2-17-2012

Librarians and Statistics: Thoughts on a Tentative Relationship

Amy S. Van Epps
Purdue University, vanepa@purdue.edu

Librarians and Statistics: Thoughts on a Tentative Relationship

Amy S. Van Epps
     Purdue University

Abstract

Librarians are not trained as original researchers during library school. As a result, librarians as authors may succumb to common statistical misconceptions and use errors, thus it is important for librarians to know how to recognize them. A quick discussion of what a researcher should be aware of to avoid poor methods and inaccurate use of statistics is included. A review of statistics and research methods courses currently offered in library and information science programs helps determine if the lack of training observed in earlier studies has changed. A list of recommendations for authors and reviewers of Library and Information Science (LIS) literature is presented.

Corresponding author: Amy S. Van Epps vanepa@purdue.edu   i

Introduction

For years the library profession has perceived that librarians and statistics have an uneasy relationship. The seeming lack of confidence in using statistics is considered by some to stem from most librarians having a humanities background and not being numbers people (Dilevko, 2007; Hiller, 2002; Thornton, 2008). The perceived problem was surveyed by Powell, Baker, and Mika (2002) who found that 58.6% of library and information science practitioners felt they were not adequately prepared by their library and information science (LIS) programs to conduct research. The same study showed that 36.9% of the librarian practitioner-researcher population also feels unprepared to read and understand research publications. The level of discomfort discovered by Powell et al. (2002) supports Wallace's (1985) findings showing that 74% of the literature published in library and information science does not include any statistics. Recent studies of library science programs and their inclusion of research methods coursework showed that only about 60% of programs require master's students to take a research course before graduation (O'Connor & Park, 2001). A series of studies of the LIS literature, beginning as early as 1978, shows that the use of inferential statistics in library publications has been increasing (Dilevko, 2007; Wallace, 1985; Wyllys, 1978), which necessitates an understanding of statistics to read them, compounding the discomfort with statistics of many LIS practitioners. To improve the statistical preparation of LIS professionals, existing shortcomings need to be adequately described. This paper is a step in that direction.

Purposes of This Paper

General errors in using statistics that I have encountered on more than one occasion in LIS literature will be addressed, such as using descriptive statistics to make inferences about the data. A few insights on the sorts of issues that await the unwary user of statistics will be provided, such as assuming an observed change is significant without testing for statistical significance.

In addition, I take a quick look at the statistics preparation future library professionals may or may not receive as part of their LIS degree and suggest the lack of statistics in our publications reflects a lack of training.

Definitions

A variety of terms will appear throughout this discussion and in the referenced readings which need to be defined. When working with statistics, there are two categories of statistical methods to consider: descriptive and inferential. *Descriptive statistics* are used to summarize the collected data, for example totals, median, mean (average), and range (Byrne, 2007; Stephen & Hornby, 1995; Wyllys, 1978). Descriptive statistics encompass the type of calculations most of us are comfortable with based on early introduction during grade school and regular use of these figures in our lives, both professional and personal. For example, we all know what an average is, how to calculate it, and how it describes a set of numbers, be they grades, prices, or number of people in the library daily. When librarians include statistics in our professional literature, the tendency is to work within our comfort zone and use descriptive statistics. Wallace (1985) shows that within library science publications only 26% of the titles he studied used any statistics (20% descriptive, 6% inferential), with a clear preference for using only descriptive statistics. Wallace (1985) also demonstrates that the percentage of descriptive statistics in LIS literature is higher than other, similarly practice based professions – business, education, and social work – use in their published research (p. 406). For the four disciplines studied, only LIS had the descriptive only statistics as the second most frequently occurring type of article after no statistics; the others all had inferential statistics as their second most frequent category. Descriptive statistics can help show a change that has occurred or describe a shift in use patterns; they do not support decisions or thoughts on why those changes happened.

Inferential statistics, simply defined, are those used to draw conclusions and infer characteristics about a population (a large group), based on the data collected from a sample (a representative smaller part of the large group). Without getting into specifics about how to use these tests, examples of statistical methods used for making inferences include t-test, chi-square test, ANOVA, and Pearson correlations (Byrne, 2007; Wyllys, 1978). Inferential statistics allow a researcher to test a question or theory with data collected and to make a statement about a larger user group or population with a certain level of confidence. Inferential statistics help a researcher pose a question about why an observed pattern is happening or if there is a difference between expected behavior and observed behavior. Examples of research questions that could be addressed with each type of statistics are offered for clarification of their use.

Descriptive statistics:

Was a change in reference desk traffic seen? – This question can be determined

from simple comparison of sums or averages of number of transactions by week or month.

