2022\_05\_27-Learning objectives

For module (1) Logic of hypothesis testing, errors of inference, power, interpretation of statistical significance and confidence intervals

Explain how to correctly use the term "effect."

Research is often characterized as examining whether or not some factor affects some outcome. An example would be comparing the benefits of a new drug to the benefits of an old one. If there is a difference in the benefits, then that is sometimes called "the effect of the new drug". But that is not exactly correct. The "effect" isn't just the benefit of the new drug, because both drugs may have some benefit. What matters is if there is a difference between those benefits. The difference could be that the new drug is better, or that the new drug is worse. Of course, in this case, the researchers would probably want to see an advantage for the new drug. That is, they would want to see the new drug providing more benefit compared to the old drug. But regardless of whether the new drug is better or worse, because it is the difference that matters, that difference should be called the "effect of the type of drug." Then, if you need to refer to whether the new drug was better or worse, you could refer to "the beneficial effect of the new drug," or "the detrimental effect of the new drug."

Distinguish between an effect in a sample and a true effect in a population.

The researchers testing the drugs would like to be able to take the whole population of people who use that drug, randomly assign them to get either the old drug or the new drug, and then compare the outcomes. But it is generally impossible to test an entire population. The researchers have to select some volunteers from the population and use them as a sample in their experiment. There is a problem for the researchers that arises because there is variability in the population. For some individuals, the benefit of either drug is going to be different than for other people. If the researchers see an effect of the type of drug in their experiment, it could be because there actually is an effect of the type of drug. That would be called a true effect, that is, an effect in the population. But it could also be that they would have seen no effect if they had been able to test the whole population. The effect in their experiment could have occurred only by chance. That is, maybe when they randomly assigned the individuals to the old and new drug, they, just by chance, assigned more people who would have benefited more from either drug to the new drug, and more people who would have benefited less from either drug to the old drug. Therefore, an effect in the sample doesn't necessarily mean that there is any true effect in the population.

Explain the logic of hypothesis testing.

Fortunately for researchers, even though an effect in a sample doesn't prove that there is a true effect in the population, there is nevertheless a way to come to a conclusion, with some amount of confidence, about whether they should believe there is a true effect in the population. This belief rests on how unlikely it is that an effect of some size would occur in the sample if there were no effect in the population. The researchers can start by saying, let's suppose there is no effect in the population. That's called the "null hypothesis." Even though the null hypothesis might never be absolutely true, we can still use it as the logical basis for our test. The reason we can do that is because, hypothetically, if the null hypothesis were true, then when we look at our sample, even if we see an effect, it should be only a small effect, and it should be only due to the variability in the population. If we see a larger affect, that is, large in relation to the amount of variability in the population, then we know that, even though that is possible, it would have been unlikely if the null hypothesis were true. The larger the effect we see in the sample, the more unlikely it is that it would have occurred if the null hypothesis were true. But if we see with our own eyes something that we know would be extremely unlikely if a certain hypothesis were true, then we can logically doubt the truth of the hypothesis. In other words, the evidence goes against the null hypothesis. That doesn't mean we have proven that the null hypothesis is false, but the stronger the evidence, the greater the doubt.

Imagine a police officer who pulls over a driver that is driving erratically. The police officer suspects that the person was drinking alcohol. If the police officer had been in the bar where the driver was indeed drinking, and the police officer watched him the whole time, then he could say that proves this person shouldn't be driving. But when he pulls the guy over, the driver claims he hasn't been drinking. The police officer smells alcohol on his breath. The driver says someone spilled beer on him. The police officer sees open bottles of beer all over the passenger seat. The driver says someone else put them there. The police officer says the driver is slurring his speech. The driver says he always talks that way. Finally, the police officer gives him a breathalyzer test that shows .154 (.08 is the legal limit). The driver says that the breathalyzer could be wrong because it's been used since it was last calibrated. The driver insists that the police officer can't prove he's been drinking too much to drive because the officer didn't see him drinking. But the officer knows that it's extremely unlikely that a person who was perfectly capable to drive would smell of alcohol, slur their speech, have empty beer bottles next to them, and get a .154 on a breathalyzer. Therefore, the officer has a strong doubt that the driver's explanations are true. It's strong enough that the driver is arrested for DUI – and convicted.

