(n-1, 0.025).
```

Again, the interpretation is that for $i = 1, \ldots, 100$ sample means $\bar{x}_i$, we expect 95 of the confidence intervals, one constructed for each $\bar{x}_i$, to contain μ. Of course, we aren't going to calculate 100 sample means in practice – we're going to calculate one sample mean with all of our data. Relative to the specific confidence interval that we actually calculate:

- **Correct Interpretation:** The 95% confidence interval calculated from this sample includes the true population mean μ with probability 0.95. *(Notice that the probabilistic statement is about the interval, which is random, but not about μ.)*
- **Incorrect Interpretation:** There is a 0.95 probability that μ lies within the 95% confidence interval calculated from this sample. *(Notice that the probabilistic statement is about μ, but μ is not random – it's an unknown constant.)*

Note that less confidence gives a smaller interval. Think back to the interpretation of a confidence interval: a 90% confidence interval means that a *smaller* percentage of constructed intervals will actually contain μ, so it makes sense that the corresponding interval is a tighter one. (We're less confident about hitting a smaller target, in a sense. Or another way of

thinking about it: to be really confident that the interval contains μ, it must be a really big interval.)

Two-Sided Hypothesis Testing

Suppose we have some guess about what the population mean μ is. If it's a good guess, then intuitively it should be "close" to the sample mean $\bar{x}$, because we expect $\bar{x}$ itself to be "close" to μ for a large enough sample size (the law of large numbers). Hypothesis testing is a way of formalizing "closeness."

We start with a **null hypotheses**. This is our guess for what μ is. Let μ_0 be that guess. We express the null hypothesis as

$$H_0 : \mu = \mu_0.$$

In English: my null hypothesis H_0 is that the population mean μ equals my guess μ_0.

We need to test the null hypothesis against something – we call this the **alternative hypothesis**. The simplest case is that our guess is wrong, which we express as

$$H_1 : \mu \neq \mu_0.$$

Here is how the test proceeds in narrative terms. We assume that our guess is true. Then we compute a difference between our guess and the sample mean. If we've made a good guess, then the difference should be nearly zero. If the difference is big (in either positive or negative direction), then our guess was probably bad, so we reject our guess.

Now let's carry the test out. The way to quantify "closeness" is with the expression

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \equiv t,$$

where the number t is referred to as a ***t* statistic**, a specific type of **test statistic**. If the null hypothesis is true, then the t statistic is $T(n-1)$ distributed (because it has the exact same form as in the sampling distribution). By definition, we know that 95% of the draws from a $T(n-1)$ distribution will fall within the interval

$$[-t_{n-1,0.025}, t_{n-1,0.025}].$$

Numbers $-t_{n-1,0.025}$ and $t_{n-1,0.025}$ are called **critical values**. If the test statistic falls beyond the critical values – in the **rejection region** – then we *reject the null at significance level 0.05*. Such is our **rejection rule**.

In English: If my guess is true, then 95% of these test statistics should fall within this interval. But what if my test statistic doesn't lie within this interval? There's only a 5% chance of that actually happening if my guess is actually true, which is pretty unlikely. So my guess is probably bad.

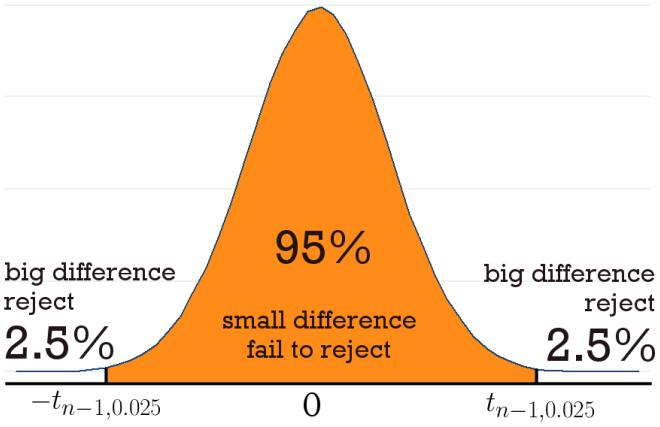
If the guess does lie within the interval, then we *fail to reject the null hypothesis at significance level 0.05*. We never say "we accept" or "we confirm" the null hypothesis due to the logic employed. To illustrate, the following two statements are logically equivalent:

- If the null is true, then t is probably close to zero. (If A, then B.)
- If t is not close to zero, then the null is probably not true. (If not B, then not A.)

