Evidence Based Library and Information Practice

Commentary

A Case for the Use of Nonparametric Statistical Methods in Library Research

Megan Hodge
Assistant Head for Teaching & Learning
VCU Libraries
Virginia Commonwealth University
Richmond, Virginia, United States of America
Email: mlhodge@gmail.com

Received: 28 Feb. 2019 **Accepted:** 1 May 2019

DOI: 10.18438/eblip29563

If called upon to name statistical methods, the average librarian would likely reply with examples such as correlation and *t*-test. Those with more experience conducting research might name more comparatively exotic tests such as MANCOVA or factor analysis. These are all examples of what are known as *parametric* statistical tests: tests designed to identify a limited number of things about a data set where most characteristics of the data are already known or assumed. As parametric tests depend upon these assumptions, librarians who intend to use parametric tests must take care to collect data with these assumptions in mind.

While these assumptions vary somewhat from test to test, some are common to most parametric tests. One is the assumption that the data at least approximately resembles a normal distribution (also known as a "bell-shaped curve"), with most scores falling around a central score, fewer scores falling further from that central score, and few or no extreme scores. Another is the assumption that data is measured on a continuous scale, with the dependent (observed) variable measured on a real numerical scale, such as GRE scores or number of program attendees. A third common assumption is that there is a minimum number of participants in each group.

If even one of these assumptions is not true, then it is likely that parametric tests should not be used. Library research often violates these assumptions, with data that may be heavily skewed, with a tendency towards larger or

smaller values; a variable of interest that is often categorical, or non-numeric (e.g., emotions elicited by the library: anxiety, gratitude, wonder, frustration), or that may have an order but not a numerical value naturally associated with that order (e.g., degree of comfort using a database: not comfortable, neutral, comfortable); finally, sample sizes in library research are often small.

There are strategies that researchers can use to reduce the likelihood of these issues: rewrite questions to use a continuous rather than non-continuous scale; develop a plan to recruit a larger number of participants; remove outliers from the data. Sometimes these strategies are not feasible, however: the data may already have been collected and the sample size cannot be increased; the variable of interest cannot be measured on a continuous scale; the outliers are valid, if inconvenient, scores.

Fortunately, there is an alternative to parametric statistical tests: nonparametric statistical tests. Nonparametric tests tend not to rely upon the same assumptions required by parametric tests. Instead, nonparametric tests rely upon the median as a measure of a data set's central tendency, rather than the mean, which is the measure used by parametric methods. The mean is influenced by outliers in the data set; the median is not. Nonparametric alternatives exist for most common parametric methods, including ANOVAs, Pearson product-moment correlations, and *t*-tests.

Using a parametric statistical test when one or more of that test's core assumptions have been violated compromises the validity of the inferences that can be drawn from the test results and, by extension, the rigor of the research. In this case, the term 'inferences' refers to the conclusions that may be drawn from a test's results. It is usually not possible to survey or test every member in the population of interest (for example, academic librarians who have advanced into middle management positions within the last five years), and as such, inferential statistical tests may be used on a much smaller sample of that population to make inferences (generalizations) about that larger population. Parametric tests can have greater power to detect statistically significant differences and effects than their nonparametric equivalents; in other words, they can be more sensitive to effects and differences that are smaller in scale. However, using a parametric test when one or more of its assumptions has been violated may result in an inaccurate representation of the data — for example, when the mean is skewed far from the centre of the data by a few extreme values — which in turn means the inferences made about the larger population from which the sample was drawn may be flawed or inaccurate. Therefore, nonparametric tests, when called for, increase the rigor of a study's conclusions and the extent to which such conclusions are justified for use in evidence based practice.

A number of research scenarios common to library scholarship warrant the use of nonparametric statistical methods. Their use may be called for in order to increase a study's internal validity (the extent to which the study is able to investigate the topic of interest), to increase the study's statistical rigor, or both. Several of these research scenarios are described below.

Surveys are a popular research method for librarians, as is evident from the number of requests for participation that come through email discussion lists. Many of the sorts of questions that are asked in librarian-designed surveys would best be analyzed with nonparametric statistical methods, as our research interests often tend to elicit categorical or ordinal data. For example, librarians often employ Likert scale questions. These sorts of questions ask participants to respond on a five-point scale whether they strongly disagree, disagree, neither disagree nor agree, agree, or strongly agree with the question stem. Likert-

type questions, which use similar scales but which may have more points or ask about frequency rather than agreement, are also common. Ideally, in addition to identifying the construct(s) or variable(s) of interest, survey designers will have also identified all of the subconstructs making up the construct(s). For example, the construct of library anxiety might have attitudinal, cognitive, and behavioral subconstructs. A rigorous survey will have at least three questions that speak to each subconstruct of library anxiety; the survey designer will have determined the survey's construct validity (the extent to which the subconstructs do or do not represent all aspects of the construct itself); and evaluated whether the questions themselves adequately speak to each subconstruct. To analyze data collected from Likert scales, the response options are converted to artificial scores, with, for example, a 'strongly agree' converted to a one, an 'agree' converted to a two, and so on. Responses to Likert or Likert-type questions designed in this way allow for responses for each subconstruct to be combined, resulting in data on an interval scale that may be analyzed with parametric statistical methods.