The driver's claim that he hasn't been drinking is like a null hypothesis. All the evidence that the police officer observes is like a large effect in a sample. The evidence doesn't prove that the null hypothesis is false, but it casts great doubt on it. The doubt occurs because of how unlikely the evidence would be if the null hypothesis were true. Don't make the mistake of saying that it is unlikely that the null hypothesis is true. The only probability here is the probability of the evidence. Doubt is not the same thing as a low probability. You doubt the truth of the null hypothesis not because the truth of the null hypothesis is unlikely, but because the presence of the evidence would be unlikely if the null hypothesis were true. And you can't doubt the evidence, because you see it with your own eyes.

Explain why the fact that it might be the case that the null hypothesis is never true, is not a problem

Some people point out that even before you do a hypothesis test it's very doubtful that the null hypothesis is ever actually true. For example, suppose there is a population of people who take a certain drug for a medical condition. Everyone who takes the drug gets some benefit, but there is variability. That is, some people get more benefit, and some less. If there is a new drug that, in fact, is no better than the old drug, people taking that new drug will also benefit with some variability. Therefore, if you could test the whole population by randomly assigning half of it to the old drug and the other half to the new drug, the average benefit of both groups might be close to the same, even very very close. But if you carried out the calculations of the two averages to enough decimal places, you would probably at some point see a difference. The difference might be tiny, but it still means that, technically, the null hypothesis is not true. Fortunately, that's not a problem, because the logic of a hypothesis test does not rely on the necessity that the null hypothesis is actually true. It only relies on the logic of the implications that would arise if, hypothetically, it were true.

To see that it is perfectly possible to use the implications of something that is only hypothetical as a way to draw a conclusion, consider this example. A developmental psychologist conducted this study. Second graders and eighth graders were given the following scenario. Suppose you are walking down a country road, and out in a field you see a herd of flying buffalo grazing. You know that if a flock of birds flies over your head, especially if they have just eaten, you are likely to get a bird dropping plopped on your head. That's gross enough. But imagine as you are walking by the grazing herd of flying buffalo, they suddenly take off and begin flying overhead. Would you want to be standing right underneath where those buffalo were flying? The eighth graders generally answered, "Eww! It's bad enough to get a bird dropping on your head, but imagine if it were a buffalo!" The second graders generally said, "Buffaloes can't fly." In other words, they had not developed enough cognitively to realize that just because the scenario was hypothetical, and in this case, not only hypothetical but counterfactual as well, you can still draw out the implications that would arise if it were true. You can do that even if it were obviously not true. You can still use the logic of what would happen if it were true to draw a conclusion.

Explain why "statistically significant" does not necessarily mean "practically important."

When a sample is used in a null hypothesis significance test, the sampling distribution of the sample statistic (for example, the distribution of the mean or the distribution of mean differences) is converted to a probability distribution of a test statistic ( for example, a *z* or a *t* distribution). Then, the value of the test statistic for that sample is calculated, and the probability distribution is used to determine how unlikely that test statistic would be if the null hypothesis were true. The larger the absolute value of the test statistic, the more unlikely it is. A low probability value, called alpha, (for example, only .05 or .01) is selected to use as the criterion to be used to decide if the sample result is unlikely enough to stand as good evidence for doubting the null hypothesis. If the criterion is reached, that is if the probability of the test statistic from the sample is less than the select alpha (for example *p* < .05, or *p* < .01), then that casts enough doubt on the null hypothesis so that we say we reject it. When we reject the null hypothesis, we say the result is "statistically significant." We say that when the size of the effect in the sample is large. But it is not correct to say that "the effect is statistically significant," unless we specify whether we are referring to the effect that we see in the sample, or the effect that we are concluding exists in the population. When we say something is "significant" we often mean that it is important. But when we say an effect is "statistically significant" we are referring to the effect in the sample. That effect in the sample is our evidence. We are saying that the evidence from the sample is significant, that is, it is strong enough to support our conclusion about the effect in the population. But what is our conclusion about the effect in the population? Unfortunately, one way to explain what the conclusion is requires a double negative. It is that we conclude that there is not no effect in the population. We can convert that to a positive statement by saying we conclude that there is an effect in the population. But we must remember that "not no effect" only means "an effect" which actually only means "at least some effect." It says nothing about the size of that effect. Therefore, it tells us nothing about whether the effect in the population is large enough to have any practical importance.

Those who point out that it is very doubtful that the null hypothesis is ever actually true, are saying that it is very believable that there is always at least some effect. Therefore, if we get a "statistically significant" effect, our reaction should be: “Of course. There's little doubt that there's always at least some effect.” And when we fail to get a "statistically significant" effect our reaction should be it's very doubtful that we came to the right conclusion. Some people say that this conundrum shows that null hypothesis significance testing is useless. However, it's not entirely useless, but it's only a first step. Obviously we need something more. We need something that allows us to draw a conclusion about not only whether there is "at least some effect" in the population, but how large that effect might be. Fortunately, we can use the same logic that we used to decide whether to reject the hypothesis of no effect at all, to see if we can confidently reject a whole range of hypotheses about how large the effect might be. On the reverse side, that enables us to confidently say what range of hypotheses we can not reject. In other words, that is the range of hypotheses about the size of the effect in the population that we can be confident about.

