Statistics is for Everyone (*Yes, Even You*): Straightforward Statistical Concepts to Help Instruction Librarians Help Their Students (Part 1)

Savannah L. Kelly, University of Mississippi

Last semester I found myself working with a student at the reference desk who was writing a research paper on test anxiety for his composition course. He chose an article from the *Journal of Educational Psychology* entitled "Test anxiety and academic performance in undergraduate and graduate students." The title seemed promising but the student became discouraged when he realized that statistics were incorporated throughout; he told me that as an English literature major, he hated numbers. I told him not to worry, that I could help—it's not as hard as it looks.

I wasn't always so confident, and as library professionals, we may feel the same way as that student. It is relatively uncommon for instruction librarians to have had formal training interpreting statistical analyses. Personally, I had a single course—educational statistics—for my bachelor's degree in psychology, *sixteen years ago*, and no stats or math-related content as a graduate student in library science. For librarians, this lack of experience with statistics may position us at a disadvantage when we have to guide students through the academic literature in the numerous disciplines where statistics are methodologically essential. I have worked with many students over the years who are completely unfamiliar and intimidated by this landscape.

Prior to my learning about statistics, my advice to students was one of avoidance—just ignore the method and results sections and skip the analyses, I'd say. What I came to realize, however, was that this sent a dangerously dismissive message: either quantitative analysis was inconsequential, or more likely, just too difficult for "regular" people to understand. As students from a wide-variety of majors continued to need help interpreting quantitative articles at the reference desk, in the classroom, and in research consultations, I found myself reconsidering my avoidant approach. And yet, I was just as mystified as the students: What did those Greek symbols mean? What exactly was a *p*-value anyway?

To mitigate my own confusion, I decided to enroll in a graduate-level statistics course a couple of years ago at my university. I thought surely I could learn something beneficial if I just took Statistics 1. I have since completed six statistics courses and will enroll in a seventh course next semester. As it turns out, statistics is not nearly as difficult as I had feared. Most importantly, this new knowledge has helped me help my students.

What I learned throughout this process was that the more familiar I became with statistics and grasped its basic underlying principles, the less intimidating and challenging it felt. The purpose of this article is to share some of those foundations that helped me—someone without a numbers background—feel more comfortable with the field of statistics, and to discuss how that knowledge has enhanced my ability to relate to both my students and to the quantitative literature.

Two Fundamental Distinctions

Descriptive vs. Inferential Statistics

A critical starting point for understanding statistics is to recognize that there are two interrelated branches: descriptive statistics and inferential statistics. The purpose of descriptive statistics is to describe what you already know about something, whereas inferential statistics is about trying to make the leap from what you know to what you do not know. For example, if I had a randomly-selected group of thirty students, I could describe a lot about them: their mean (average) score on an assignment, the median (middle point) height of students, and their modal (most frequently occurring) hair color. These descriptive statistics are all measures of central tendency (mean, median, and mode) that would help me summarize what I already have in front of me. However, if I wanted to know if these measures that I observed in my group of thirty students were representative of the entire population of thousands of students, that's when I would need to consider the realm of inferential statistics.

Before we consider how inferential statistics work, we need to visit another important concept in descriptive statistics: variance. Most people have heard the term standard deviation, but what is less known is that standard deviation is simply the square root of the variance. Variance describes how far your data is from a central point (usually the mean). You can think of the standard deviation as the average amount of variance; the larger the standard deviation, the more spread the data is from the central point; the smaller the deviation, the more clustered the data is around the central point. Descriptive statistics uses these central points to make estimations, so if there is a lot of variability in your data (e.g., large standard deviations) then it can be more challenging to compute accurate inferential statistics.

A final difference between descriptive and inferential statistics is that you can calculate descriptive statistics without wanting or needing to calculate inferential statistics, but it doesn't work the other way around. This is because inferential statistics are based on descriptive sta-

tistics, which allow you to estimate from the sample you have to the population you don't have.