If, however, there are only one or two questions that speak to each subconstruct or construct, the data created will be ordinal in scale: the difference in strength of feeling between one respondent's "strongly agree" and "agree" may not be the same as the difference between that respondent's "agree" and "neither agree or disagree." Further, the differences in strength of feeling are likely to differ between respondents. For example, a respondent who only slightly agrees with the question stem, and another who wholeheartedly agrees but does not consider their agreement 'strong,' may both choose a response of "agree." And, if there are few survey respondents, there may not be enough responses to meet the minimum number required for the anticipated parametric statistical test (for example, 15 per group for a *t*-test), or the data may not have a normal (bell-shaped) distribution: responses may be heavily skewed. All of these scenarios warrant the use of nonparametric tests.

Another common type of question on librarian-designed surveys are ranking questions. For example: "Please rank the following methods of receiving information from ACRL in the order in which you are most likely to use them." "Please rank the usefulness of each of the topics you learned about in today's webinar." "Please rank the following mediums of professional development in order of their desirability." Before analyzing the statistical significance of the data distribution, it is important to first assess whether respondents agree in their rankings: a given item may appear to be the most popular, but upon reviewing the data it may be revealed that the item was ranked last by a good number of respondents. This requires a nonparametric test that, essentially, tests inter-rater reliability on a large scale.

Quasi-experimental studies that evaluate the effectiveness of a program or instructional strategy are also common in the library literature. In all but the rarest of occasions, however, library studies do not have sufficient participants to meet the minimum threshold for the parametric tests librarians commonly use for these research designs, such as a *t*-test or ANCOVA (at least 15 per group, or 30 in a single group). Parametric statistical methods are influenced by outliers and therefore require a minimum number of participants to counteract the effect of any outliers. Additionally, most if not all parametric methods assume independence of observations: that each participant has received the treatment independent of all other participants. Independence of observations is important both because it ensures participants do not influence each other's scores, but also because it mitigates the risk of systematic bias in the scores. Systemic bias could be introduced in many ways: a fire alarm, resulting in all students in a class missing the same piece of content; a discussion that takes

place in one class section but not another; or seemingly minor differences in delivery between classes offered by different librarians. In short, a librarian who wishes to evaluate the effectiveness of a lesson taught to one class of 20 students has an *n* of 1, not 20. Unless the librarian is teaching for a course that has many sections, such as a first-year writing course, or is willing to collect data over multiple years (which introduces validity threats of its own), it is likely that the librarian will have a very small *n*. Data collected from these small or radically non-normal samples should be analyzed using nonparametric methods such as the Mann-Whitney *U* test or the sign test (alternatives for the independent samples *t*-test and paired samples *t*-test, respectively), which do not rely upon the assumption of a normal distribution.

Further, nonparametric tests may also increase librarians' understanding of the practical significance of their research. Statistical significance, or the likelihood of the findings *not* being due to chance, can be manipulated by increasing sample size; with a sufficiently large sample, most measured relationships/differences will be found to be statistically significant. A nonparametric test such as Kendall's *W* evaluates agreement among a large number of raters, is not affected by sample size, and will, for example, allow the researcher to determine the extent to which survey-takers agree on the order of ten ranked items.

One explanation for nonparametric tests' relative obscurity may be their lack of power (ability to detect small differences/associations between groups) when compared with their parametric counterparts. However, data that do not meet the assumptions undergirding parametric tests can in some cases be more powerfully analyzed with nonparametric tests. More statistical power results not just in a greater ability to detect differences or associations between groups, but in the statistical significance of the difference/association, or all-important *p*-value, to be much stronger.

These are just a few of the reasons that nonparametric statistical methods are more appropriate than parametric tests for many of the research designs favored by librarians. When used appropriately, nonparametric statistical methods can result in research findings of greater statistical validity and explanatory power. The subscription-based (but inexpensive) website Laerd Statistics is recommended as a resource for librarians wishing to identify nonparametric alternatives to specific parametric tests or learn more about nonparametric methods.