york_11.wp1 7/27/11

RUNNING HEAD: Simpson's Paradox

Simpson's Paradox as a Challenge in Library Assessment: An Example Using LibQUAL+™ Long and LibQUAL+™ Lite Data

Bruce Thompson

Texas A&M University and Baylor College of Medicine

Martha Kyrillidou

Association of Research Libraries

Paper presented at the 9th Northumbria International Conference on Performance Measurement in Libraries and Information Services, York, England, August 22, 2011.

Bruce Thompson is distinguished professor of educational psychology and CEHD distinguished research fellow, and distinguished professor of library science, Texas A&M University, College Station, TX, and adjunct professor of allied health sciences, Baylor College of Medicine, Houston, TX. He may be contacted via e-mail at: bruce-thompson@tamu.edu.

Martha Kyrillidou is director of statistics and service quality programs at the Association of Research Libraries, Washington, DC. She may be contacted at: martha@arl.org.

Abstract

Many librarians may be unaware of Simpson's Paradox (Simpson, 1951). Simpson's Paradox involves the phenomenon that relationships or patterns in score means may disappear, or even reverse, when nested data are analyzed at different hierarchical levels (e.g., for users across all libraries completing a library service quality assessment, versus separate analyses computed at each individual library; or for users across all branches of a library, versus separate analyses computed for each branch with the same library). Librarians need to be aware of this phenomenon, lest they reach wildly incorrect interpretations of their assessment results. Our paper explains and illustrates Simpson's Paradox.

Many librarians may be unaware of Simpson's Paradox[1] (Simpson, 1951). Simpson's Paradox involves the phenomenon that relationships or patterns in score means may disappear, or even reverse, when nested data are analyzed at different hierarchical levels (e.g., for users across all libraries completing a library service quality assessment, versus separate analyses computed at each individual library; or for users across all branches of a library, versus separate analyses computed for each branch with the same library). Librarians need to be aware of this phenomenon, lest they reach wildly incorrect interpretations of their assessment results. The purpose of our paper is to explain and illustrate Simpson's Paradox.

Mean Ratings on LibQUAL+® Long versus LibQUAL+® Lite

Most scholars of library service quality are well aware of the LibQUAL+® Lite protocol. An overview of research related to LibQUAL+® Lite can be obtained by consulting Cook, Thompson and Kyrillidou (2010), Kyrillidou (2009), Kyrillidou, Cook and Thompson (2010), and Thompson, Kyrillidou and Cook (2009a, 2009b, 2010a, 2010b).

LibQUAL+® Lite is a protocol under which a given library user completes one "linking" item from each of the three LibQUAL+® Long

---

[1]Simpson's Paradox is not to be confused with Simpson's in the Strand. The first is a statistical phenomenon, while the second is a well known restaurant located next to the Savoy Hotel in London. Simpson's in the Strand features lovely roast beef carved at tableside from a silver trolley by a waiter dressed in starched white cottons. Simpson's in the Strand was founded in 1828:

http://www.simpsonsinthestrand.co.uk/history.php

and Stephen Towne will take you there, if you are very nice to him.

scales (i.e., Service Affect, Information Control, and Library as Place), plus five additional items randomly selected from the remaining 19 items (i.e., 22 items - 3 linking items = 19 remaining items). A given user completes (a) two items randomly selected from the remaining eight (i.e., 9 Service Affect items - 1 linking item = 8 remaining items), (b) two items randomly selected from the remaining seven (i.e., 8 Information Control items - 1 linking item = 7 remaining items), and (c) one item randomly selected from the remaining four (i.e., 5 Library as Place items - 1 linking item = 4 remaining items). Although a given user completes only eight of the 22 core items, nevertheless all 22 items are completed across the users of a given library.

Previous Research Findings

Researchers (Cook, Thompson & Kyrillidou, 2010; Kyrillidou, 2009; Kyrillidou, Cook & Thompson, 2010; Thompson, Kyrillidou & Cook, 2009a, 2009b, 2010a, 2010b) have conducted randomized clinical trials (RCTs) comparing, within given institutions, the mean total, scale, and item ratings of users randomly assigned either the LibQUAL+® Lite or the LibQUAL+® Long protocol. The following conclusions have in general been supported in this research:

1. Mean ratings tend to be lower (i.e., less favorable) on the LibQUAL+® Lite protocol, apparently because more users are willing to complete the shorter protocol, and thus the views of more users tend to be collected, including more people with somewhat less positive views of library service quality; and

2. The smallest differences in mean ratings tend to occur on the Service Affect as against the Information Control and the Library as Place scales.

