Statistics is for Everyone (*Yes, Even You*): Straightforward Statistical Concepts to Help Instruction Librarians Help Their Students (Part 2)

Savannah L. Kelly, University of Mississippi

Welcome back to the second article in a two-part series on understanding statistical concepts for instruction librarians. If you are unfamiliar with the first article you may want to review it briefly as this second article will build on the topics discussed previously: descriptive vs. inferential statistics, samples vs. populations, and null hypothesis significance testing (NHST). The purpose of this second article is to help academic librarians conceptualize how data corresponds (or not) to particular statistical tests. Essentially, how does your data influence your statistical analysis options? This idea impacts everyone—whether you are doing the research yourself, or if you are helping a student analyze a paper and wondering why the paper's authors chose the test that they did.

In general, the types of questions researchers use statistical tests for are basically endless—from large-scale with copious amounts of numbers (e.g., "How does the attainment of a college degree affect occupational wage trajectories?") to more localized with fewer numbers (e.g., "Did the creation of a service-learning component impact students' engagement with the biology curriculum at this particular university?"). In our field, some common questions are "Do students perform better on an English assignment if there was a librarian intervention beforehand?" or "Is there a relationship between how many books a student checks out of the library and the types of citations the student uses in her research paper?" Regardless of the question, you want to understand some principal concepts prior to selecting and running your statistical tests.

Prior to Statistical Analysis

The first issue to consider is the relationship between your research question and the available data. It's important to know up front what you are trying to investigate. For example, you might want to know if there are differences between groups, or how one variable impacts another, or if your survey is adequately measuring an underlying latent variable (e.g., students' self-efficacy). What you want to know will directly influence the type of statistical test you ultimately select. Here are a few types of questions in our field that can be answered using common statistical tests that will be discussed later in this article:

- Is there a difference in information literacy scores between students who had library instruction this semester and those that did not (t-test, ANOVA)?
- Can I predict what the library collections budget should be at our institution based on other institutions in our consortium (regression)?
- What is the relationship between students' self-reported engagement scores and their final class grades (correlation)? Can a student's engagement score predict his ability to pass (or not) the class (logistic regression)?
- Do students' classification—first-year, sophomore, junior, senior—and their declared major interact in influencing students' critical thinking scores (factorial ANOVA, moderated regression)?
- Do the items in my survey measure an important underlying latent variable (EFA, CFA)?

The second issue to consider is a little more complex. The type, the number, and the distribution of each variable will also affect the statistical tests you can use. For the sake of simplicity, I'm going to break down variables into two types: continuous or categorical. A continuous variable is numerical; a categorical variable is not. So, taking from the examples above, a student's score on an information literacy test is continuous but a student's classification and their major are categorical variables. These are important distinctions because there is much more you can do mathematically (i.e., statistically) with a numerical variable than you can with a categorical variable, and many statistical tests necessitate that you use continuous variables. With that said, it is important to note that researchers will sometimes "dummy code" their categorical variables into numerical variables. This is done by changing categorical variables into binary ones (e.g., for Field A every record for a freshman is given a value of 1, and every non-freshman is a 0; for Field B every record for a sophomore is a 1, and every non-sophomore is a 0; and so on); by doing this, a researcher can expand the number of available statistical tests they can employ.

The number of variables you want to include in your analysis also plays an important role in selecting the appropriate statistical test. Essentially, this is asking about the relationship between the independent and dependent variable(s). Remember that independent variable(s) are what help explain the dependent variable(s). Univariate tests (e.g., t-test) have only *one* dependent variable while multivariate tests (e.g., MANOVA) can have *more than one* dependent variable. As for independent variables, that is much more flexible. It is very common to have multiple independent variables explaining a single dependent variable.

What did I mean earlier when I mentioned the distribution of a variable? This is where it can get a little tricky because we need to talk about statistical *assumptions.* There are parametric and nonparametric statistical tests. Parametric tests, which are the only types of tests discussed in this article, must meet certain requirements before you can run them, while nonparametric tests require fewer assumptions (and are often referred to as distribution-free tests). Many of the statistical assumptions for parametric tests relate to the distribution of variables and their corresponding error variances. One of the most common requirements is the assumption of normality; that is, when you plot the values, the resulting image should look like a bell curve. As an example, if you take a measure of students' information literacy scores and all their scores are "piled up" on one side of the

distribution, then you might have an issue. There are a variety of approaches that researchers take when a statistical assumption is violated. If normality is violated, for example, researchers will often transform the data by taking the log or square of the original scores. Without getting too much in the weeds, however, what you want to remember is this: *Statistical tests come with assumptions about the data. If your data does not meet one or more assumption, then you can try to transform your data to fix it, you can argue that the assumption was not critical for that particular analysis, or you can run a nonparametric test.*

Univariate Tests (and correlation)

Now that I've discussed issues to consider prior to running an analysis, let me get to the fun part. What are the common statistical tests & what are they generally used for?

