Simpson's Paradox in Nonparametric Statistical Analysis: Theory, Computation, & Susceptibility in Public Health Data

James Boudreau[1], Justin Ehrlich[2], and Shane Sanders[3]

[1]Department of Economics, Kennesaw State University
[2,3]Falk College of Sport & Human Dynamics, Syracuse University

Abstract

This study establishes sufficient conditions for observing instances of Simpson's (data aggregation) Paradox under rank sum scoring (RSS), as used, e.g., in the Wilcoxon-Mann-Whitney (WMW) rank sum test. The WMW test is a primary nonparametric statistical test in FDA drug product evaluation and other prominent medical settings. Using computational nonparametric statistical methods, we also establish the relative frequency with which paradox-generating Simpson Reversals occur under RSS when an initial data sequence is pooled with its ordinal replicate. For each 2-sample, n-element per sample or 2 x n case of RSS considered, strict Reversals

al. (2016) nd evidence of aggregation paradox instances in randomized clinical trial data. Evidence of aggregation paradox has also been found in the settings of large-scale registry data (Gron et al. 2016), meta-analyses of an academic literature (Kuss 2016), and clinical risk reclassi cation (Cook et al. 2017).

In one respect, the Paradox can be viewed as a robustness check on a given statistical result. When the Paradox occurs, it follows that a given result is at least partly a function of data scale or sample size. As noted, the Paradox has been shown to occur for the Wilcoxon-Mann-Whitney (WMW) Rank Sum Test. However, there exists no computational or empirical evidence as to the frequency with which instances of the Paradox occur for the WMW Rank Sum Test and little such evidence for non-parametric statistical tests overall. Are Simpson Reversals pervasive or only a marginal concern for the WMW Test? Even previuos research as to the incidence of the Paradox for parametric statistical tests is scarce and provides somewhat contrasting conclusions. There are two studies that directly estimate the incidence of Simpson's Paradox for parametric tests: one pertaining to contingency tables and the other pertaining to path models. Speci cally, Pavlides and Perlman (2009) nd that a Simpson Reversal occurs for one-sixtieth (1.67%) of all 2 x 2 x 2 contingency tables. Kock (2015) estimates the likelihood of a Simpson Reversal in path models as approximately 12.8%.

test and data under consideration.

2 Material and Methods

2.1 Rank Sum Scoring and Simpson's Aggregation Paradox: De nitions and a Theorem

Let us formally de ne 2-group rank sum scoring. Consider two groups, A and B. Each group is de ned as a rank-ordered sequence of n individual elements, where n is some integer greater than 1 ($n \in Z^+$). For example, A is de ned as $A = (a_1; a_2; a_3; :::; a_n)$, where the element a_i represents the i^{th} ranked element in A. We de ne an event as an objective process of comparison that generates a complete rank-order sequence of individuals across more than one group (i.e., both within and between groups). An event might be de ned as a competition or as a statistical test. Consider an event in which elements of A and B are compared. If A and B are each composed of n elements, for example, then the event generates a rank-ordered outcome sequence of 2n elements. One possible outcome sequence for the case in which n = 3 is $F_{AB} = (a_1; b_1; b_2; a_2; b_3; a_3)$. If a_i precedes b_j in the outcome sequence, we say $a_i \quad b_j$ (a_i ranks higher than b_j). For simplicity, we assume that rank-order equality between two elements is not possible, an outcome that would obtain given continuous measurement of underlying parameter values. For any $a_i \in A$ and $b_j \in B$, that is, we have that $a_i \quad b_j \quad b_j \quad a_i$ is a tautology.