Contradictory Canadian Consortium Results

In 2010, 47 Canadian libraries as a consortium participated in LibQUAL+®. Related comparisons of LibQUAL+® Lite and LibQUAL+® Long mean ratings for these data were presented by Kalb, Hong, Czarnocki and Champagne (2010). Kalb et al. (2010) interpreted their results as showing that LibQUAL+® Lite scores tend to be higher than LibQUAL+® Long scores, a result which contradicts previous findings (Cook, Thompson & Kyrillidou, 2010; Kyrillidou, 2009; Kyrillidou, Cook & Thompson, 2010; Thompson, Kyrillidou & Cook, 2009a, 2009b, 2010a, 2010b).

Table 1 presents selected findings from the Kalb et al. (2010) report. Note that for postgraduate students the mean LibQUAL+® Lite mean rating was slightly higher than the LibQUAL+® Long mean rating.

Table 1
2010 Long and Lite Perceived
Quality Scores for All of Canada

Group	n	M	(SD)	Cohen's d
Undergraduates				
Long	6194	6.930	(1.127)	0.024
Lite	10627	6.902	(1.166)	
Postgraduates				
Long	1532	**6.918**	(1.108)	-0.020
Lite	9159	**6.940**	(1.128)	
Faculty				
Long	871	7.073	(1.133)	0.249
Lite	3309	6.783	(1.199)	

What explains the contradictory findings reported by Kalb et al. (2010) versus the findings in the other reports? One explanation may be that the previous reports (Cook, Thompson & Kyrillidou, 2010; Kyrillidou, 2009; Kyrillidou, Cook & Thompson, 2010; Thompson, Kyrillidou & Cook, 2009a, 2009b, 2010a, 2010b) each involved randomized experiments in which users were randomly assigned to receive one of the two protocols.

On the other hand, in the Kalb et al. (2010) study, only two of the 47 institutions used both protocols. Among the 47 institutions, (a) two institutions randomly assigned both protocols, (b) 11 institutions used only the LibQUAL+® Long protocol, and (c) 34 institutions used only the LibQUAL+® Lite protocol. Thus, the mean ratings may differ across the two LibQUAL+® protocols because the institutions using one versus the other protocol themselves differed.

However, another (not necessarily contradictory) explanation involves the fact that Kalb et al. (2010) computed means for the data as a whole, aggregated across all institutions, while in the previous studies means were computed and compared only within institutions, and not across the institutions being studied. Thus, Simpson's Paradox may also explain the Kalb et al. (2010) findings.

Simpson's Paradox

Thompson (2008, pp. 9-11) explained Simpson's Paradox in the following way. Consider a hypothetical study in which a new medication, Thompson's Elixir, is developed to treat patients with serious coronary heart disease. The results of a randomized

clinical trial (RCT) five-year drug efficacy study of the new medication are presented below. Patients randomly assigned to "Treatment" receive Thompson's Elixir, while patients randomly assigned to "Control" receive a placebo sugar pill.

Outcome	Control	Treatment
Live	110	150
Die	121	123
% Survive	47.62%	54.95%

The initial interpretation of the results suggests that the new medication improves five-year survival, though the elixir is clearly not a "cure all" for these very ill patients. However, mindful of recent real research suggesting that a daily aspirin may not be as helpful for women as for men as regards heart attacks, perhaps some inquisitive women decide to look for gender differences in these effects. They might discover that for women only, these are the results:

Outcome	Control	Treatment
Live	58	31
Die	99	58
% Survive	36.94%	34.83%

Apparently, for women considered alone the elixir appears less effective than the placebo.

Initially, men might rejoice at this result, having deduced from the two sets of results (i.e., combined and women only) that Thompson's Elixir therefore must work for them. However, their joy is shortlived once they isolate their results.

