Overcoming Simpson's Paradox

James W. Boudreau*
Shane D. Sanders†

August 5, 2021

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 response rate of 40%. Which is the truth? This is an example of Simpson's Paradox (Simpson, 1951).

Separately the two samples suggest one trend, but when combined the trend is reversed. This phenomenon is well-known in the statistics literature, and was pointed out by Yule (1903) before Simpson formalized it. It presents a challenge to researchers, particularly those in fields that rely on clinical trials, due to the simple reality of scarce resources and limited data. Suppose, for example, that instead of two separate rounds of the trial in our example, there was only the one combined study as presented in Table 3. It could be the case that the results would have been reversed–a *Simpson reversal*–had the study been separated into two groups, perhaps with the first consisting of only patients under the age of 65 and the second consisting of only patients over the age of 65. Or perhaps it could have been men in one group and women in the other. The reality is that there are an unknown number of latent variables that could have been controlled for, but may not have been due to issues such as funding. Is there a way to examine the data to determine whether or not the larger group should be split into sub-samples?

Alternatively, suppose these groups were generic and a larger, combined sample would have led to a

[1] This example is based on Tables 4-6 of Heydtmann (2002). Similar ratios have been used to introduce the paradox elsewhere.

Table 1: Example 1, First Round Results

	Response	No Response	Response Rate
X	180	120	$180/300$=60%
Y	70	30	$70/100$=70%

Table 2: Example 1, Second Round Results

	Response	No Response	Response Rate
X	20	80	$20/100$=20%
Y	90	210	$90/300$=30%

Table 3: Example 1, Combined Results

	Response	No Response	Response Rate
X	200	200	$200/400$=50%
Y	160	240	$160/400$=40%

more accurate result, but there was only funding for the first round of the trial? If only round one of the trial occurred can we determine whether or not a reversal is possible with a similar sample?

Although the pitfalls of the Simpson Paradox have been explicitly acknowledged in areas besides statistics including medical research journals (e.g. Rojanaworarit, 2020; Ameringer et al., 2009; Heydtmann, 2002), options for addressing it have been limited until now. While conducting more finely controlled experiments that identify all possible variables is always recommended, we are developing a robustness test that can allow researchers to determine whether Simpson reversals are even possible in their data. This in turn will help them to determine whether additional trials are needed, or whether sub-sampling is required. Our approach is based on combinatorics, though to put it more simply we just look for patterns in the data based on how two groups differ with one another.

2 Another Example: Detecting the Paradox

Our approach is based on the *ranking* of one group as compared to another. The following simple example demonstrates the logic of that approach, though a more full explanation is provided in Boudreau, Ehrlich, and Sanders (2020).

Consider a medical trial comparing two different treatments, A and B. The results reported indicate the

Table 4: Example 2, Combined Results

#Aches:	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}
A	21	24	12	75	72	54	64	51	48	15	18	42	45	33
B	55	58	25	6	0	36	61	27	30	70	67	9	3	39

Table 5: Example 2, Combined Results by Rank

#Aches:	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}
A	12	15	18	21	24	33	42	45	48	51	54	64	72	75
$R(A_i)$	5	6	7	8	9	13	16	17	18	19	20	24	27	28
#Aches:	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}	B_{12}	B_{13}	B_{14}
B	0	3	6	9	25	27	30	36	39	55	58	61	67	70
$R(B_i)$	1	2	3	4	10	11	12	14	15	21	22	23	25	26

number of occurrences of a specific symptom in the patients receiving each different treatment over a given period of time. For the sake of example, suppose it's the number of headaches each patient experiences over the course of six months. Table 4 reports those results, with P_i referring to patient i in each of the two groups.

Now there a number of ways we can compare these two groups of data. And keep in mind, again, that we are keeping the sample size of this example extremely small for illustrative purposes. One popular method used to compare two groups of data in scientific research, however, is the Mann-Whitney U test, also known as the Wilcoxon rank-sum test (hereon the $MWW - U$), which is based on comparing the *ranks* of the two groups of data. To demonstrate this idea, in Table 5 we present the same data from Example 2, but simply re-ordered, and also include the overall rank of each patient's result in the total data set. The patient from group B who had zero occurrences of symptoms, for example has a rank of 1 since they had the lowest number in the data set, while the patient from group A who had 75 occurrences is given a rank of 28 since they had the highest number and so forth. In other words, the data points are simply ordered first to last.

In Table 5 we have also arbitrarily re-labeled the patients in each group just for reference. For example, patient A_1's overall rank in the data, $Rank(A_i)$, is 5, since their number of reported headaches was 12. Patients B_1, B_2, B_3, and B_4, meanwhile had ranks of 1 through 4 respectively since their reported numbers were ranked accordingly.