Inferential statistics:

Is the number of questions received at the reference desk a function of the number of people entering the library? – A correlation between number of questions with number of people entering the library (door count) can help determine a relationship and infer if having more people in the library leads to more questions.

Similarly, the roles of creating versus reading research literature have a few terms to define. *Active users* of statistics are also referred to as *producers of research statistics*. Both terms refer to individuals who are doing research and reporting statistics. *Passive users* are also called *consumers of research literature*, those who read the professional literature. Wyllys (1978), Liebscher (1998), and Powell et al. (2002) indicate the need for library professionals to be knowledgeable consumers of research to stay abreast of developments in the profession.

Discussion of Statistical Errors

In the past, when reading some articles that addressed collection analysis for electronic publications, it became apparent the article author(s) exhibited a lack of understanding about how to gather, analyze, and present quality data. These articles may represent a small, homogenous sample, but my fear is that they are indicative of a larger problem in the LIS professional literature. There is culpability on the part of any author(s) who publishes papers including poor statistics. The fact that the papers encountered all appear in peer-reviewed, professional LIS publications from a couple of well-known and respected publishers raises questions about the review process regarding articles containing statistics and data analysis. The problem seems larger than a particular author or two, a particular reviewer, or a single publication. If poor statistics are being pointed out by a reviewer and are not addressed by the author, it might be that the editors are allowing the 'sketchy' statistics to be published despite concerns. When limited to reading the final product, a reader is unable to know where problems may arise prior to publication.

Some errors I found while reading, both statistical and methodological, include the following items, each of which will receive some discussion later in the text.

- Use of any and all data returned from a survey with no mention of required sample

size for generalizations.
- No discussion of required survey response rates for drawing valid conclusions.
- No discussion of non-response bias in low response rates.
- No response or commentary regarding partial survey response (incomplete surveys).
- No comment on data discrepancies between that shown in tables of data with that presented in the text.
- Use of descriptive statistics to explain a situation.
- Lack of careful understanding or analysis of statistics gathered and supplied by a third party.
- Poor survey design leading to poor data.

Historically, Wyllys (1978) and Liebscher (1998) noted that the skill set of information professionals needs to include an understanding of both qualitative and quantitative data analysis, without which they would be unprepared to do their own research or be critical consumers of published research. A more recent article by Dilevko (2007) reiterates the same needs for library and information professionals and cites numerous previous studies which make the same assertion. Based on what has been encountered recently, it seems LIS programs in the United States have continued to produce inadequately prepared information professionals with regard to research and statistics skills. Dilevko's (2007) findings indicate that the trend in using inferential statistics in LIS articles is increasing, and so the need for librarians to be able to read and understand the articles using inferential statistics is increasing.

Working with Survey Data

The first three items in the errors listed above (use of any and all data returned from a survey with no mention of required sample size; no discussion of required survey response rates for drawing valid conclusions; no discussion of non-response bias in low response rates) are all about working with surveys and the data received from them. Sample size is a regularly used term that describes how many responses need to be received to gain a representative subset of the population. Similarly, response rate or return rate speaks to how many surveys have been received as a percentage of the total sent. A researcher tries to send out enough surveys so the response provides a sufficient sample size. For example, if a study requires a sample size of 200, and a return rate of 20% is expected, 1,000 surveys would be distributed (five times more than the required sample size) to ensure a sufficient sample. Librarians frequently work with data generated by surveying patrons or some

subset of a patron group. Unfortunately, frequently we do not consider sample size when conducting surveys and analyzing the data returned, choosing instead to give surveys to just those who enter the library or to students in a single department and use that data to make conclusions about all students. We use any responses received and consider ourselves lucky to get even a small set of answers and then analyze the data gathered, unaware of the validity, or lack thereof, of our conclusions.

In her short statistical primer, Byrne (2007) provides a concise description of sample size and sampling methods that can be used to generate non-biased samples for surveying. As Byrne states, in most cases the larger the sample size and the higher the survey return rate, the better. There is no easy formula for determining sample size for a given study, since sample size depends on the study design, the population (large group) to be sampled, and several other characteristics. In short, if the number of responses to a survey is too small, any conclusions made about the population as a whole will not be valid since the number of data points (sample size) is not sufficient to accurately represent the entire population (large group). When a researcher goes ahead with analysis, regardless of sample size and/or response rate, any non-response bias which could be introduced into the data is typically ignored. If the non-response bias is high enough due to low response rate, it can skew and possibly invalidate any research conclusions.