Formally, we represent the rank of an element $a_i \in A$ in the outcome sequence F_{AB} as $r(a_i \mid F_{AB})$. Let $x_i^+(F_{AB}) = \{x \in F_{AB} : x \quad a_i\}$ be the set of elements in F_{AB} that rank better than a_i. Then, $r(a_i \mid F_{AB}) = \mid x_i^+(F_{AB} \mid +1$. From elemental rankings, we generate a rank sum score for each group as follows. The respective scores for A and B for outcome sequence F_{AB} are $S(A \mid F_{AB}) = \sum_{a_j \in A} r(a_j \mid F_{AB})$ and $S(B \mid F_{AB}) = \sum_{b_j \in B} r(b_j \mid F_{AB})$, where it must be that $S(A \mid F_{AB}) + S(B \mid F_{AB}) = \frac{2n(2n+1)}{2}$. That is, the sum of ranks for a 2n element sequence simply equals the sum of integers from 1 to 2n. Where ties are allowed, [illegible]

$B \succ_{FF'} A$ (i.e., that A ranks strictly higher than B in F and F', but B ranks strictly higher than A in FF') or that $B \succ_{F} A$ and $B \succ_{F'} A$, but $A \succ_{FF'} B$ (i.e., that B ranks strictly higher than A in F and F', but A ranks strictly higher than B in FF').

Table 1: Sufficient Condition for Presence of Reversal and Observation of at Least One ...

...ne Reversal	Sufficient Condition for Presence of Reversal	Observation of at least ...		
			F	T
		F	162	12
		T	0	78

...of at Least One... Table 2: Sufficient Condition for Absence of Reversal and Observation

...	Sufficient Condition for Absence of Reversal	Observation of at least one Reversa...		
			F	T
		F	84	90
		T	78	0

...of at least one Reversal... ...Observation...

Table 4: Sufficient Condition for Absence of Reversal and Observation of at Least On...

Of the 252 initial sequences, F_{AB}, Table 1 tells us that the su cient condition for presence of at least one Reversal across all poolings of F_{AB} and F^{0}_{AB} holds for 78 of those sequences. Empirically, we observe at least one Reversal for each of those sequences. The second table shows that for a distinct 78 of the 252 initial 2 x 5 sequences, F_{AB}, the su cient condition for absence of Reversals across all poolings of F_{AB} and F^{0}_{AB} holds. Empirically, we do not observe a Reversal in any of those sequences. For the 2 x 5 case, then, the su cient conditions from Theorems 1 and 2 assure us whether or not Reversal is possible for 156 of the 252 initial sequences (61.9%).

Tables 3 and 4 deal with su cient conditions for the 2 x 6 case. Of the 924 initial sequences, F_{AB}, for the 2 x 6 case, Table 3 shows that the su cient condition for presence of at least one Reversal across all possible poolings of F_{AB} and F^{0}_{AB} holds for 220 of those sequences. Empirically, we observe at least one Reversal for each of those sequences. The fourth table shows that for a distinct 364 of the 924 initial 2 x 6 sequences, F_{AB}, the su cient condition for absence of Reversals across all poolings of F_{AB} and F^{0}_{AB} holds. Empirically, we do not observe a Reversals for any of those sequences. For the 2 x 6 case, then, the su cient conditions from Theorems 1 and 2 assure us whether or not Reversal is possible for 584 of the 924 initial sequences (63.2%). In each observed case, the su cient conditions determine unambiguously whether an initial sequence is susceptible to Reversal in more than three- fths of cases. Therefore, we can usually assess the general robustness of a rank sum result in terms of

susceptibility to Simpson Reversals. As such an assessment can determine whether a given result is scale-variant, we conclude that Theorems 1 and 2 can usually combine to o er a \quick and dirty" robustness check on a rank sum result.

