Statistical Significance

You might have heard of the phrase: ”statistically significant difference." We know that significance is indicated by a p-value < .05, but what does that actually mean? What does *significance* really mean?

What does p < .05 really mean?

Does it mean:

A. There is a 95% chance that the alternative hypothesis is true
B. This finding will replicate 95% of the time
C. If the study were to be repeated, the null hypothesis would be rejected 95% of the time

Nope. It means *none* of those things.

The **p-value** is defined as the probability of finding these results *IF* the null hypothesis is true. We consider the outcome as being "statistically significant" at p < 0.05 if the outcome we observed was *unlikely* to have occurred by chance.

What p < .05 really means

1. It tells us that the results were **statistically significant.** But it does NOT tell us whether the results are **practically significant.** In other words, do these results really matter in the real world?
2. It tells us that the treatment **had an effect.** But it does NOT tell us **how large** that effect was. And finding a p-value that is very small, like 0.0001, does NOT make that finding "more significant." There is no such thing as being "more significant" — results are either statistically significant or not.
3. It tells us to **reject the null hypothesis**. But it does NOT prove that the alternative hypothesis is correct. Whenever we make a decision, we can make a mistake. And to avoid making mistakes, we ought to rely on evidence *in addition* to p-value so that we can make informed decisions.

So, what can we do to address questions of effect size and practical significance?

1. Measures of effect size, like Cohen's D and R squared
2. Confidence intervals
3. Power of a test

What you want to know…

- The probability that the alternative hypothesis is true, given the evidence
- In other words, "Did the treatment have an effect?"

…but what you get is:

- The probability of the evidence, *assuming that the null hypothesis is true*
- If the treatment really did work, how likely is it that we would have found these results?

The big problems with this approach:

1. The null hypothesis is never true. The treatment always works. The treatment always does *something*, even if that something is trivial. And means are never truly equal. If you go enough decimal places, they *will* diverge. So the null hypothesis is never actually true.
2. Improbable things happen all the time. If you run enough tests with the same ineffective treatment, some of those means will in fact be statistically significantly different. What does that mean? Well, $p < 0.05$ means a 5% error rate. That means that 5 times out of 100, or 1 time out of 20, we will make a mistake. We will mistakenly attribute a result to having statistical significance, when in reality, it's not. This is called a type 1 error.
3. Statistical significance is a function of sample size. With enough people, every comparison will be statistically significant. Even if it is *practically meaningless*. We must consider the **power** of the test we are running.

In sum…

Statistical significance tells us that the differences that we found are unlikely to be due to chance or luck, but possibly not. And, even if the differences are real, it may not mean anything in the real world. And that is why, whenever we run a statistical test, we should also compute measures of effect size, confidence intervals, and power.

===

Questions:

1. In a recent study, scientists have demonstrated statistical significance that injecting 20 rats with heroin increases their running speed compared to 20 rats with saline solution. What conclusions might you draw from this finding?
2. What are some potential limitations of this study, and what suggestions would you offer to address them?