An Introduction to Statistics for Librarians (Part Three): An Introduction to Statistical Tests

Caitlin Bakker, MLIS, AHIP-D[a],

[a]Discovery Technologies Librarian, Dr. John Archer Library, University of Regina, Regina, Saskatchewan, Canada, https://www.orcid.org/0000-0003-4154-8382, Caitlin.Bakker@uregina.ca

Cite as: Bakker C. An Introduction to Statistics for Librarians (Part Three): An Introduction to Statistical Tests. Hypothesis. 2024;36(1). doi:10.18060/27969

Data Bytes is a new column derived from the Statistics for Librarians series of columns in Research Mentor[1,2]. The purpose of this column is to provide librarians and information professionals with brief introductions to specific data- and statistics-related topics in a way that is accessible and relevant for their research and practice. The intention of the column is not to cover all aspects of statistics or data, or to turn information professionals into statisticians. Instead, the focus here is to help librarians and information professionals become savvier consumers of scholarly literature and to potentially enhance their own research practices.

Some Statistical Tests to Consider

In the previous two columns, the different types of data, central tendency, and frequency distributions were discussed[1,2]. In this column, we're going to build on the previous two columns to consider some of the different statistical tests that are available. The intention of this column isn't to state definitively which test should be used in any given situation, but rather to highlight how the topics from the previous two columns can be used to help you identify potential tests.

The below table, based on the original work of Samantha Carlson and Tanya Hoskin[3,4], is meant to introduce some tests to consider. This isn't a comprehensive inventory of all tests and isn't meant to definitively identify the "correct" test to use. The test you choose depends on the data that you have. Beyond the type of variable you have for your outcome (sometimes referred to as the dependent variable), you should also think about how big your sample size is, how closely your sample represents the overall population, and whether a parametric or non-parametric test is more appropriate.

Parametric tests assume that the underlying data are normally distributed, or that the data forms a bell curve. Parametric tests may also have other assumptions, such as requiring that the values of one observation aren't related to or influenced by the values of another observation, which is known as independence of observations. Before determining whether a parametric or non-parametric test is most appropriate, consider not only the distribution of the data but also confirm that all assumptions of the test are met.

Table 1. A selection of parametric and non-parametric tests for categorical and continuous outcome variables.

Number of Groups	Type of Outcome (Dependent Variable)	
	Categorical	Continuous
1	Binomial test	Single sample t-test One sample Wilcoxon signed rank test*
2	Chi-square test Fisher's exact test*	T-Test (paired or two sample) Wilcoxon signed rank test*
More than 2	Chi-square test Fisher-Freeman-Halton test*	ANOVA (one way or two way) Kruskal Wallis test*

* For non-parametric data or data with small sample sizes

Scenarios to Consider

To help operationalize the above table, let's look at some different scenarios in which it may be appropriate to conduct a statistical test and walk through the process of selecting tests for each.

Scenario #1: A librarian wants to determine if their class on health literacy is effective. At the end of the class, they give students a 10-question quiz and give one point per question.

Thinking back to the first column in this series,1 the librarian can determine that this is ratio (continuous) data because there is a natural zero, meaning that if a student were to get 0 out of 10, it would mean that there was an absence of any correct answers. There is only a single group, which limits the librarian's options for statistical tests. Statistical tests with a single group at a single point in time can be challenging, because it's necessary to have some reference point to which we are comparing that data.

The librarian investigates the underlying distribution of data to determine if the data is normally distributed. If so, the librarian could do a single sample t-test, which could be used to determine if the mean score on the quiz differed from either a known or hypothesized population mean. For example, if this was a quiz that had been developed elsewhere and the authors of the original quiz had published an article in which they reported the mean score on the quiz of a large, representative cohort, the librarian might use this as a hypothesized mean, or a value that is assumed to the mean score of a population. They could then compare the mean quiz scores for students who have participated in their class to that reference point.

Recalling the previous column on central tendency and distribution,2 we see that, because this test is comparing means, it would not be appropriate if the data were not normally distributed. If instead, the quiz scores skewed either positively or negatively, the librarian might instead decide to conduct a non-parametric test, such as a one-sample Wilcoxon signed rank test. In a one-sample **Wilcoxon signed rank test**, the librarian would be testing to see if there was a statistically significant difference between the sample's median and the hypothesized median. For more information on the t-test and its alternatives, consult Herzog, Francis & Clarke [5].

Scenario #2: That same librarian continues to teach the class on health literacy but starts introducing new modalities while continuing to give students the same 10-question quiz. After several years, they have quiz scores from students who participated in hybrid classes, in-person classes, and online classes. The librarian wants to know whether the modality of delivery has any impact on the quiz scores.

The librarian is still assessing continuous data as we were in scenario 1, but there are now more than two groups to consider. Because there are three different groups (hybrid, in-person and online), the librarian might consider doing an **ANOVA**, or an Analysis of Variance, which is a parametric test. An ANOVA can be either a one-way ANOVA or a two-way ANOVA depending on the number of independent variables, or factors, being included.

In a one-way ANOVA, there is one categorical independent variable (i.e., the teaching modality) and one continuous dependent variable (i.e., the quiz scores). The one-way ANOVA shows whether there are any statistically significant differences in the means of the different groups. In comparison, a two-way ANOVA can include two categorical independent variables. For example, the librarian may have also had two different lengths of classes, a 60-minute session and a 90-minute session. If the librarian wanted to assess whether the impact of both the teaching modality (e.g., hybrid, in-person, online) and the length of session (e.g., 1 hour, 1.5 hours) has any impact on quiz scores, they could conduct a two-way ANOVA. The ANOVA is comparing the mean scores and assumes that the underlying data are parametric. If the data were non-parametric, the librarian might consider conducting a **Kruskal-Wallis test**. To learn more about ANOVAs and Kruskal-Wallis, consider reviewing Sarty[6].