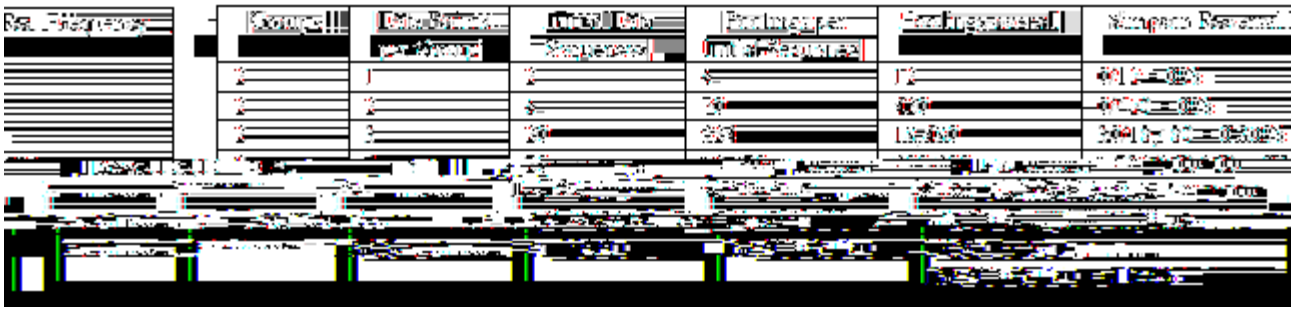
2.2 The Sample Space: A Combinatorial Description

For the 2 x n case, there are $\frac{(2n)!}{(n!)^2}$ initial sequences, F. We are arranging 2n elements | n elements from each of 2 groups | where we do not distinguish between respective objects of a given group. For each initial sequence, we then ask in how many ways F can be pooled with its ordinal replicate, F^{0}. This is equivalent to a \stars and bars" combinatorial problem, in which we are placing 2n \stars" or elements from F^{0} into 2n \bars" or potential pooling positions amongst the elements of F. From this characterization, there are $\frac{(4n)!}{([2n]!)^2}$ poolings for each initial sequence and $\frac{(2n)!}{(n!)^2}$ initial sequences. The number of poolings for a given 2 x n case equals the product of the number of initial sequences and the number of poolings per initial sequence, or $\frac{(2n)!}{(n!)^2} \quad \frac{(4n)!}{([2n]!)^2}$, for each case, 2 x n. For example, in the 2 x 7 case, there are $\frac{(2\ 7)!}{(7!)^2} = 3;432$ initial sequences, F. Moreover, there are $\frac{(4\ 7)!}{([2\ 7]!)^2} = 40;116;600$ poolings per initial sequence. As such, there are $3;432 \quad 40;116;600$ or approximately 137.68 billion possible poolings for the 2 x 7 case. We provide the sample space for each 2 x n case in Table 5 of the subsequent section.