T-test and ANOVA

An independent *t*-test is used to understand the difference in an outcome between two groups. The two groups can be almost anything—STEM majors or social science majors, morning people or night people, workshop A attendees or workshop B attendees—the important thing is that there are *only* two groups. If you have more than two groups, then you can still run this type of analysis but it's called an ANOVA, which stands for analysis of variance. An ANOVA will look for differences in an outcome between two *or more* groups. Both t-tests and ANOVA have categorical independent variables and a continuous, single outcome variable. Basically what a t-test and an ANOVA do is compare the variability from the mean within the groups compared to the variability between the groups. If you have more variability between the groups than within, then you're more likely to find a statistically significant difference.

Correlation

Although correlation is not technically considered a univariate test, it is foundational to understanding other univariate tests so it is included here. Correlation is a bivariate analysis that examines the strength and direction of a relationship between two continuous variables. The higher the correlation, which ranges between 0 and 1, the stronger the relationship (e.g., 0.78 is a stronger relationship than 0.23). Let's say we are trying to determine the relationship between points scored on a pre-test vs. a post-test—the two variables are students' pre-test scores and their post-test scores. The sign of the number (negative or positive) indicates the direction of the relationship. If it is a positive (+) sign, then both variables move in the same direction; if the sign is negative (-), then as one variable moves in one direction (e.g., up), the other variable moves in the opposite direction (e.g., down). And remember: correlation does **not** imply causation.

Regression

My favorite statistical tool is OLS (ordinary least squares) regression; it is considered the workhorse of statistical analysis. Regression is the statistical method of choice when researchers want to better understand the *relationship between variables*. This is frequently pitted against ANOVA, which is categorized as the analysis you want to use when understanding the *relationship between groups.* What is actually fascinating is that ANOVA is simply a form of regression, but with categorical variables. So why do we think they are different? Because regression and ANOVA were developed separately in disciplinary silos. ANOVA was favored in fields where experimentation was possible (e.g., psychology) while regression was the tool of choice in fields where correlational data was prevalent (e.g., economics, epidemiology). Regardless, they are the same mathematically. Traditionally OLS regression requires that all variables are continuous, but as mentioned previously, researchers will dummy code their categorical variables into continuous ones in order to use regression analysis. You can also use regression for prediction. You can run a regression, then use the equation from the analysis to predict another scenario with different inputs. For example, by performing a regression you could argue for a larger collections budget that is more similar to schools with parallel inputs.

Factorial ANOVA / Moderated Regression

These are the final univariate analyses that I will mention in this article. These type of analyses deal with an *interaction* between independent variables in explaining a dependent variable. What does that mean? When two independent variables *interact* that implies that the outcome of the dependent variable varies across the other independent variable. Here's an example: Let's say you want to understand the relationship between sex (e.g., male, female) and disciplinary branches (e.g., STEM, humanities, social sciences) on students' information literacy scores. Sex and disciplinary branch are the independent variables and the information literacy score is the dependent variable. A factorial ANOVA / moderated regression is when the information literacy scores for males and females differ based on the other independent variable, the disciplinary branch. So it may be that women in STEM have better information literacy scores than men in STEM but men in the humanities have better scores than women in the humanities. The condition is not the same across the variables. This type of analysis is more complicated than the ones mentioned previously, but also more precise.

Multivariate Tests (and correlation)

Now I'm going to give a brief overview of some of the most common multivariate tests you may encounter.

MANOVA

The first is MANOVA, which stands for multivariate analysis of variance and is an extension of ANOVA. Whereas ANOVA has a single dependent variable, with MANOVA, you can have more than one dependent variable. Rather than running multiple ANOVAs, each with a different dependent variable, a MANOVA will allow you to run an analysis with the understanding that your dependent variables are related (and often correlated).

Logistic regression

Logistic regression is an odd little bird. Whereas traditional OLS regression explains a continuous outcome, lo-

(Statistics is for Everyone...Continued on page 10)

campus, and many faculty expressed excitement about using IL resources to build IL into their syllabi and teaching practice. This workshop was held again in 2019, and plans are currently underway to design a "roadshow" version to offer to departmental faculty, chairs, associate deans, and deans.