Sample Statistics vs. Population Parameters

In order to communicate the estimation process used in inferential statistics, I need to clarify the distinction between a sample and a population, and a statistic and a parameter. A population is what you really want to know about, whereas the sample is a smaller subset of that population. If you had unlimited access, time, and resources to measure the full population that you want to know about (e.g., all undergraduate students at R1 universities; all graduate students who were mothers; all students who use academic libraries), then you don't need inferential statistics, because you can ask the entire population. If, however, you need to work with a smaller group (as is typical because time and resources are limited), then you have to figure out if your sample can be generalized to the larger population of interest.

This article is too brief to discuss sampling designs, but it is important to note that the best sample for generalizability is a random one. The term random sample gets tossed around a lot. A *random sample*, unlike its name, is actually not random at all—it's a planned sample where each individual in the population has an equal chance of being placed in the sample. That's why, if an appropriate sampling design is followed, a seemingly small sample of undergraduates (250) can reasonably represented a much larger (10,000) undergraduate population at a university. A properly-designed random sample is based on probability and it ensures that differences between individuals are equally distributed across groups. Importantly, it randomly distributes natural error (i.e., the difference between sample and population values) across groups.

Another term that is often confused is *statistics*. This is where things can feel a little tricky, because, technically, a *statistic* refers to something from a sample, while a parameter refers to that same thing, but in a population. For example, the symbol $\bar{x}$ (literally "x bar") represents a sample mean, while μ(pronounced "mu") represents a population mean. In inferential statistics, you are using the sample statistic (e.g., $\bar{x}$) that you have as an estimate for the population parameter (e.g., μ) that you do not have. Let's say I ask a random sample of students how many ounces of coffee each person drinks the Tuesday before finals. I might get a sample mean of 35 ounces—that's my $\bar{x}$. Now I'll use that sample mean to estimate an accurate population mean (e.g., μ) so that I can be reasonably sure to order enough coffee for the entire campus for that particular Tuesday. This is what inferential statistics is all about, and demonstrates how inferential statistics are based on descriptive statistics. Now, let us consider how this information can help us approach the quantitative literature.

Decoding Quantitative Literature

Hypothesis Testing, P-Values, Confidence Intervals, and Statistical Significance

When it came to understanding quantitative literature, the very first thing I wanted to know was the meaning of a *p*-value (e.g., $p < 0.05$). This is something you and your students will come across in almost all quantitative articles. In order to make sense of a *p*-value, you have to understand two things: inferential statistics and null hypothesis significance testing (NHST). Recall that the process of inferential statistics is when you use a sample to make estimations about an unknown population; hypothesis testing is part of that estimation process. It seems somewhat counterintuitive, but the beauty of statistics is that when you create a hypothesis, you are making a hypothesis about the larger population, not the sample. And when you test the hypothesis, you are testing this unknown population based on what you have in your sample. The underlying theory can get complex, but what you need to know is that these hypotheses tests are essentially informed guesses about what you don't know (population parameters) based on what you do know (sample statistics).

A research hypothesis is a prediction by a researcher about what a particular outcome might look like. For example, if you are comparing two groups—students who are exposed to library instruction and students who are not—then your research hypothesis is that the outcome (e.g., information literacy skills) will be different between the groups. But that's not what you test. In statistics, you do not directly test the research hypothesis; you test the straw man, the null hypothesis, and see if you can knock it down. I like to think of the null hypothesis as *nothing's going on here*. In the previous scenario, you would test whether the outcome is *no different* between the groups. And you want to be able to determine whether you can reject that null hypothesis. This is what is referred to as null hypothesis significance testing and it predominates the behavioral & social sciences literature.

Here's where the *p*-value comes in. The *p*-value is related to the null hypothesis. Historical convention dictates that most researchers employ a *p*-value threshold at 5%, which is why you often see this in the literature: $p < 0.05$. The *p*-value is about probability. Here's what it means: the *p*-value is the probability that you will get the observed results you have (or more extreme) *given that the null hypothesis is true*. If the *p*-value is less than your predetermined threshold (0.05 or, if you want to be even more strict as is increasingly the case, 0.01 or 0.001), then you reject the null hypothesis and lean towards your

(Statistics is for Everyone…Continued on page 7)

(Statistics is for Everyone...Continued from page 3)

research hypothesis. You can think of this outcome as: *that's a low probability that I would get these results if nothing was going on here*. When the outcome of a *p*-value (e.g., 0.04) is less than the researcher-designated threshold (e.g., 0.05), the researcher considers that finding *statistically significant*.