Consider a situation where the librarian-researcher receives survey responses from only those campus researchers who are unhappy with the library at the time of the survey. Now it is possible to recognize how the non-responses plays an important role in understanding the data. Without a complaint to offer, users who are happy with the services chose not to participate in the survey. When a researcher receives a low response rate or has a small sample size, it does not mean the study should not be published, just that the conclusions drawn are relevant only for the respondents and the data cannot be used to make statistically supported inferences about the entire group of people represented in the sample (population). It is important for a researcher to be clear about what is being concluded based on the data and what group the data describes. For example, studies using qualitative research methods, such as interviews, focus groups, and case studies, frequently have small sample sizes with a goal of finding information or patterns that could be applied in other settings or situations (transferable). The purpose of this type of study is not to use inferential statistics to create broad statements about an entire user group or population.

Clarity of Data Presentation

Failing to discuss, or at least mention, an anomaly in the data is poor practice. An author should discuss the apparent discrepancy in the text at the time, if that is appropriate, or mention that the discrepancy has been noted and will be discussed later in the text. The discrepancy should not be ignored; the reader will wonder what is causing the observed inconsistency.

When working with statistics, careful use of language is warranted so the reader is not misled. 'Significant' is a word that is used casually and can result in confusion when statistics are involved. In our regular speaking and writing, significant means important. When working with inferential statistics, the term significant has a specific meaning, which indicates that a tested research question is probably true with a certain degree of confidence. Using descriptive statistics to make claims about the level of importance for different values in a set of data can be dangerous. The difference between two numbers does not always need to be tested statistically for a discussion, unless the author wishes to determine if the difference is statistically significant. At the same time an author should be wary of calling the difference between two numbers slight without providing the full context of the data. With a small data set, a seemingly large difference between numbers is less likely to be important, and possibly significant in a statistical sense, than with a large data set, where the error term in the analysis will be smaller, making it easier to find significance. People who are not comfortable with statistics can easily fall prey to drawing conclusions from data which do not support their claims. It is possible to draw superficially plausible, but actually unwarranted, conclusions (e.g., a difference in numbers signifying a change that appears significant) if the data is not understood or not properly analyzed (Wyllys, 1978).

Survey Response Data

Lack of identification of partial or inaccurate survey responses (identified in initial list of errors) is part of a larger category of errors encompassing data verification in survey responses. Survey data is often messy, generating partial responses and invalid responses to certain questions. Messy data needs to be acknowledged in the context of the paper, including stating if incomplete data was removed from the sample analyzed. When an author goes ahead and uses the data available, she may end up reporting the information received for a given question that does not reflect the full sample (including voids in survey responses), and yet percentages are calculated using the full sample set size. The opposite problem can happen with invalid responses to questions, generating more data

than the size of sample set (for instance, a question where a respondent is asked to select one of three choices and yet some responses have more than one option selected). If this error is not accounted for, then the response rate on that question is greater than 100% and is meaningless. There are cases where a respondent is asked to choose all appropriate answers, but in that case a percentage of 100, as determined using the survey response sample set size, is not a valid method for representing the response rate for that question. In such a case, the percentage needs to be determined based on the number of responses to that particular question, not the number of surveys returned.

A related problem to using whatever data is returned from a survey of library users is that librarians are not schooled in creating valid survey tools any more than in analyzing the data the surveys return. Librarians, along with other professionals, use surveys because they can easily generate numerical or easily coded information, providing numbers for quantitative analysis. If a survey question is not well designed to gather the information a researcher needs, unwarranted conclusions may be drawn from the data (Hiller, 2002). The possibility of drawing unwarranted conclusions is at the heart of why it can be difficult to use a data set created by another researcher or a third party, such as a database provider. All research is context sensitive. The original researcher designs data collection for the specific need of the question(s) being investigated; if data is fully documented, and can be adequately repurposed to the new research question being investigated, then using available data can speed up answering a question.

Third-Party Data

Using data and statistics provided by a vendor is tempting to perform use-analysis studies of electronic materials, because it is a readily available set of numbers and may be the only source for such information. When using data from an external source it is particularly important to understand all the data points, what they mean, and what they measure (Hiller, 2002). Vendors all use similar terms, but often count very different things in those categories. The lack of clarity in vendor statistics led to the COUNTER project, which is an attempt to make vendor statistics more consistent and more easily comparable ("About COUNTER"). Despite that effort, it is still critical to check directly with a vendor to determine what is counted and how it is measured and reported. Once the researcher has clarity about the data, this information needs to be explained in the discussion of the data.