2.3 Computational Methods Materials

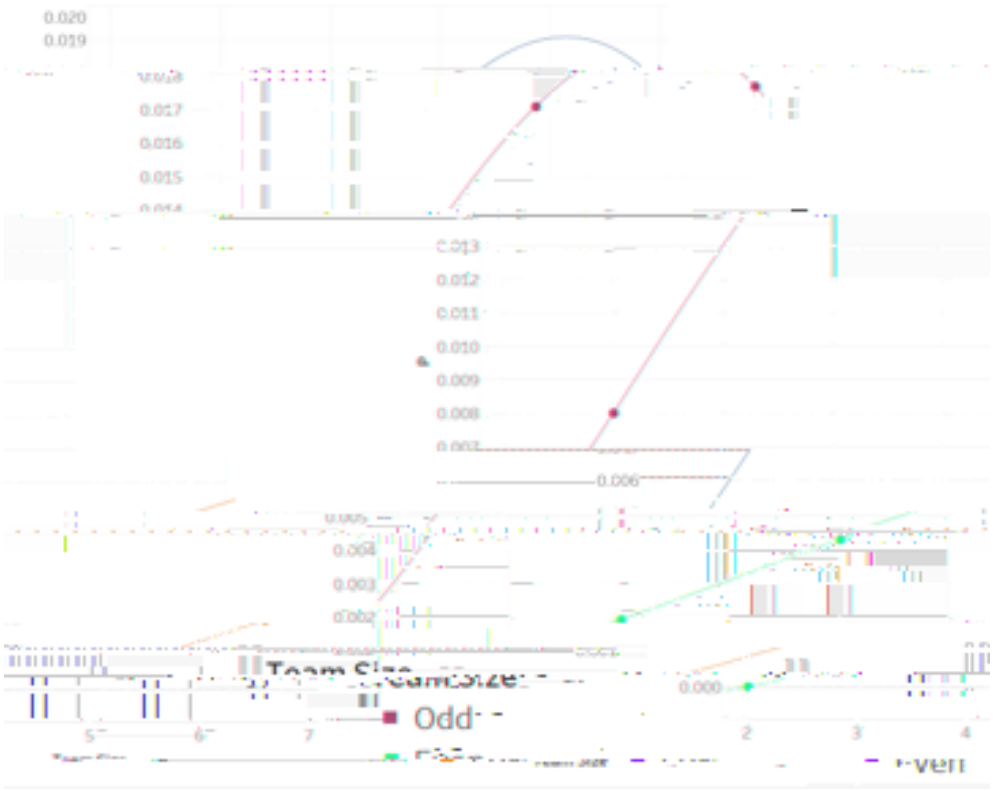
We wrote a computational algorithm in Java by which to search the sample space of each case where $0 < n(2\ Z^{+}) < 7$: The algorithm is shown in Appendix 1. It programmatically generates all possible initial sequences, F_{AB} (F^{0}_{AB}), for a case, then creates all possible pooled sequences, FF^{0}_{AB}; for each initial sequence. For each initial sequence, rank sum scores for A and B are computed. This scoring task is then repeated for each pooling FF^{0}_{AB} of F_{AB} and F^{0}_{AB} and iteratively for each pooling of each initial sequence. Then, instances of Simpson Reversal are checked using the

Table 5: Relative Frequency of Simpson Reversal by Case



We observe that Simpson Reversals are not possible for su ciently small n (i.e., n < 3). In the context of Theorem 1, the largest possible is not su ciently large to motivate a strict Simpson Reversal in these cases. For the 2 x 1 and 2 x 2 cases, a group that is strictly outranked in F_{AB} cannot have a positive , and therefore a strict Simpson Reversal is not possible for these cases. We can also consider computed cases where n > 2. From even to odd case, the results suggest a wavelike movement in the likelihood of a Simpson Reversal. In general, there is a lower likelihood of strict Simpson Reversal in even cases than in neighboring odd cases due to the possibility of ties for n-even cases of pairwise rank sum scoring (but not for n-odd cases). With some probability mass allowing for a pairwise tie in the n-even case, strict Simpson Reversals are less likely. This result also holds for other social choice violations (e.g., violations of Transitivity and of IIA; see Boudreau et al. 2014). To evaluate the marginal e ect of increases in n, as distinct from the e ect of changes from even to odd case, one should compare the iterative trend between n and n + 2 rather than that between n and n + 1. We do this for the even and odd cases, respectively, in Figure 1.

