The Transition from Descriptive Statistics to Inferential Statistics

Elliott Hammer Xavier

University of Louisiana New Orleans, Louisiana

Overview

One of the reasons that other sciences sometimes disparage psychology is that we often admit that we can't "prove" support for our theories. Proof is virtually impossible for psychology researchers to attain because controlling enough variables to isolate a relationship definitively is, well, virtually impossible. In other sciences, researchers can regulate conditions for their experiment to a greater degree than can researchers studying behavior. Chemicals rarely "have a bad day," and the structure of a cell is pretty much the same for one person as for any other. People, however, (and nonhuman animals, for that matter) operate much less in a vacuum than do the subjects of lab research for biologists, chemists, and physicists. As a result, psychologists are subject to issues of mood, relationships, traffic, and the like that affect the performance of our participants.

Assessing Dependability

In the absence of proof, the claims that psychologists make must rely on assessment of the likelihood that one's results are dependable. That likelihood never reaches 100 percent, but it can get very close. How close is close enough to count on a finding is a matter of some debate, but most researchers agree to a reasonable degree on 95 percent; such findings would mean that I would have only a 5 percent chance of being wrong in a claim that I might make about a relationship between a couple of variables. I'll discuss where this number comes from in a bit, but keeping that 5 percent value in mind can be helpful. First, let's consider a bit further why proof is unattainable and how we can assess the dependability of our results. Consider a researcher who wants to know if a new way of teaching math can help students learn math more effectively. The first thing that researcher might want to do is to assess the math proficiency of the general population. Even this step presents a bit of a problem in that it would be impossible to get every member of a population to take a math test. So, what

would the researcher do? First, he or she would gather a sample from the population that he or she hopes to make a statement about, and that sample would serve to represent the population for the purposes of the study. This sample would take the test and provide something of a baseline. Does that baseline perfectly match the population? Probably not, but it's the closest we can get, considering the circumstances. Obviously, the bigger and more random our sample is, the more closely it resembles the population, and the more dependable that sample's mean score is going to be. This effort to gather a sample to represent the population is indicative of our goals in conducting such research. We typically don't care about our sample in and of itself. Instead, what we care about is the population that that sample represents. It's one thing to make a statement about a group of 50 or 1,000 people, but it's another to be able extrapolate what we learn about them to the greater population, who are those people who didn't participate as part of our sample. This is what inferential statistics is all about. We infer what the population is like, based on what we know about our sample. Note that we actually can prove something about our sample. If we find that our sample averaged 47.5 correct answers on the test, we have proof about the sample mean. We can be sure how many questions each test-taker got right in the sample (as long as we calculated correctly). What we can't know for sure is what this says about everyone else— the people who weren't in the sample. So we take what we know (sample statistics) and infer what we don't know about the population (inferential statistics). What we're doing is estimating what the population's values would be, based on what the sample is like. This process is called parameter estimation because a parameter is a measurement for the entire population, as opposed to statistics, which refer only to samples. Because we can rarely know what the population is like, we estimate, through our inferential statistics.

Using Inferential Statistics

Once we have a mean for our sample, that mean can serve as our best guess of what our population is like; we have no reason to assume that our sample mean over- or underestimates the population mean. So, if our sample mean is 47.5, our best guess of the population mean (parameter estimation again) is 47.5. Why do we care? Well, this estimated mean might be helpful, but when it really comes in handy is when we're trying to do some research, for example, to see if the new