Scenario #3: The librarian is now working with a new program and is adapting their teaching and assessment strategies. The librarian can't give the same 10-point quiz in this program, but they are still offering course-integrated instruction. In addition to this class, the librarian has also developed an optional homework activity for students. The librarian wants to assess whether completing the optional homework activity has any impact on whether students pass or fail.

There are now two groups (students who complete the homework, students who do not complete the homework) and a categorical outcome (passing or failing the course). If a parametric test is appropriate, the librarian could consider using a **chi-square test** to evaluate whether the distribution of the outcome (passing or failing) is significantly different between students who did and did not submit the homework. In a chi-square test, data would be organized into a table known as a **contingency table**, which shows the frequencies of observations for each combination of variables (i.e., students who complete the homework and pass and those who complete the homework and fail). A chi-square test compares the actual number of observed counts in the table to what you would expect to see if there was no relationship between the different variables. To learn more about chi-squared analyses, consult Jarman[7].

However, rather than concluding that the homework activity has an impact on whether students pass or fail, we would instead be asking whether there is an association between submitting the optional homework and the likelihood of passing or failing the course. This is because, although we can establish whether there is some sort of relationship between the

variables, we cannot establish a cause-and-effect relationship. In statistics, **causation** means that a change in the independent variable will cause a change in the dependent variable, or outcome. Statisticians will often note that "correlation does not imply causation," which emphasizes that the existence of a relationship between variables does not mean that causation has been established. In observational studies, it can be difficult to establish causation because there may be confounding variables (e.g., students who don't submit the optional homework may have more commitments outside of school) and the researcher may not be able to manipulate variables directly (e.g., the homework is optional, so the librarian cannot force one group to do the homework while the other does not).

Interpreting Results

Sometimes, the challenge with statistics can be understanding the results of the analysis. What exactly does it mean to be "statistically significant"? P-values are ubiquitous, but that doesn't mean that they're always interpreted correctly. To understand p-values, we need to understand two concepts: a **null hypothesis** and an **alternative hypothesis**. A null hypothesis is the assumption that there is no difference. For example, in Scenario #2 above, the null hypothesis would be that there is no difference in the quiz scores from the different teaching modalities, and that any difference is just due to chance. An alternative hypothesis states that the null hypothesis is untrue (i.e., that there is a difference in the quiz scores from the different modalities).

But what does this have to do with p-values? The p in p-value stands for probability. It describes the probability that, in a world where the null hypothesis is true, you would observe data like this. So, we assume that we live in a world where the teaching modality doesn't have a relationship with the quiz scores, and then we look at the data from the quizzes and see how likely it is we would see those scores in that world. One important thing to remember is that a **p-value** doesn't tell you that the alternative hypothesis is true–it's not testing whether the teaching modality is related to quiz scores. Instead, a small p-value indicates that it's unlikely we would see this data if the null hypothesis were indeed true (i.e., if there was no difference).

What exactly the threshold for a "small" p-value should be is controversial. The most widely used value is less than .05. This essentially means that, in a world where the null hypothesis is true, there's less than a 5% chance of getting your results. Although this cutoff point is standard, it's also somewhat arbitrary. For more information about some of the history and controversy around p-values, consult Amrhein, Greenland McShane[8] and Kennedy-Shaffer[9].

The most important thing to keep in mind when we think about p-values is this: a low p-value isn't proof, it's evidence. A p-value of .001 does not definitively prove that something is absolutely true, it just means that it's highly unlikely that we would see this data in a world where it's not. Conversely, a p-value of over .05 doesn't mean that something isn't real or doesn't matter, it just means that it's more likely. Just as we have to be intentional about choosing which tests we use, we have to be cautious about our interpretations.

References

1. Bakker CJ. An Introduction to Statistics for Librarians (Part One): Types of Data. Hypothesis Res J Health Inf Prof. 2022;34(1). doi:10.18060/26428

2. Bakker C. An Introduction to Statistics for Librarians (Part Two): Frequency Distributions and Measures of Central Tendency. Hypothesis Res J Health Inf Prof. 2023;35(1). doi:10.18060/27162

3. Carlson S. Data analysis: Making sense of the numbers. Presented at: Department of Family Medicine Community Health, University of Minnesota; 2016; Minneapolis, MN.

4. Hoskin T. Parametric and Nonparametric: Demystifying the Terms. Mayo Clinic. https://www.mayo.edu/research/documents/parametric-and-nonparametric-demystifying-the-terms/doc-20408960

5. Herzog MH, Francis G, Clarke A. Variations on the t-Test. In: Understanding statistics and experimental design: How to not lie with statistics. Springer International Publishing; 2019:51-59. Accessed May 28, 2022. http://library.oapen.org/handle/20.500.12657/23029

6. Sarty GE. Introduction to applied statistics for psychology students. University of Saskatchewan Open Press; 2022. Accessed December 30, 2023. https://openpress.usask.ca/introtoappliedstatsforpsych/

7. Jarman KH. Bunco, bricks, and marked cards: Chi-squared tests and how to beat a cheater. In: Beyond Basic Statistics: Tips, Tricks, and Techniques That Every Data Scientist Should Know. Wiley; 2015:47-68.

8. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305-307. doi:10.1038/d41586-019-00857-9

9. Kennedy-Shaffer L. Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing. Am Stat. 2019;73(sup1):82-90. doi:10.1080/00031305.2018.1537891