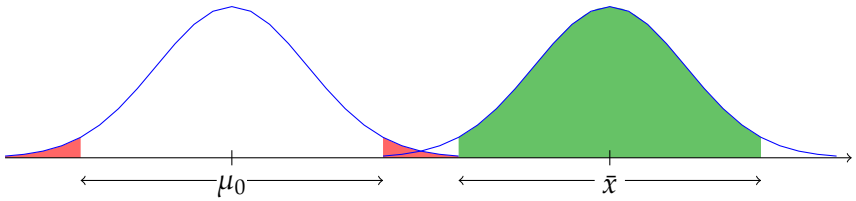
Some final words about hypothesis testing

Confidence intervals and hypothesis tests: two sides of the same coin?

We have discussed scenarios in which a sampling distribution (which includes null distributions) will be approximately normal, things like $n \geq 30$, or np and $n(1-p)$ at least 10, samples from two groups are independent, etc. Let us assume such *normality* is in effect.

In hypothesis testing, we use the null distribution, placing the null value (marked as μ_0 below) at the center. If the signficance level α represents the area in the tails (marked red), the observations not in the tails represent the nearest $100 \times (1-\alpha)\%$ of observations to μ_0. When the test statistic/point estimate (marked as $\bar{x}$) is far enough out to be in the red (called the **rejection region**), that leads to a P-value smaller than α.

On the other hand, in the construction of a $100 \times (1-\alpha)\%$ confidence interval, we place the *point estimate* in the center, extending out in either direction enough so that the width of the confidence interval (green) corresponds to the width of the **non-rejection region** (non-shaded region around μ_0). The instances in which μ_0 is not in the confidence interval are precisely those which land $\bar{x}$ in the rejection region.



So, say a random sample is collected. The sample statistic, perhaps $\bar{x}$, could be used to construct a confidence interval for the parameter, perhaps μ, but it also could be used as a test statistic in the 2-sided test of hypotheses

$$\mathbf{H}_0\colon\ \mu = \mu_0, \qquad \mathbf{H}_a\colon\ \mu \neq \mu_0.$$

But no matter which of these it is, confidence interval or hypothesis test, the one informs the other. Examples of the ways include these:

- If μ_0 is not in a 95% confidence interval, then the P-value of the hypothesis test is smaller than 0.05.
- If μ_0 is in a 90% confidence interval, then the P-value of the hypothesis test is larger than 0.1.
- If the P-value from the hypothesis test is 0.07, then μ_0 is in the 95% confidence interval, but not in the 90% confidence interval.

Cautions about multiple testing

Remember what was said earlier: setting $\alpha = 0.05$ for all your hypothesis tests means that, in cases where the null hypothesis is true, you will commit Type I error 5% of the time, *mistakenly* rejecting $\mathbf{H}_0$. That's a Type I error rate of 1-in-20. Any researcher conducting numerous statistical tests with $\alpha = 0.05$ should keep this in mind, and should maintain a healthy suspicion if about 5 percent of her tests are yielding statistically significant results. If, over the last year, 40 hypothesis tests have been conducted and 3 have been statistically significant, that is right near what we might expect to happen even if *none* of the null hypotheses in those tests of significance have been false.

Statistical significance is different from practical importance

Referring to the picture above, statistical significance amounts to our test statistic being far enough from the null value (μ_0) that it lands in the rejection region, nothing more. This may be evidence enough to reject the null hypothesis in favor of the alternative $\mathbf{H}_a: \mu \neq \mu_0$, but it does not necessarily follow that the true value of μ is far away. You may have evidence that is statistically sifnificant for showing that a drug does not leave blood pressure unchanged in those who suffer from high blood pressure even if its effect is only to decrease systolic pressure by 1.

Ellenberg has useful illustrations of hypothesis testing in Chapters 6 and 7 of his book, "How Not to Be Wrong: The Power of Mathematical Thinking." While I mention here his point that the word *significance* in statistics is meant in a technical sense that English language speakers are likely to misconstrue, I will let Ellenberg have the burden of hammering it home.