Suppose the researchers in our drug example know that the medical problem for which the drug is a treatment is due to having too little of some substance in a person's blood. There is a test to measure how much an individual has. Suppose the measurements could range from 100 to 900 of some unit of measure (for the purposes of this example it doesn't matter what the unit of measure is.) Let's say that a good number is 500, but anywhere between 400 and 600 is okay. The old drug raises the number, on average, 50 points. Therefore, any person with a measurement of anything above 350 will benefit from the drug. But some people are down close to 300. Therefore, the researchers hope that the new drug will raise the number 100 points. That is, they hope that the new drug is 50 points better than the old drug.

In the experiment that they conduct with a sample of patients, they see a 60 point difference (in the right direction) between the drugs. They do a significance test, and find that the difference is statistically significant. By itself, that only means they can be confident that in the population there is at least some true effect. But that only means they can be confident that the true effect in the populations isn't zero. Now suppose that they could do another hypothesis test, but this time instead of testing against the null hypothesis, they test against the hypothesis that there is a 1 point effect. If the evidence from the sample still shows that the sample effect would only have a less than alpha probability of occurring if the 1 point effect were true, then they can reject that hypothesis. Suppose they test against an effect of 2 points, then 3 points, and so on, and each time they find that they can reject the hypothesis they are testing. However, when they get up to testing against an effect of 55 points, the evidence is not strong enough to reject that effect. That means that they can be confident that, even though the true effect in the population isn't necessarily 60 points, as it is in the sample, they can be confident that it is at least 55 points. As a matter of fact, if they continue testing larger and larger effects, they will find that they can not reject the hypothesis that the true effect in the population is 61, 62, 63, 64, or 65 points. When they tested against the hypothesis that the true effect in the population was anywhere below 5 points lower than the 60 point effect that they saw in the sample, they could reject those hypotheses. And when they tested against the hypothesis that the true effect in the population was anywhere above 5 points higher than the 60 point effect that they saw in the sample, they could reject those hypotheses as well. The logical conclusion, then, is that they can be confident that the true effect in the population is somewhere between 55 and 65. That's called the confidence interval. Fortunately, the calculation for the lower and upper boundaries of the confidence interval is one calculation. We don't actually have to do a whole series of hypothesis tests. But it is important to understand that when we calculate a confidence interval, the logic used is the same logic that we use for the single null hypothesis significance test. That will help us avoid some misunderstandings. For example, we need to remember that, just as we don't prove the null hypothesis is false, we haven't proven that the true effect in the population is in the confidence interval, we could have made an error of inference. We can't even say that it is likely that it is in the confidence interval, because the only probability we can work with is the probability that the sample result would have occurred if some hypothesis were true. And if that probability is low, it stands as *evidence* against that hypothesis, and if the probability is not low, it does not stand as evidence against that hypothesis. Because we can't talk about the probability of the hypothesis, only of the sample evidence, we use the word "confidence" to talk about the range of hypotheses that we can't reject. We say that we are confident that the true effect is in a range bounded by some lower and upper value.

Explain the importance of a confidence interval

Explain what Type I and Type II error are

Explain what power is

We can use the confidence interval to solve the problem that the hypothesis test only tells us whether we have strong enough evidence to conclude that there is "at least some effect," in the population, which doesn't tell us anything about the practical importance of that effect. Before we even conduct the hypothesis test, we can decide how big the effect in the population would have to be to be important. We can call that the "minimum practically important" effect (MPI). That decision can not be made by statistics. It has to be made on the basis of our knowledge about the specifics of whatever we were testing. But once we have decided on the MPI, we can proceed by first doing the hypothesis test. Of course, if the test leads us to reject the null hypothesis, in reality, we could either be correct in concluding that there is at least some effect, or, because we haven't proven that the null hypothesis is false, we could have made a Type I error. The Type I error means that we concluded that the null hypothesis should be rejected, when, in fact, the null hypothesis is true. Of course, as mentioned earlier, some people say that the null hypothesis is never actually exactly true. If you take that position, then you are saying the you never really make a Type I error. That emphasizes the importance of understanding the importance of having the confidence interval to allow us to draw a conclusion about not only whether there is "at least some effect" in the population, but how large that effect might be.