In addition, because faculty highly value IL for their learners as students and for themselves as professionals, librarians can use this to work with faculty to identify higher-order and highly-relevant IL learning sessions, modules, or resources that demonstrate to departmental faculty the range of IL teaching and learning librarians can support.

Finally, several collaborations and partnerships have arisen from networking opportunities provided by OIE. By continuing this partnership with the campus IRO office, the library has increased capacity to make evidence-informed choices about instruction, outreach, and resource/service development, in partnership with departmental faculty, in support of Fresno State students and their learning.

References

California State University Fresno Office of Institutional Effectiveness. (2018). Data. Retrieved from http://www.fresnostate.edu/academics/oie/data/

Hewitt, G. J., & Hewitt, R. T. (2010). Ability, assistance, and collaboration in academic library assessment. *Library Philosophy & Practice*, 1-6.

Kennedy, M. R., & Brancolini, K. R. (2018). Academic librarian research: An update to a survey of attitudes, involvement, and perceived capabilities. *College & Research Libraries, 79*(6), 822.

Ofori-Attah, K. D. (2002). Institutional research in higher education. In J. W. Guthrie (Ed.), *Encyclopedia of Education* (2nd ed., Vol. 4, pp. 1144-1146). New York, New York: Gale.

Swing, R. L., & Ross, L. E. (2016). Statement of aspirational practice for institutional research. *Association for Institutional Research, Tallahassee, Florida. Accessed October, 15*, 2017.

Volkwein, J. F., Liu, Y., & Woodell, J. (2012). The structure and functions of institutional research offices. In R. D. Howard, G. W. McLaughlin, & W. E. Knight (Eds.), *The handbook of institutional research* (pp. 22-39). San Francisco: Jossey-Bass.

(Statistics is for Everyone ...Continued from page 5)

gistic regression predicts a binary, categorical outcome. For example, you might want to predict whether or not a student will pass (1) or fail (0) a class based on a student's overall high school GPA, SAT score, and fall semester attendance record. You can do that with logistic regression. The interpretation of logistic regression results are very different than many other common analyses, which is why I think of it as an odd little bird. If you want to know more about probability, odds, and odds ratios, then you're probably going to love diving into logistic regression!

Exploratory Factor Analysis (EFA) / Confirmatory Factor Analysis (CFA)

The final multivariate test I want to mention is factor analysis. The primary function of these types of analyses (EFA, CFA) is to identify factors that underlie manifest variables. What does that mean? Imagine that there is an underwater geyser. You can't see the geyser itself, but you can see the bubbles on the surface of the water so you know the geyser exists. The bubbles are the manifested variables, which you can see and measure like items on a scale, and the geyser is the unobservable factor (also referred to as latent variable) that you cannot directly measure. Here's a concrete example: Let's say you want to create an instrument that measures the various components of the construct information literacy. As you can imagine, measuring constructs is a tricky business, but you decide that information literacy is comprised of three different underlying factors (the geysers) and you want to measure that via a thirty-item instrument (the bubbles). What you would then do, after administering the instrument to a group of students, is analyze the scores by looking at the correlations between items to identify the underlying factors. Generally speaking, if you had a hypothesis that information literacy was composed of a certain number of factors, say three, then you use confirmatory factor analysis to extract three factors and see how well that worked in explaining the construct. If you do not have a hypothesis of how many factors to extract, then you use exploratory factor analysis.

Conclusion

The purpose of this second article in this two-part series for instruction librarians was to provide an overview of common statistical tests and to explain the choices researchers make for why certain tests are employed over others. It is important to acknowledge that this article did not cover all the available types of statistical analyses, nor did I rely on external resources as I composed this piece. I decided that it was best to write as if I was talking to someone in person, librarian to librarian. However, as a librarian, I would be remiss if I did not end this article with some of the resources that I consistently rely on when working with students or advancing my own knowledge. The online content at UCLA's Institute for Digital Research and Education (IDRE) are phenomenal. Check that out! Also, I'm a big fan of Andrew Hayes (*Regression Analysis and Linear Models; Introduction to Mediation, Moderation, and Conditional Process Analysis)*, Geoff Cumming (*Understanding the New Statistics)*, and Jeremey Miles and Mark Shevlin (*Applying Regression and Correlation).* If you prefer something a little more advanced, I highly recommend Singer and Willett's *Applied Longitudinal Data Analysis* for librarians who are interested in measuring change across time.