1 Confidence Intervals

A 95 percent **confidence interval** is an interval constructed from a random sample in such a way that approximately 95 percent 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 percent chance that the population mean lies within the interval. Explained later.)

1.1 Constructing a Confidence Interval

We want to construct some values A and B, which depend on our data, such that

$$\Pr\left(A < \mu < B\right) = .95.$$

One way to approach this is to appeal to the central limit theorem. For sufficiently large sample size, we know it is approximately true (and sometimes exactly true) that

$$T \equiv \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T(n-1), \tag{1}$$

and hence we can make probabilistic statements involving its actualization

$$t \equiv \frac{\bar{x} - \mu}{s/\sqrt{n}}. \tag{2}$$

Because the $T(n-1)$ distribution is symmetric, there must exist some value $t_{.025,n-1}$ such that there is a 95 percent probability that anything drawn from this distribution lies within the interval $(-t_{.025,n-1}, t_{.025,n-1})$. This is illustrated in Figure 1.

Hence we can write

$$\begin{aligned}
.95 &= \Pr\left(-t_{.025,n-1} < \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{.025,n-1}\right) \\
&= \Pr\left(-t_{.025,n-1} \times \frac{s}{\sqrt{n}} < \bar{x} - \mu < t_{.025,n-1} \times \frac{s}{\sqrt{n}}\right) \\
&= \Pr\left(-\bar{x} - t_{.025,n-1} \times \frac{s}{\sqrt{n}} < -\mu < -\bar{x} + t_{.025,n-1} \times \frac{s}{\sqrt{n}}\right) \\
&= \Pr\left(\bar{x} + t_{.025,n-1} \times \frac{s}{\sqrt{n}} > \mu > \bar{x} - t_{.025,n-1} \times \frac{s}{\sqrt{n}}\right).
\end{aligned}$$

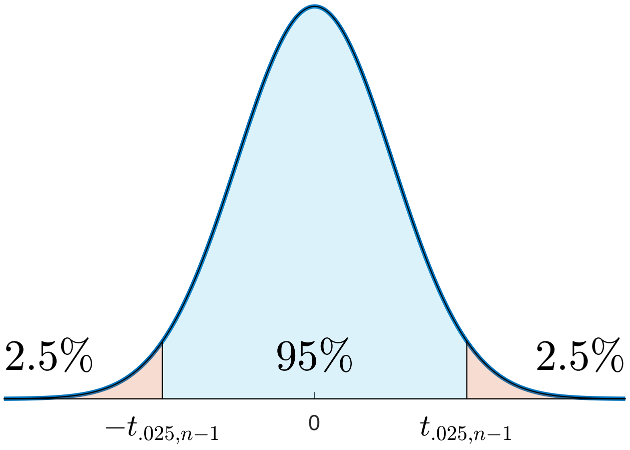

FIGURE 1: 2.5 percent of the $T(n-1)$ distribution lies beneath $-t_{.025,n-1}$ and 2.5 percent lies above $t_{.025,n-1}$.

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 percent confidence interval for μ,

$$\left(\bar{x} - t_{.025,n-1} \times \frac{s}{\sqrt{n}}, \bar{x} + t_{.025,n-1} \times \frac{s}{\sqrt{n}}\right). \tag{3}$$

That's the formula to use, and you will be seeing it repeatedly. The R command for $t_{.025,n-1}$ is either `qt(.025, n-1, lower.tail=FALSE)` or `qt(1-.025, n-1)`.

1.2 Interpretation

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 percent confidence interval calculated from this sample includes the true population mean μ with probability .95. *(Notice that the probabilistic statement is about the interval, which is random, but not about μ.)*[1]
- **Incorrect Interpretation:** There is a .95 probability that μ lies within the 95 percent confidence interval calculated from this sample. *(Notice that the probabilistic statement*

[1] The interpretation is actually even more subtle still. If you are interested, read more here. The least controversial interpretation is "we expect 95 percent of the confidence intervals to contain μ." Don't dwell too much on this; go with the correct interpretation above.

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 percent 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.)

2 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."

2.1 Null and Alternative Hypotheses

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

$$H_0 : \mu = \mu^*. \tag{4}$$

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

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

$$H_1 : \mu \neq \mu^*. \tag{5}$$

Here is how the test proceeds in narrative terms. We assume that our guess μ^* 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.

2.2 Critical Values and Rejection Region

Now let's carry out the test. The way to quantify "closeness" is again by appealing to the central limit theorem and using the actualization

$$t \equiv \frac{\bar{x} - \mu^*}{s/\sqrt{n}}, \tag{6}$$

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 drawn from the $T(n-1)$ distribution.

By definition, we know that 95% of the draws from a $T(n-1)$ distribution will fall within the interval $[-t_{.025,n-1}, t_{.025,n-1}]$. The numbers $-t_{.025,n-1}$ and $t_{.025,n-1}$ 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 .05*. Such is our **rejection rule**. This is illustrated in Figure 2.

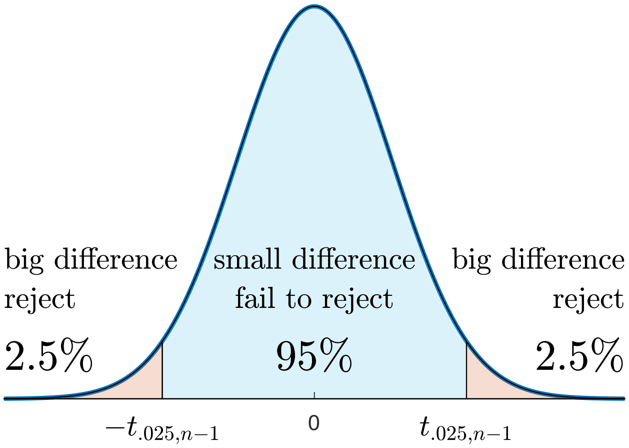

FIGURE 2: The null hypothesis is to be rejected when the t-statistic is too far from zero. For 5 percent significance, 2.5 percent significance goes in each tail.

