Title:

"Overcoming Simpson's Paradox

Author(s):

James Boudreau, Bagwell Cente
Affiliated Faculty

Shane D. Sanders, Syracuse
University

Overcoming Simpson's Paradox

James W. Boudreau
Shane D. Sanders[†]

January 11, 2022

Abstract

Simpson's paradox (loosely speaking) occurs when multiple samples of data exhibit the same trend when examined separately, but that trend is reversed when the data is combined. This presents a potential problem for researchers, particularly in applications such as clinical trials, because samples are inevitably finite. In this brief we demonstrate the paradox and the problems it presents, then introduce an approach to identify when the paradox is a concern–or perhaps more importantly when it is not. Our approach is based on combinatorics, but is essentially a method of finding patterns in data, and relates to a widely-used statistical test in scientific research, the Mann-Whitney U (Wilcoxon rank-sum) test.

Kennesaw State University, Email: jboudre5@kennesaw.edu.

[†] Syracuse University, Email: sdsander@syr.edu

1 Introduction

To begin, consider a simplified example of a medical trial.[1] Suppose there are two possible treatments for a condition that are being tested, treatment X and treatment Y. In the first round of the trial, 300 patients are given treatment X and 100 are given treatment Y. The results are reported in Table 1, listing the number of patients that either did or did not respond to each treatment. While this is an admittedly stylized example, based on the results of the first round, treatment Y seems to have the higher response rate, with 70% versus 60% for treatment X.

Now suppose there's a second round of trial, the results of which are presented in Table 2. In that trial, 100 patients are given treatment X and 300 are given treatment Y. The results are not identical, but again treatment Y has the higher response rate, 30% versus 20%. But here's the trick.

Suppose instead this were one combined trial, with 400 patients in each group and the same results. These are presented in Table 3. Now treatment X has a response rate of 50% while treatment Y has a

the rankings of group B is 55. Thus, while group B has the lower total sum of rankings in the combined sample, group A has a lower sum of rankings if we compare either group on its own. A Simpson reversal. Why does this occur? And how can we know when to look for such occurrences?

Effectively what we've done here is taken the original full set of data from Example 2 and placed it into an ordinal sequence by rank:

$B_1; B_2; B_3; B_4; A_1; A_2; A_3; A_4; A_5; B_5; B_6; B_7; A_6; B_8; B_9; A_7; A_8; A_9; A_{10}; A_{11}; B_{10}; B_{11}; B_{12}; \ldots$
$A_{12}; B_{13}; B_{14}; A_{13}; A_{14}.$

Knowing what patterns to look for, we can then isolate one distinct group from another that we know can lead to a Simpson reversal. In this case:

$B_1; B_2; B_3; B_4; A_1; A_2; A_3; A_4; A_5; B_5; B_6; B_7; A_6; B_8; B_9; A_7; A_8; A_9; A_{10}; A_{11}; B_{10}; B_{11}; B_{12}; \ldots$
$A_{12}; B_{13}; B_{14}; A_{13}; A_{14}.$

The group highlighted in red represents the second trial when the two were separated for this simple example. Had we simply used a brute-force method to assemble separate combinations of the two groups of equal size, checking for the existence of a Simpson reversal each time, the sheer number of possibilities would make the time necessary for a complete search impractical for even small sample sizes, and virtually infeasible for the larger data sets that are more typically used in scientific studies.[3]

Instead, our research focuses on narrowing and guiding the search for possible Simpson reversals by

References

Ameringer, S., R. C. Serlin, and S. Ward (2009). Simpson's paradox and experimental research. Nursing Research 58(2), 123–127.

Boudreau, J., J. Ehrlich, and S. Sanders (2022). Prevalence of simpson's paradox in nonparametric statistical analysis of medical and other scientific data: Theoretical and computational analysis. Working Paper, Bagwell Center for the Study of Markets and Economic Opportunity, Kennesaw State University.

Boudreau, J., J. Ehrlich, S. Sanders, and A. Winn (2014). Social choice violations in rank sum scoring: A formalization of conditions and corrective probability computations. Mathematical Social Sciences 71(1), 20–29.

Heydtmann, M. (2002). The nature of truth: Simpson's paradox and the limits of statistical data. Q J Med 95(4), 247–249.

Mann, H. B. and D. R. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematics and Statistics 18(1), 50–60.

Rojanaworanit, C. (2020). Misleading epidemiological and statistical evidence in the presence of simpson's paradox: An illustrative study using simulated scenarios of observational study designs. Journal of Medicine and Life 13(1), 37–44.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society, Series B 13(2), 238–241.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. Biometrika 2(2), 121–134.