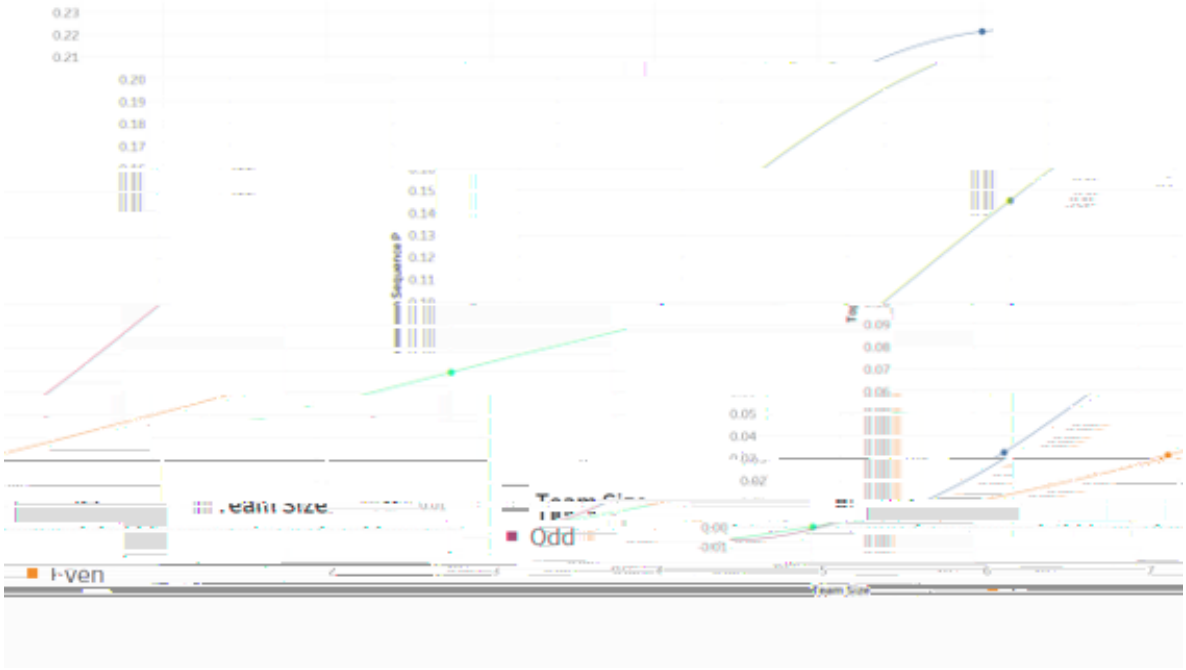
Figure 1: Frequency of Simpson Reversal by Case



Over the set of cases computed, the relative frequency of Reversal rises for both the even and odd sets of cases. For the 2 x 8 case, we run a simulation to estimate whether this trend might continue. Speci cally, we randomly

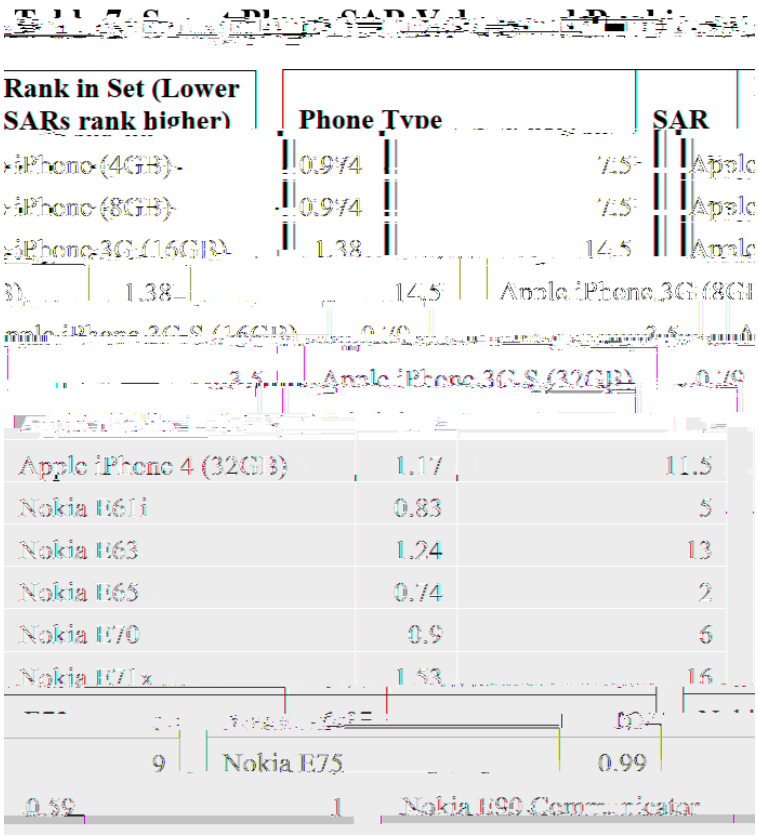
select and generate one-quarter of all possible initial sequences, F_{AB}, (without replacement) for this case and then replicate each selected initial sequence. For each selected initial sequence and its replicate, we then randomly select approximately 0.1% of all possible poolings, or a little more than 600,000 poolings per sampled initial sequence. For each pooling, we check for Reversals