In English: If my guess is true, then 95 percent 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 percent 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 .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.[2]

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 $|t| > t_{.025,n-1}$.

2.3 *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 $|1|$.

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 R via command `2*pt(abs(t), n-1, lower.tail=FALSE)` or equivalently with `2*(1-pt(abs(t), n-1))`.

[2] Statistics, and much of science more generally, can falsify but not confirm. See: philosopher of science 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. If you have data about 99.9999999% of swans and they are all white, that does not allow you to confirm that 100% of swans are white: you can only be more or less confident, but never totally certain, about the claim that 100% of swans are white. (When the phrase "black swan" was coined, people really didn't think black swans existed, so it was used for statements of supposed impossibility. And then someone eventually found a black swan. Oops.)

The p-value is illustrated in Figure 3 as the sum of the two shaded areas.

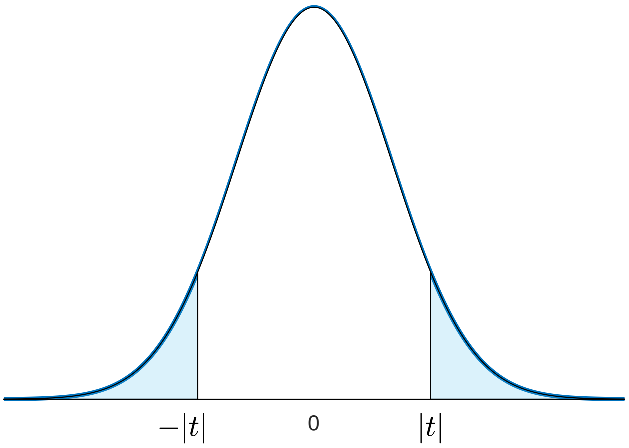

FIGURE 3: The p-value is the probability of observing a t-statistic as large in magnitude as the one we have observed, if the null hypothesis were true.

Note that a p-value less than .05 means there's a less than 5 percent chance of observing the $\bar{x}$ we've calculated 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 $\mu^* \neq \mu$, and your assertion would be **statistically significant**. In other words, we can conclude that μ is statistically significantly different from μ^*.

So we have three completely equivalent rationales for rejecting the null and asserting statistical significance:

- The t-statistic, in absolute value, is greater than the critical value $t_{.025,n-1}$;
- The p-value is less than .05;
- The value of the null hypothesis μ^* does not lie within the 95% confidence interval.

3 Errors

We're dealing with uncertainty, and because there's uncertainty, we might come to the wrong conclusion. We either reject the null or do not reject the null; our decision to reject could be right or wrong; and our decision to not reject could be right or wrong.

3.1 Type I Error

A **type I error** or **false positive** occurs when a true null is rejected. Think about how a hypothesis test at $\alpha = 5$ percent significance works. We assume that the null is true. If the null is true, then there's only a 5 percent probability that we observe a t-statistic in the rejection region. We consider that to be sufficiently unlikely, so if we do observe a t-statistic in the rejection region, we conclude that the null hypothesis is probably wrong, and so we reject it.

But still, there's a 5 percent probability that the null is in fact true and we just happened to get a t-statistic in the rejection region, and therefore our rejection of the null hypothesis is a mistake. The significance level α then is the probability of committing a type I error, that is, the probability of rejecting a null hypothesis even though the null hypothesis is correct; we also call this the **size** of a test.

3.2 Type II Error

A **type II error** or **false negative** occurs when a false null hypothesis is not rejected. The null hypothesis is wrong, but our sample is weird and we get a t-statistic that does not fall in the rejection region. In that case, we are failing to reject a null hypothesis even though the null hypothesis is wrong.

The **power** of a test is the probability of rejecting a null hypothesis when it is false,

$$\text{Power} \equiv 1 - \Pr(\underbrace{\text{failing to reject } H_0 \text{ when } H_0 \text{ is false}}_{\text{type II error}}).$$

Notice that there is a fundamental tension between type I and type II errors. To illustrate, consider the following extremes. We can completely avoid making a type I error if we *never* reject a null hypothesis by having $\alpha = 0$ percent; but then we'll never reject a false null hypothesis either, so there will be zero power. On the other hand, we can completely avoid a type II error by *always* rejecting a null hypothesis and having 100 percent power; but then we'll be rejecting every true null hypothesis, so the significance of a test will be 100 percent.

In practice, the significance of the test is chosen by the researcher. The common choice of 5 percent is arbitrary. Then, upon having chosen the significance of the test, the power of the test is maximized (that is, the *most powerful test* is chosen) by using the proper testing procedure. We won't worry too much about power.

4 Foreshadowing

Later we will be interested in the relationship between two variables via linear regression analysis. One thing we ask is, "is there a statistically significant relationship between x

and y?" Let β_2 be the number that captures that relationship. Our null hypothesis will be $H_0 : \beta_2 = 0$, that is, our null hypothesis is that there is no relationship between x and y. If there is sufficient evidence to the contrary, then we will end up rejecting the null in favor of $H_1 : \beta_2 \neq 0$ and therefore concluding that there is a statistically significant (i.e. statistically distinguishable from zero) relationship between x and y.