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 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.)

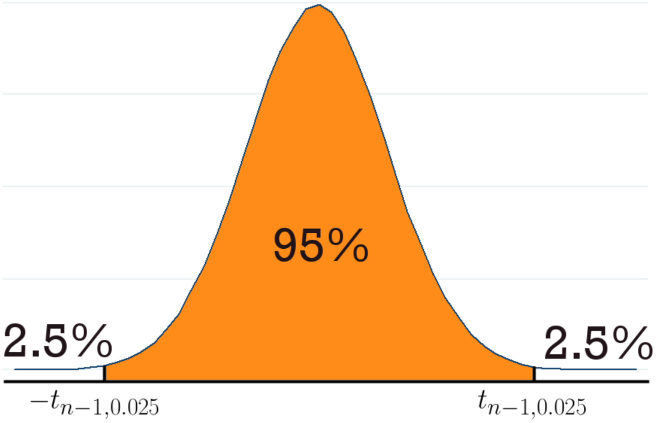
Ultimately we are trying to construct some value A, which depends on our data, such that

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

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

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

Thus, we know 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

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

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 turn the μ term positive.

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 R command for $t_{n-1,0.025}$ is either of the following:

- `qt(1-0.025, n-1)`
- `qt(0.025, n-1, lower.tail = FALSE)`

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. So consider a confidence interval based on $\bar{x}_i$, say, $[-A, A]$. The number A is just that, a *number*, determined from $\bar{x}_i$ and s_i and N; there is nothing random about it. Similarly, μ is just some number – just because we don't know what it is doesn't make it random. So whether μ is in $[-A, A]$ or not is not a probabilistic statement: it either is or it isn't. Thus it is not correct to say "there is a 95% chance that μ lies within $[-A, A]$." This is subtle, so don't panic if you don't quite understand it. But you should still be aware of the distinction.

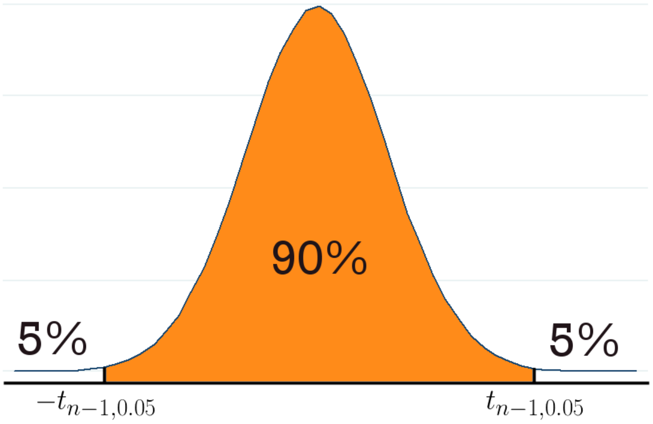
Problem 1. Suppose you calculate sample mean $\bar{x} = 50.06$ with sample standard deviation of $s = 11.27$. There are $n = 185$ observations. Find the 95% confidence interval for μ.

Answer 1. We need to find `qt(1-0.025, 184)`, which is about $t_{184,.025} = 1.97$. Thus the 95% confidence interval is

$$\left[50.06 - 1.97 \times \frac{11.27}{\sqrt{185}}, 50.06 + 1.97 \times \frac{11.27}{\sqrt{185}}\right] \quad = \quad [48.43, 51.69].$$

Problem 2. Now find the 90% confidence interval for μ.

Answer 2. The logical is exactly the same, we are just considering different values corresponding to 5% tails instead of 2.5% tails. So we are now considering a picture like the one seen below:



This warrants using R command `qt(1-0.05, 184)`, which is about $t_{184,.05} = 1.65$, giving

$$\left[50.06 - 1.65 \times \frac{11.27}{\sqrt{185}}, 50.06 + 1.65 \times \frac{11.27}{\sqrt{185}}\right] \quad = \quad [48.69, 51.43].$$

Notice that less confidence gives a smaller interval. Think back to the interpretation of a confidence interval: 90% means that a smaller percentage of our 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.)