Table 6: Example 2, Trial 1 Results by Rank

#Aches:	A_1	A_2	A_3	A_4	A_5	A_6	A_{14}
A	12	15	18	21	24	33	75
$R(A_i)$	3	4	5	6	7	11	14
#Aches:	B_1	B_2	B_5	B_6	B_7	B_8	B_9
B	0	3	25	27	30	36	39
$R(B_i)$	1	2	8	9	10	12	13

Table 7: Example 2, Trial 2 Results by Rank

#Aches:	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
A	42	45	48	51	54	64	72
$R(A_i)$	3	4	5	6	7	11	14
#Aches:	B_3	B_4	B_{10}	B_{11}	B_{12}	B_{13}	B_{14}
B	6	9	55	58	61	67	70
$R(B_i)$	1	2	8	9	10	12	13

The $MWW-U$ test is based on the following simple premise: sum the ranks of each group and compare them. In our Example 2, for example, the sum of ranks for the group A treatment is 217 while the sum of ranks for group B is 189. Loosely speaking, this shows that group B has more patients that rank lower in terms of their number of symptoms, suggesting that group is different from the other in that regard. Again, loosely speaking, if the sample were larger and the results were statistically significant[2], if one member were chosen from each sample, group B would be more likely to show a better result. But of course, we've constructed this example for a reason.

It turns out we can subdivide this full sample into two samples of equal sizes and demonstrate a Simpson reversal, and we know that because of the specific patterns the ranking exhibit. Specifically, consider same exact data, but separated into the two trials in Tables 6 and 7 above. Note that since the size of each trial is half the original, rather than ranking each from 1 through 28, we rank them each separately 1 through 14, ranking them only in terms of their rank within the trial.

One trick here is that both trials have been separated such that one is an *ordinal replicate* of the other. That is, comparing the rankings of groups A and B in the first trial, they are the same as those in the second trial. Furthermore, the sum of the ranking of group A in either case is only 50 while the sum of

[2]We omit a full description of the $MWW-U$test here for brevity, but see Mann and Whitney (1947).

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 first establishing boundaries on the difference in the sums of ranks between two groups that guarantee when those reversals are even possible. In particular, if the sum of ranks for one group is sufficiently larger than another, we know that any reversal is impossible. That boundary changes as the size and composition of the groups change, but knowing them can eliminate the need to search at all.

Furthermore, even when Simpson reversals are possible, we can narrow down the search of which two "trial groups" to look for by knowing which patterns of sequences of ranks allow for reversals and which do not. In Example 2 above, for example, one may notice the way a few B components lead, followed by a string of A components, then some alternating but with A components trailing at the end. This is an

[3]For a detailed explanation of how many possible combinations there are for any two groups of equal size, see our working paper, Boudreau et al. (2020). As explained in that paper, however, even for two groups with seven elements each, the number of possible combined sequences is staggering. Specifically, there are approximately 137.68 billion possible combinations for the 2×7 case.

extremely vague description, but by knowing how the sums of components in order can or can not lead to reversals, we can narrow down the search for possible Simpson reversals to a practical time. This process is helped by the fact that it turns out such combinations are statistically rare, so if we know what types of combinations are even possible and needed to be checked for, we do not have to engage the brute force approach of examining the entire data set.

In the end, then, our hope is that researchers will be able to check their data for the likelihood of Simpson reversals–and thus the potential for spurious conclusions–in a practical way. Rather than having to worry that some sub-trial groups need to be split according to latent characteristics that may not have been thought of, the use of combinatoric logic–the various ways different numbers can be combined–can be used to determine whether or not it is necessary.

3 Discussion on the Origins of our Methods

A final note on our methods that may clarify how they work (or at least be of interest to readers) concerns their origins. Years ago we began doing work on scoring in team cross-country running meets. In these competitions two teams have groups of runners that all run one race, and at the end there is one final order of finishers. The standard method of scoring used is rank-sum scoring, whereby the first-place finisher scores one point for their team, the second-place finisher scores two points for their team, and so on. A process, it turns out, qutie related to the $MWW - U$ test, though we didn't know it at the time.

Our original work at that time (e.g. Boudreau et al., 2014) was interested in features of the scoring method itself. Questions such as the possibility and likelihood of things like cycles, whereby one team could beat another, that team could beat a third team with the same runners, but the third team could beat the first with its same runners. In working on the methods necessary to study these more esoteric questions both mathematically and computationally, however, we realized that they had far more broad applications than originally intended. This highlights the progression of research itself, as new methods can spring from seemingly unexpected sources.

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 (2020). Prevalence of simpson's paradox in nonparametric statistical analysis of medical and other scientific data: Theoretical and computational analysis. Technical report, *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.