Before we consider the implications of statistically significant findings, I want to take a quick detour to address a common misunderstanding. Let's return to the previous example of students receiving library instruction. If you were comparing the difference in means between these two groups (i.e., samples) of students and the instruction group had a mean score of 3.65 for information literacy skills and the control group had a mean of 3.40, you might be inclined to assume that there is a statistically significant difference in means because 3.65 is clearly *not* 3.40, and therefore you do not need inferential statistics. Although it is true that 3.65 is different from 3.40, remember that a hypothesis test is comparing the means in the two unknown *populations*, not between the two known *samples* you have, and your samples have random error. Statistician Nate Silver's popular work is titled *The Signal and the Noise* for a reason: in statistics, you're trying to find the signal, and your sample has noise. The power of inferential statistics is the ability to recognize that noise (i.e., sampling error) and still be able to estimate and compare the population means.

If a researcher has a statistically significant finding (or not), is that the end of the conversation? In other words, is statistical significance the holy grail? Well, no. The reality is that the ability to find statistical significance is contingent on a number of different factors such as sample size, measurement error, *p*-value specification (e.g., 0.05 vs. 0.01), and statistical power. You may not have statistical significance because your sample size was small, which resulted in a lack of power to detect significant differences. Alternatively, you can obtain statistical significance for very small differences if you have an extremely large sample size. The truth is that you can "get" your results to be statistically significant if you tinker enough – that's called p-hacking and it is unethical. This nuance of additional complexity is key to convey to students, who tend to look for absolutes and easy metrics (e.g., "*p*-value is less than 0.05? Must be a worthwhile study!" is similar to "From a .org site? Must be a good resource!") when they are just learning how to analyze the literature.

The limitations of NHST, which is reliant on *p*-value interpretations, is one reason why researchers have advocated reporting confidence intervals (CI) and effect sizes in research articles. A confidence interval shows a range of values for the estimated population parameter. Consider this: 95% CI [0.14, 2.12]. If this was our confidence interval for the difference between our two groups—those who received library instruction and those who did not—then the confidence interval tells us that the true difference in the two populations is somewhere between 0.14 and 2.12; this is our "margin of error." (Recall that our samples had 3.40 and 3.65, which is a difference of 0.25.). Whereas *p*-values tend to present statistical findings as binary—significant or nonsignificant—confidence intervals allow the researcher to present something that is a little closer to the messy reality of statistics: an estimated interval.

Another important concept to understand to give nuance to a study is effect sizes. An effect size, unlike statistical significance, is an attempt to measure the magnitude of an effect, and indicates practical significance. A finding might be statistically significant (e.g., $p < 0.05$) and yet functionally meaningless. Effect sizes (e.g., Cohen's *d*, Cohen's *f*, and R^2) provide an evaluation of how important the results are in practical terms. Returning to our example, you might discover that there is a statistically significant difference between those who receive library instruction and those who do not, but that the effect size is very small and thus perhaps not a meaningful enough difference to justify the effort to implement that particular type of library instruction. Takeaway: *Don't throw a party over statistical significance; throw a party if you have statistical significance and large effect sizes*.

So how do you get around all of this when working with statistics-illiterate students? Encourage students to think about descriptive statistics as what we know and inferential statistics as *using what we know to figure out what we don't know*. Null hypothesis significance testing, *p*-values, and confidence intervals (CI) are ways that we try to make sense of that inferential leap from sample to population. Remind students that statistical significance is not the holy grail, and that effect sizes are often more practically relevant. These distinctions were the foundational concepts that I found most helpful during my statistical training, and I hope that they can provide some guidance to other instruction librarians who, like me, might have felt intimidated by the quantitative literature.

Statistics is a fascinating discipline, and you'll find that there's much more to explore than what was covered here. There will be a second, forthcoming *LOEX Quarterly* article that will dive into variable measurement types (interval vs. discrete) and associated statistical tests (e.g., t-test, ANOVA, regression). I hope you'll hang around for that!