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:

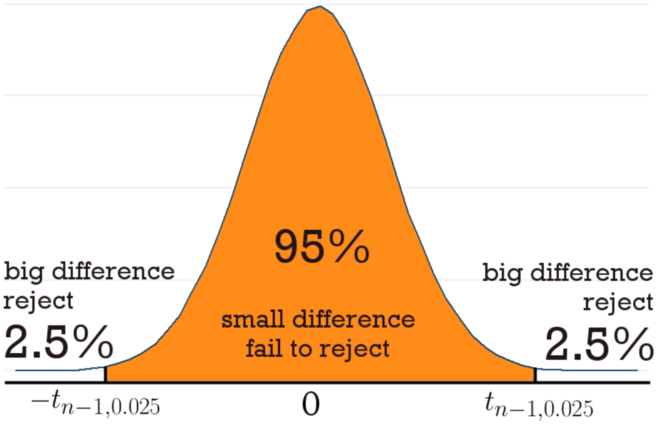
- If the null is true, then t is probably small. (If A, then B.)
- If t is not small, then the null is probably not true. (If not B, then not A.)

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 small, 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]

[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, usually implying $n \to \infty$, which in practice we rarely do.



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 $|t| > t_{n-1,0.025}$.

Problem 3. Suppose you calculate sample mean $\bar{x} = 50.06$ with sample standard deviation of $s = 11.27$. There are $n = 185$ observations. Test $H_0 : \mu = 52$ against $H_1 : \mu \neq 52$ at 95% confidence (i.e at 5% significance).

Answer 3. The null hypothesis is $H_0 : \mu = 52$, so our guess is $\mu_0 = 52$. Thus we are working with the test statistic

$$t \equiv \frac{50.06 - 52}{11.27/\sqrt{184}} \approx -2.34.$$

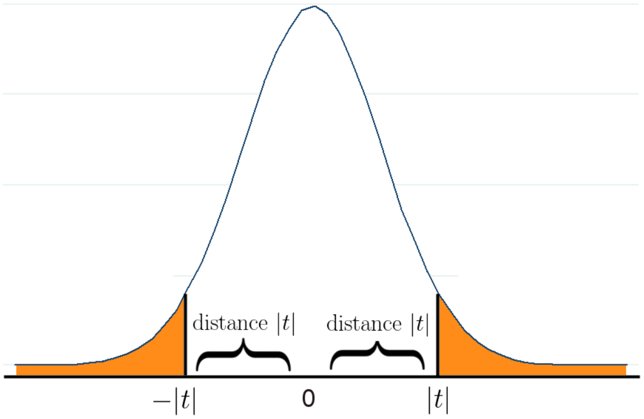
The critical values are given in R by command `qt(1-0.025, 184)`, which is approximately $t_{n-1,0.025} = 1.97$. Clearly -2.34 does not lie within the interval $[-1.97, 1.97]$, so we reject the null that $\mu = 52$.

Alternatively, $|-2.34| \geq 1.969$, so we reject.

p-values

The p-value tells you the probability of observing a number more extreme in magnitude 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,

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

The last two lines are what you'd use for R. See below.

Problem 4. Suppose you calculate sample mean $\bar{x} = 50.06$ with sample standard deviation of $s = 11.27$. There are $n = 185$ observations. Find the p-value of the test in problem 3.

Answer 4. We know from problem 3 that $t = -2.34$. Formulate the p-value as

$$\begin{aligned} p &= \Pr(T_{184} < -2.34) + \Pr(T_{184} > 2.34) \\ &= 2 \times \Pr(T_{184} > 2.34) \\ &= 2 \times \Pr(T_{184} < -2.34). \end{aligned}$$

This can be evaluated with either of the two following R commands:

- `2*pt(-2.34, 184)` $2 \times \Pr(T_{184} < -2.34)$
- `2*pt(2.34, 184, lower.tail = FALSE)` $2 \times \Pr(T_{184} > 2.34)$

They give roughly $p = 0.02$. Thus there is about a 2% chance of seeing sample mean as "extreme" as $\bar{x} = 50.06$ if our guess $\mu_0 = 52$ is actually correct. "2%? That's pretty low. My guess is probably bad." So reject the null, just as in problem 3, because this approach is equivalent to the approach taken in problem 3.

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**.