way of teaching math is any good. We would predict that the sample who goes through the treatment (the new technique, say, teaching math through pictures instead of numbers) will have a higher mean than the assumed population mean of 47.5. But how much higher would the mean have to be in order to be meaningful? Consider that even if the treatment does absolutely nothing, the treatment group will still almost certainly not score exactly 47.5 because of a bit of chance fluctuation or mild irregularities in the sample (called sampling error), such as some especially lucky guesses. In fact, if we are reasonably certain that the group's mean will not be 47.5 exactly, there's a 50 percent chance that it'll be higher, and a 50 percent chance that it'll be lower. So we can't just look at the mean and say, "Oh, the treatment sample's mean is 47.8, so the treatment works." Most of us would not be inclined to intuit that such a difference was important, so we can accept that. However, we also can't look and say, "Oh, the treatment sample's mean is 85, so the treatment works." Regardless of how big the difference is, we need to see how likely that sample's mean is to have occurred even if the treatment did nothing. So how do we assess the likelihood that the score from our sample would have occurred simply by chance and not because the treatment has an effect? Well, first we have to establish our null hypothesis, which is our assumption that the treatment has no effect. We hope to reject this null hypothesis; doing so would support our research (or, alternative) hypothesis, which is our real prediction in the study. Our statistical calculations of the probability of rejecting the null hypothesis usually depend upon several factors. The one that we typically have the most control over is the sample size, symbolized with N. Most statistical formulas adjust so that the larger the N is, the larger the statistic we are calculating; larger statistical values are more likely to be beyond our critical value (so called because it's the value that our statistic must reach in order to be considered significant), leading us to reject our null hypothesis. We therefore would like to have a fairly large sample, despite obvious limits on how many people we can get into our sample. We can also enhance our ability to reject the null hypothesis by having a fairly strong manipulation and fairly sensitive measures. Now let's get back to the .05 value discussed at the beginning. In research, the number commonly refers to our alpha level, which the researcher sets; most researchers are willing to set alpha as high as .05, but typically not higher. We can think of the value as the probability of rejecting the null hypothesis when we shouldn't; to do so would be committing a Type I error

(or alpha error, for obvious reasons). Clearly, we would like this probability to be as low as possible because we would like to avoid committing a Type I error. So why not set it even lower? We usually don't set it lower at the outset because there is a major trade-off. If, for example, we set alpha very low (say, .0001), we obviously have a very low probability of committing a Type I error. That's good. But, we also make it extremely difficult to reject the null hypothesis at all because the critical value in such a situation is so large; our calculated value (r, t, F, etc.) would have to be untenably high in order to be statistically significant. It's so high and difficult to obtain that we may miss some real effects, and we don't reject the null hypothesis even if we should. Missing those real effects is called committing a Type II error (or beta error). By setting alpha at .05, we are liberal enough that we are confident about finding an effect if one is really there, but we're conservative enough that we'll be wrong only 1 out of every 20 times that we reject the null hypothesis. Given the trade-offs, most researchers abide by this convention. The table below summarizes how our decisions to reject or retain the null hypothesis can be correct or incorrect, given the true situation that we are trying to infer.

		In Reality	
		Null Hypothesis is True (treatment doesn't work)	Null Hypothesis is False (treatment does work)
Decision Based on Statistical Information	Reject Null Hypothesis	**Incorrect Decision**; Type I Error (probability = alpha)	Correct Decision
	Retain Null Hypothesis	Correct Decision	Decision Incorrect Decision; Type II Error

What does all this have to do with the proof problem that elicits disrespect from some researchers in other scientific disciplines? In reality, the null hypothesis is either true or false. That is, the treatment really does have an effect, or it doesn't. Two variables are either really related to one another, or they aren't. How do we know? We don't, because the null hypothesis pertains to the population, which we'll never know for sure. As a consequence, when we make our decision to

reject the null hypothesis or not, we can't be sure we're making the right decision. The right decision for a false null hypothesis would be to reject it. However, we make this decision based on our sample, not the population, so rejecting the null hypothesis is a bit of an educated guess (based on those probabilities) of what the population is like. It's not the same as knowing what the population is like. Similarly, when we retain the null hypothesis, we don't know for sure if we've made the correct decision either. Why? Once again, because we don't know whether the null hypothesis is, in fact, true, which is our assumption when we retain the null hypothesis.

Conclusion

In reality, we may not be able to prove our predictions, but we can determine how reliable our results are. If a researcher set a reasonably low alpha level and conducted the study appropriately, then critics would not have a very strong case in countering the findings. So even in the absence of proof, we can support the conclusions to which our data lead us, and we can determine the likelihood of our being wrong. This absence of proof also serves as encouragement for replication, which further strengthens our claims. Just as a large sample can be more convincing than a small one, a lot of studies can be much more convincing than one.