Figure 3: Highest Initial Sequence Level Reversal Likelihood by Case



2	1	0/6	NA	NA	NA
2	2	0/0	NA	NA	NA
2	3	30/924 = 3.25%	abbaab	10 - 11	1
2	4	402/12870 = 3.12%	abbbaaab	18 - 17	1
2	5	26,879/184,756 = 14.54%	aabbbbaaab	27 - 28	

Table 6 shows that reversals are more likely given sequences that feature both close rank sum scores and uninterrupted clusters of one group and then of the other within the rank sequence. Note that the maximum Reversal likelihood generating sequence for each case is not unique. In each case, one could transpose the elements 'a' and the elements 'b' to obtain the same Reversal likelihood. We find that the maximum Reversal likelihood generating sequence also generates the closest margin of victory in each case (i.e., 1 rank sum unit for n-odd cases and 2 rank sum units for n-even cases). While the overall likelihood of Reversal



Rank in Set (Lower SARs rank higher)	Phone Type	SAR	
iPhone (4GB)	0.974	7.5	Apple
iPhone (8GB)	0.974	7.5	Apple
iPhone 3G (16GB)	1.38	14.5	Apple
8)	1.38	14.5	Apple iPhone 3G (8GB
mple iPhone 3G S (16GB)	0.96	2.5	Apple
	2.5	Apple iPhone 3G S (32GB)	0.96
Apple iPhone 4 (32GB)	1.17	11.5	
Nokia E61i	0.83	5	
Nokia E63	1.24	13	
Nokia E65	0.74	2	
Nokia E70	0.9	6	
Nokia E71x	1.53	16	
9	Nokia E75	0.99	
0.59	1	Nokia E90 Communicator	

From this data, we nd that Nokia E Series phones from this time period rank higher than Apple iPhones in terms of emitting lower levels of radiation. The rank sum score for the 8 Nokia (Apple) phones is 62 (74). We also compare subsets of these two mobile phone series. For example, we compare the 7 (6; 5; 4; 3; 2; 1) most recently released Nokia E phones in the dataset with the 7 (6; 5; 4; 3; 2; 1) most recently released Apple iPhones. For each of these subsets, Nokia E Series phones also rank better than Apple iPhones under rank sum scoring. Given these subset results, we might expect Simpson Reversals to not occur in this application data.

In this application setting, there are two main ways in which to think of Simpson Reversals. One can think of them in the speci c: Is there an alternative set of data comparing the two phone series such that, when pooled with the original data, yields a Reversal? Alternatively, one can think of them generally: For what proportion of poolings of this data and its ordinal replicate does a strict Reversal arise? Though the speci c question dominates applications in the previous literature on Simpson Reversals, the general question has certain conceptual advantages. Under the general question, one can determine how globally robust a given data is against Reversal when pooled with an ordinal data that individually generates an identical test result. When one ordinally replicates a data set, no new information is introduced by which to evaluate the two groups. By de nition, the original data and its ordinal replicate yield the very same rank sum test result. By considering incidence of Reversal under pooling of the two data sets, one can determine the general robustness of the original result by considering to what extent that result relies upon the interaction of the test itself with scale-variant features of the data. In the present application, therefore, we consider the general question as a means to determine the general robustness of the data against (susceptibility to) Reversal. In so doing, one can characterize the strength of the original result in terms of data scale invariance.