Much of the list above comes from errors encountered by the author while reading papers, but they are nicely summed up by Hiller (2002) as a list of things to think critically

about when working with numbers:
- Where do the numbers come from?
- How and why were they generated?
- What do they represent?
- Can you compare them?
- Do they make sense?
- Can we use the data to improve performance?

The observations in this paper bring me to a question for other librarians: if an author knows they lack the knowledge or skills to do a proper analysis of their data, why does she not seek advice from someone who has the skills? An answer to this question is beyond the scope of this article and may prove an interesting future study. The apparent need for additional education in the area of research and statistics lead me to investigate the options currently available as part of a typical LIS degree.

Recommendations for LIS Education and Statistical Training

Larger questions for the profession exist. Do library schools require courses in social science research methods, design of experiments, statistics and survey design, or some combination therein as part of the curriculum for all students intending to pursue careers as academic librarians? This question has been asked in some fashion for years, beginning with Wyllys in 1978. Many library and information programs do offer at least one course in research methods, so the question becomes one of requirements. Not all academic librarians are expected to publish, but it is possible for future academic librarians to end up in a tenure track position, where they need to research and publish, and thus it makes sense for that person to have research skills. Even for librarians who will never do original research, each graduate needs the ability to read and understand statistical research to stay current in the field (Dilevko, 2007; O'Connor & Park, 2001; Powell et al., 2002; Wyllys, 1978). Powell (1995) discusses the research skills needed for PhD students in LIS, which implies a lack of skills from those already holding a master's degree. PhD LIS students are by far the minority in the library profession.

Why do librarians not spend more time acquiring these skills once the need is discovered? This second question is perhaps a bit easier to respond to without further investigation. As practitioner librarians, with or without faculty rank, there is more than enough to do with existing daily tasks. The idea of taking a short course to acquire skills in statistical analysis and/or research

design may seem too large an additional load. The 'time excuse' does not negate the need, articulated by Hiller (2002) and Powell et al. (2002), for a practical approach to gaining the skills for doing research and using statistics. In the research completed by Powell et al. (2002), a strong correlation was found between research support provided on the job, primarily in the form of time, and the amount and quality of research generated by practicing librarians. Which is to say, if practicing librarians are given the time to focus on research as part of the job, they do stronger, higher quality research. Part of that time can be used to develop the skills required for high quality research. Once the skill set has been developed, it needs to be used, as ease with statistics depends on regular practice.

To get a sense of the current LIS education landscape, I did a quick review of the top 10 library and information schools in the United States as determined by *U.S. News and World Report* in 2009 ("Library and Information Studies") that turned up the following bits of information regarding research courses offered. The top schools, including those tied for 10th place, in order, are: University of Illinois – Urbana-Champaign; University of North Carolina – Chapel Hill; Syracuse University; University of Washington; University of Michigan – Ann Arbor; Rutgers; Indiana University – Bloomington; University of Texas – Austin; Drexel University; Simmons; University of Maryland – College Park; University of Pittsburgh; and University of Wisconsin – Madison. Library school webpages were analyzed for course listings and statements of required coursework for all master's level graduate students. When a program offers more than one concentration or area of study, the traditional LIS degree was used for consideration of course requirements, not the information or informatics degrees. Eleven of the 13 top schools offer at least one course on research methods of some sort, including quantitative and qualitative data analysis, survey design, or introduction to research methods. Only three of the schools require at least one research methods or statistics class for all their LIS masters graduates. The lack of required research courses for the master's degree indicates that our library and information schools are falling down on the job that Leibscher (1998) stated as the obligation of all LIS programs, "to ensure that their graduates are competent consumers of research" (p. 678). A more complete assessment of the availability and requirements of research methods courses and their content in LIS programs were conducted by Park and others (Dilevko, 2007), and I have listed those publications in Appendix A.

Educational Opportunities

For librarians who find themselves needing to be both consumers and producers of

research literature and yet inadequately prepared to do so, there are several things which can be done. Depending on where one works, some of these options will be more feasible than others: take short courses on statistical and research methods if offered, participate in the ACRL e-learning course on statistics for the non-mathematical mind (Association of College and Research Libraries, 2011), find a course in social sciences research methods on campus that can be taken or audited, or tackle a small research project and learn as the work is done. One other option would be to identify a larger research project and work with a mentor who has an accomplished research record on gathering data and doing statistics (Powell et al., 2002; Thornton, 2008). Alternately, if the librarian is a self-starter, find a collection of articles (Appendix B) or purchase a basic statistical methods textbook (one with an accompanying set of sample data), read the text, and work through the problems. With statistics, like so many things in life, practice makes one more comfortable with a topic (Byrne, 2007).