The hypothesis procedure assumes that the null is true, which is why we can use the second bullet point as a logical justification to reject the null when t is big enough. It *not* logically equivalent, however, to say that

- "If t is close to zero, then the null is probably true." **No!** (If B, then A. **No!**)

In fact, this is a logical error made commonly enough that it has its own name: *affirming the consequent*. Hence the procedure of our test gives no logical grounds for accepting the null; we can either reject or not reject.[1]



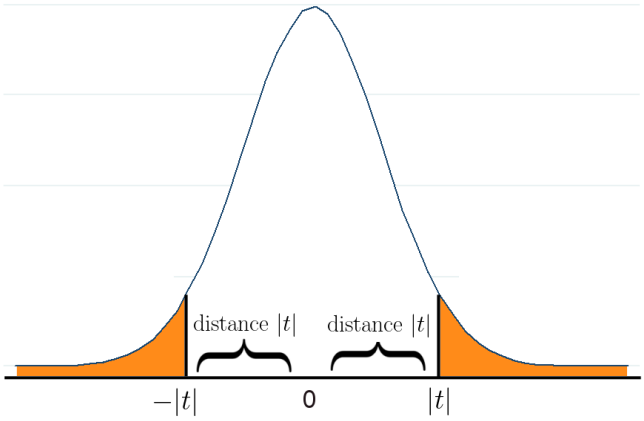
Here's another way to think about it. We're interested in the closeness of our guess to the sample mean. We can use absolute value as the "distance" between the two. If the distance is too big, then we reject the null. Then we can simplify and reject if $|t| > t_{n-1,0.025}$.

[1]Statistics, and much of science more generally, can falsify but not confirm. See: Karl Popper. We can never prove something about the entire population unless we have the entire population of data, which in practice we rarely do.

p-values

The p-value tells you the probability of observing a number more extreme in magnitude (that is, in either positive or negative direction) than the t statistic you've found, supposing that the null hypothesis is true.

Suppose you calculate your t statistic and find that $t = -1$. What is the probability of getting a random $T(n-1)$ draw, call it T_{n-1}, that is greater than $|t| = 1$ in absolute value? It's the probability of drawing less than $-|1|$ plus the probability of drawing greater than $|1|$. In pictures, it's the probability of being in the orange region below:



Note that the two tails are identical in mass because $T(n-1)$ is symmetric about zero, so we can just calculate one tail and double it. Or to put it in the maths,

$$\begin{aligned} p &= \Pr(T_{n-1} < -|t|) + \Pr(T_{n-1} > |t|) \\ &= 2 \times \Pr(T_{n-1} > |t|) \\ &= 2 \times \Pr(T_{n-1} < -|t|). \end{aligned}$$

In practice, the equation $p = 2 \times \Pr(T_{n-1} > |t|)$ is the easiest to use, and this number can be found in Stata via command

```
di 2*ttail(n-1, |t|).
```

Note that a p-value less than 0.05 means there's a less than 5% chance of observing $\bar{x}$ if the null hypothesis is true – a small enough chance that our null is probably wrong. You are able to assert with some confidence that $\mu_0 \neq \mu$, and your assertion would be **statistically significant**. In other words, we can conclude that μ is statistically significantly different from μ_0.

Exact and Approximate Distributions

Depending on the specifics, our appeal to the central limit theorem can be either exact or approximate.

(a) If σ is known and $30 < n < \infty$, then use the standard normal distribution $\mathcal{N}(0,1)$. (Approximation)

(b) If σ is known and the underlying distribution is normal, then use the standard normal distribution $\mathcal{N}(0,1)$ for any n. (Exact)

(c) If σ is not known and $30 < n < \infty$, then use the $T(n-1)$ distribution. (Approximation)

(d) If σ is not known and the underlying distribution is normal, then use the $T(n-1)$ distribution regardless of n. (Exact)

The approximate cases become exact as $n \to \infty$. On paper, if $n > 30$, then you can usually use the normal distribution instead of $T(n-1)$ because they will be very similar. (In fact, $T(\infty)$ is exactly standard normal.) If you're using Stata, then just use $T(n-1)$ anyway. In practice, we will usually have unknown σ and $n > 30$, so the $T(n-1)$ statistic is used heavily.

Note that if $n \leq 30$, then we can only do "reliable" inference if we have reason to believe that the underlying data is normally distributed. Accordingly, you should be skeptical of inference on small sample sizes.