After we have determined whether the sample result is or is not statistically significant, we can proceed to calculating the confidence interval. We need to determine where the boundaries of the confidence interval are in relation to the MPI. Even if the result is statistically significant, if the MPI is in the confidence interval, then all we can conclude that we can be confident about is that there is "at least some effect, but we still don't know if it is large enough to have any practical importance."

If the whole confidence interval is above the MPI, then we can conclude that we can be confident that there is "at least an effect large enough to be important." However, because we can't say that we have proven that the effect is large enough to be important, we have to realize that we could still be wrong. That is, we might not be wrong about there being at least some effect, but we might be wrong about it being large enough to be important. In other words, even those people who believe that the null hypothesis is never really true could call that their version of a Type I error.

If the whole confidence interval is above the null hypothesis, but below the MPI, then we can conclude that we can be confident that there is "at least some effect, but the evidence supports the conclusion that it is not large enough to be important." Nevertheless, again, we still have to realize that we could be wrong. That is, we could be wrong about the existence of at least some effect. That is called a Type II error. Or, if you are one of those people who believe that the null hypothesis is never really true, so, according to your belief, it's impossible to make a Type II error, then you still have to realize that you could be wrong in another way. Because you haven't proven that the whole confidence interval is above the null hypothesis, but below the MPI, even if you are not wrong in concluding that there is at least some effect, you could be wrong about concluding that it is not large enough to be important. In other words, even those people who believe that the null hypothesis is never really true could call that their version of a Type II error.

If the confidence interval is large enough to include both the null hypothesis and the MPI, then we haven't found strong enough evidence to tell us anything. There could be no effect in the population, a negative effect, an effect, but not large enough to be important, or an effect that is large enough to be important. This shows that the evidence from the confidence interval is much more informative than the evidence from the null hypothesis alone. But that is only true if the confidence interval is small enough to exclude both the null hypothesis and the MPI.

Explain the relationships among errors of inference, power, and the size of a confidence interval

How do researchers do as much as they can to ensure that their confidence intervals are small? Most of the same factors that increase the power of a test, also help to make confidence intervals small. The power of the test is the probability that a given sample will result in a statistically significant result, if there really is a true effect of some given size in the population. Power can be increased by decreasing the variability in the experiment by exercising as much control over extraneous sources of variability as possible. Power is also increased by increasing the size of the sample. Both of those factors increase power and decrease the size of the confidence interval. The size of the confidence interval can also be decreased by accepting a smaller level of confidence. The level of confidence is related to the size of the alpha. If the alpha level is .05, then the level of confidence calculated using that alpha level is .95. If the alpha level is .01, then the level of confidence calculated using that alpha level is .99. In other words, the level of confidence is 1 minus the alpha level.

Because it is desirable to have a small confidence interval, but it is also desirable to have high confidence, it is not desirable to decrease the size of the confidence interval by decreasing confidence. In addition, if you want the level of confidence to be consistent with the alpha level you are using for the hypothesis test, decreasing the size of the confidence interval by decreasing confidence also leads to a larger alpha level. The alpha level determines your prior probability of making a Type I error. For these reasons, it is a better choice to decrease the size of the confidence interval by increasing the sample size and exercising as much control as possible. Control can only be achieved to a certain extent. And there is no way to accurately calculate how much any given control measure increases power. However, as long as you can specify, in terms of Cohen's d, the minimum true effect in the population that has practical importance, and you can specify how much power you want, you can calculate the size of the sample that gives you that much power. Also, you don't need to know how much you have to decrease the variability to get that much power because Cohen's d takes into account whatever the variability is.

Because the alpha level determines your prior probability of making a Type I error, and you would like that to be low, it makes sense to select a low value of alpha, such as .01 as opposed to .05. However, all other factors being equal, the lower the alpha level, the lower the power, which you would like to be high. The probability of making a Type II error is 1 minus the power. So making alpha lower, which decreases your probability of Type I error, and which also makes the power lower, also increases the probability of making a Type II error. Therefore, the best strategy is to use increasing the sample size to increase power while keeping the apha as low as possible. If you are using an alpha of .05, then you have only a .05 probability of making a Type I error. If you can use a large enough sample to get .95 power, then you also only have a .05 probability of making a Type II error. Of course, to accomplish the same thing using an alpha of only .01, you would have to take a large enough sample to get .99 power, which might not be feasible. Nevertheless, by using this strategy, that is, only increasing power by increasing your sample size, you can get the best combination of as low as possible probability of Type I error and as low as possible probability of Type II error.