In the empirical exercise, we rst consider the 2 groups and 8 phone types per group case (i.e., the 2 x 8 case). We sort the data from lowest to highest SAR level to obtain SAR rankings for each of the 16 phones. We then add the 8 rank positions of Apple iPhones and the 8 rank positions of Nokia E phones, respectively, to obtain each brand's empirically-observed rank sum score. We then consider each \most-recent sub-sample" of the data. That is, the 2 x 7 case is developed by rank sum scoring the 7 most recently marketed Apple iPhones in the sample against the 7 most recently marketed Nokia E phones. The same procedure was followed inductively to obtain the 2 x n case 8 n 2 f1; 2; 3; : : : ; 6g. For each case, rank sum scores are shown in Table 8. In Table 9, incidence of empirically observed Reversal is reported for each case.

Unlike in our computational treatment, note that a single outcome for F is given (observed) in the empirical treatment. For the empirical application, then, we need only consider all possible poolings of the speci ed sequence, F, and its ordinal replicate, F^0. In the computational section, we observed that the likelihood of a strict Reversal has a high degree of variability across initial sequences. As this application selects a single sequence F based solely on market characteristics of two cellular phone product series (e.g., similar market time period, status as a popular line of phones during that time period, and number of models in series) and not on parametric properties of the underlying data, there was no a priori reason to believe that instances of strict Reversal would occur at all in the application. For two of the 2 x n cases considered, the 2 x 1 and 2 x 2 cases, we have established that Reversals are not possible for any pooling of the data. However, we observe a cluster of Reversals

[11] Haunsperger, D. B., & Saari, D. G. (1991). The Lack of Consistency for Statistical Decision Procedures. The American Statistician, 45(3), 252{255. https://doi.org/10.1080/00031305.1991.10475814

[12] Huang, L., Zalkikar, J., Tiwari, R. (2019). Likelihood-ratio-test methods for drug safety signal detection from multiple clinical datasets. Computational and mathematical methods in medicine, 2019.

[13] Kock, N. (2015). How Likely is Simpson's Paradox in Path Models?: International Journal of E-Collaboration, 11(1), 1{7. https://doi.org/10.4018/ijec.2015010101

[14] Kuss, O. (2015). Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. Statistics in Medicine, 34(7), 1097{1116. https://doi.org/10.1002/sim.6383

[15] Lin, T., Chen, T., Liu, J., & Tu, X. M. (2021). Extending the Mann-Whitney-Wilcoxon rank sum test to survey data for comparing mean ranks. Statistics in Medicine, 40(7), 1705-1717.

[16] Nagaraja, H. N., & Sanders, S. (2020). The Aggregation Paradox in Statistical Rankings. PLOS ONE, Forthcoming.

[17] Pavlides, M. G., & Perlman, M. D. (2009). How Likely Is Simpson's Paradox? The American Statistician, 63(3), 226{233. https://doi.org/10.1198/tast.2009.09007

[18] Pineiro, G., Paruelo, J. M., & Oesterheld, M. (2006). Potential long-term impacts of livestock introduction on carbon and nitrogen cycling in grasslands of Southern South America. Global Change Biology, 12(7), 1267{1284. https://doi.org/10.1111/j.1365-2486.2006.01173.x

[19] Pordanjani, S. R., Kavousi, A., Mirbagheri, B., Shahsavani, A., & Etemad, K. (2021). Spatial analysis and geoclimatic factors associated with the incidence of acute lymphoblastic leukemia in Iran during 2006{2014: An environmental epidemiological study. Environmental Research, 202, 111662.