Conclusions

The observations articulated here make the author wonder if, as a profession, librarians are so impressed or frightened by the presence of numbers and statistical analysis in an article, that they do not read critically. Perhaps, as Hiller (2002) and Dilevko (2007) state, this lack of critical reading of numbers is because librarians are simply not comfortable with numbers and statistics. Where is the training in paper review, in being able at least to look at a set of numbers and see if they make sense based on what is said about them? We are a statistically impoverished nation, and most people never actually think about the validity or usefulness of any reported numbers. It is clear this has infected our professional writing within librarianship as well.

The takeaways for authors of LIS research articles:
- Do not draw inferences from descriptive statistics.
- Do not use data sets too small to accurately reflect the population being studied, or limit conclusions to the population studied, and do not make generalizations.
- Be careful with the language, for example use of the word 'significant.'
- Fully understand the data set and what each item measures.

For reviewers, a few things to remember:
- Do not be overly impressed by the presence of number and percentages in a paper.
- Notice generalizations made from data and look for sufficient sample size.
- Make note of and question generalizations drawn from very small sample sizes.

The goal of this article lies in raising the awareness of library and information science researchers and reviewers to be more discerning in future reading and writing that uses statistics.

Acknowledgements

Appendix A – Articles on Research Methods in LIS Programs

O'Connor, D., & Park, S. (2001). Guest editorial: Crisis in LIS research capacity. *Library & Information Science Research*, 23(2), 103-106.

Park, S., & O'Connor, D. (2001). *Research methods as a core competency*. Paper presented at the ALISE Annual Conference, January 11, 2001, Washington, DC.

Park, S. (2003). Research methods as a core competency. *Journal of Education for Library and Information Science*, 44(1), 17-25.

Park, S. (2004). The study of research methods in LIS education: Issues in Korean and U.S. universities. *Library & Information Science Research*, 26(4), 501-510.

Appendix B – Statistical Primer Materials

Byrne, G. (2007). A statistical primer: Understanding descriptive and inferential statistics. *Evidence Based Library and Information Practice*, 2(1), 32-47.

Davies, B., & Kendrick, T. (1989). The Chi Square Test: Statistics for Librarians Part 3. *Assistant Librarian*, 82(12), 179-182.

Davies, B., & Kendrick, T. (1989). Correlation and Regression: Statistics for Librarians Part 2. *Assistant Librarian*, 82(10), 150-153.

Davies, B., & Kendrick, T. (1989). Statistics for Librarians. *Assistant Librarian*, 82(8), 118-122.

Bibliography

About COUNTER.   Retrieved from http://www.projectcounter.org/about.html

Association of College & Research Libraries. (2011). Statistics for the non-mathematical mind. Retrieved from http://www.ala.org/acrl/onlinelearning/elearning/courses/statistics

Byrne, G. (2007). A statistical primer: Understanding descriptive and inferential statistics. *Evidence Based Library and Information Practice*, 2(1), 32-47.

Dilevko, J. (2007). Inferential statistics and librarianship. *Library & Information Science Research*, 29(2), 209-229.

Hiller, S. (2002). But what does it mean? Using statistical data for decision making in academic libraries. In *Statistics in Practice - Measuring and Managing*, IFLA Satellite Conference. Retrieved from http://info.lboro.ac.uk/departments/ls/lisu/downloads/statsinpractice-pdfs/hiller.pdf

Library and Information Studies. (2009). *U.S. News and World Report*, (Best Graduate Schools). Retrieved from http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-library-information-science-programs

Liebscher, P. (1998). Quantity with quality? Teaching quantitative and qualitative methods in an LIS Master's Program. *Library Trends*, 46(4), 668-680.

O'Connor, D., & Park, S. (2001). Guest editorial: Crisis in LIS research capacity. *Library & Information Science Research*, 23(2), 103-106.

Powell, R. R. (1995). Research competence for Ph.D. students in library and information science. *Journal of Education for Library and Information Science*, 36(4), 319-329.

Powell, R. R., Baker, L. M., & Mika, J. J. (2002). Library and information science practitioners and research. *Library & Information Science Research*, 24(1), 49-72.

Stephen, P., & Hornby, S. (1995). The joy of statistics. *Library Review*, 44(8), 56-62.

Thornton, S. (2008). The workplace research librarian. *Library and Information Research*, 32(100), 35-42.

Wallace, D. P. (1985). The use of statistical methods in library and information science. *Journal of the American Society for Information Science*, 36(6), 402-410.

Wyllys, R. E. (1978). Teaching descriptive and inferential statistics in library schools. *Journal of Education for Librarianship*, 19(1), 3-20.