Outcome	Control	Treatment
Live	52	119
Die	22	65
% Survive	70.27%	64.67%

In short, for both women and men separately, the new treatment is less effective than a placebo treatment, even though for both genders combined the elixir appears to have some benefits.

The paradox is that any relationship between variables may be changed, or even reversed, when data are analyzed at different levels of aggregation. For example, the means of all users completing one form of an assessment survey (e.g., LibQUAL+® Lite) might be **higher** than the means of all users who completed a different form of a library service quality assessment survey (e.g., LibQUAL+® Long) when we ignore the users' institutional affiliations, and yet, for the exact same data, when we compute means separately at each institution, the means of all users completing one form of an assessment survey (e.g., LibQUAL+® Lite) might be **lower** than the means of all users who completed a different form of a library service quality assessment survey (e.g., LibQUAL+® Long).

Discussion

One implication of Simpson's Paradox is that librarians ought to make comparisons at the level of analysis for which a given protocol was designed. If LibQUAL+® was designed only to compare libraries across institutions, and not users or user groups ignoring their institutional affiliations, all analyses ought be conducted within institutions.

Simpson's Paradox tells us that we can obtain completely contradictory results for aggregated versus disaggregated comparisons, even for the same data set! And the comparisons that are most ecologically valid are those comparisons made in the

usual context in which the analyses are conducted (i.e., at the institutional rather than at the national level). In summary, our message to persons interested in studying library service quality is simply: Beware of Simpson's Paradox!

References

Cook, C., Thompson, B. & Kyrillidou, M. (2010, May). Does using item sampling methods in library service quality assessment affect score norms?: A LibQUAL+® Lite study. Paper presented at the 2nd Qualitative and Quantitative Methods in Libraries (QQML 2010) International Conference, Chania (Crete), Greece. http://www.libqual.org/documents/LibQUAL/publications/lq_gr_3.pdf

Kalb, S., Hong, E., Czarnocki, S., & Champagne, S. (2010, October). Canada Lite: Impact of LibQUAL+™ Lite on the members of the LibQUAL+™ Canada consortium. Paper presented at the 3rd biennial Library Assessment Conference: Building Effective, Sustainable, Practical Assessment, Baltimore, MD.

Kyrillidou, M. (2009). Item sampling in service quality assessment surveys to improve response rates and reduce respondent burden: The LibQUAL+ Lite randomized control trial (RCT) (Doctoral dissertation, University of Illinois). Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/14570/Kyrillidou_Martha.pdf?sequence=3

Kyrillidou, M., Cook, C. & Thompson, B. (2010, May). Does using item sampling methods in library service quality assessment affect zone of tolerance boundaries?: A LibQUAL+® Lite study. Paper presented at the 2nd Qualitative and Quantitative Methods in Libraries (QQML 2010) International Conference, Chania (Crete), Greece. http://www.libqual.org/documents/LibQUAL/publications/lq_gr_2.pdf

Simpson, C. (1951). The interpretation of interaction in

contingency tables. Journal of the Royal Statistical Society, 13, 238-241.

Thompson, B. (2008). Foundations of behavioral statistics: An insight-based approach. New York: Guilford.

Thompson, B., Kyrillidou, M., & Cook, C. (2009a). Equating scores on "lite" and long library user survey forms: The LibQUAL+® Lite randomized control trials. Performance Measurement & Metrics, 10, 212-219.

Thompson, B., Kyrillidou, M., & Cook, C. (2009b). Item sampling in service quality assessment surveys to improve response rates and reduce respondent burden: The "LibQUAL+® Lite" example. Performance Measurement & Metrics, 10, 6-16.

Thompson, B., Kyrillidou, M., & Cook, C. (2010a, May). Does using item sampling methods in library service quality assessment compromise data integrity?: A LibQUAL+® Lite study. Paper presented at the 2nd Qualitative and Quantitative Methods in Libraries (QQML 2010) International Conference, Chania (Crete), Greece. http://www.libqual.org/documents/LibQUAL/publications/lq_gr_1.pdf

Thompson, B., Kyrillidou, M., & Cook, C. (2010b, October). Does using item sampling methods in library service quality assessment compromise data integrity or zone of tolerance interpretation?: A LibQUAL+® Lite study. Paper presented at the 3rd biennial Library Assessment Conference, Baltimore, MD.