[20] Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society: Series B (Methodological), 13(2), 238{241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x

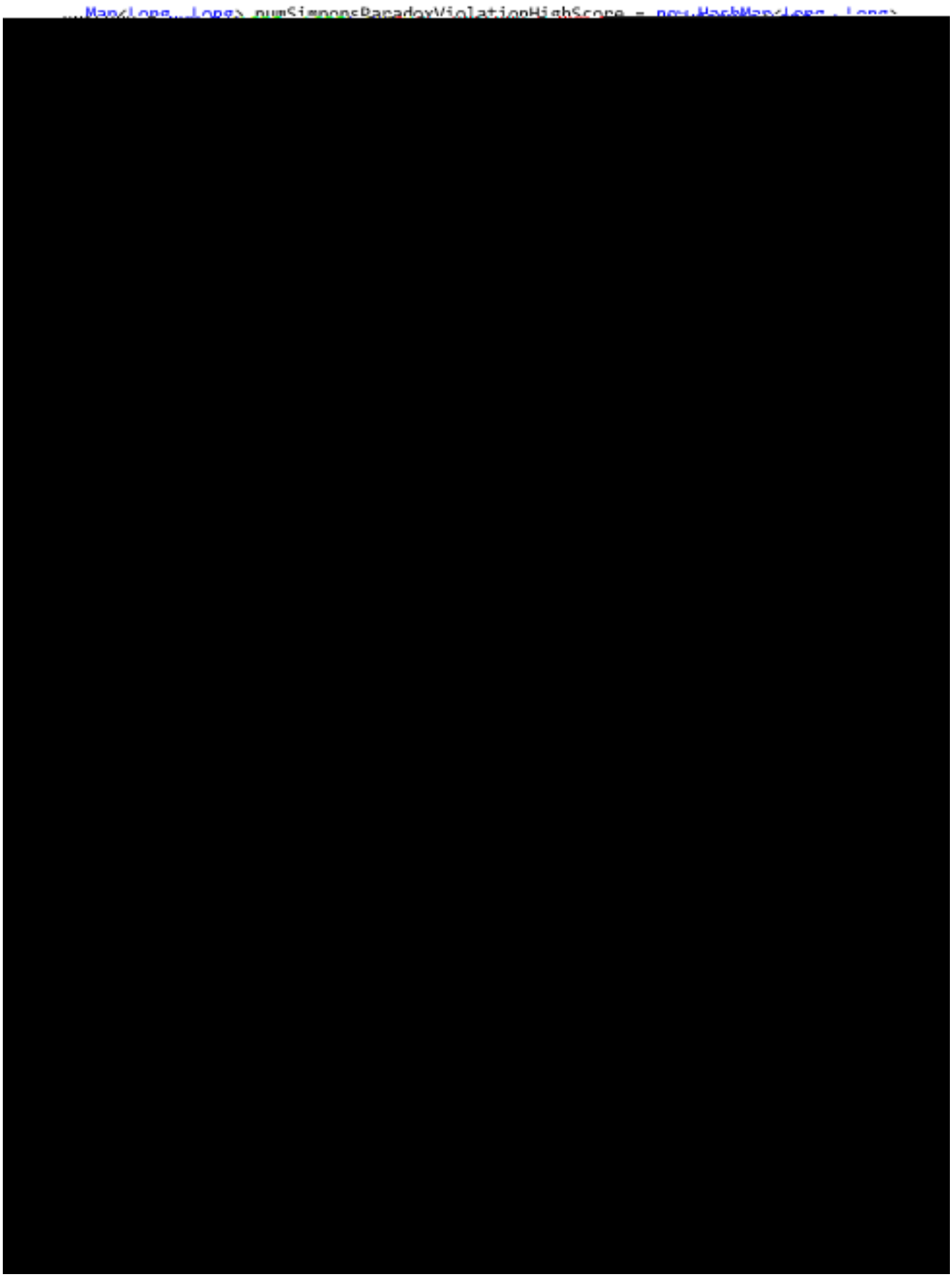
[21] Tran, P., & Waller, L. (2015). Variability in results from negative binomial models for lyme disease measured at di erent spatial scales. Environmental research, 136, 373-380.

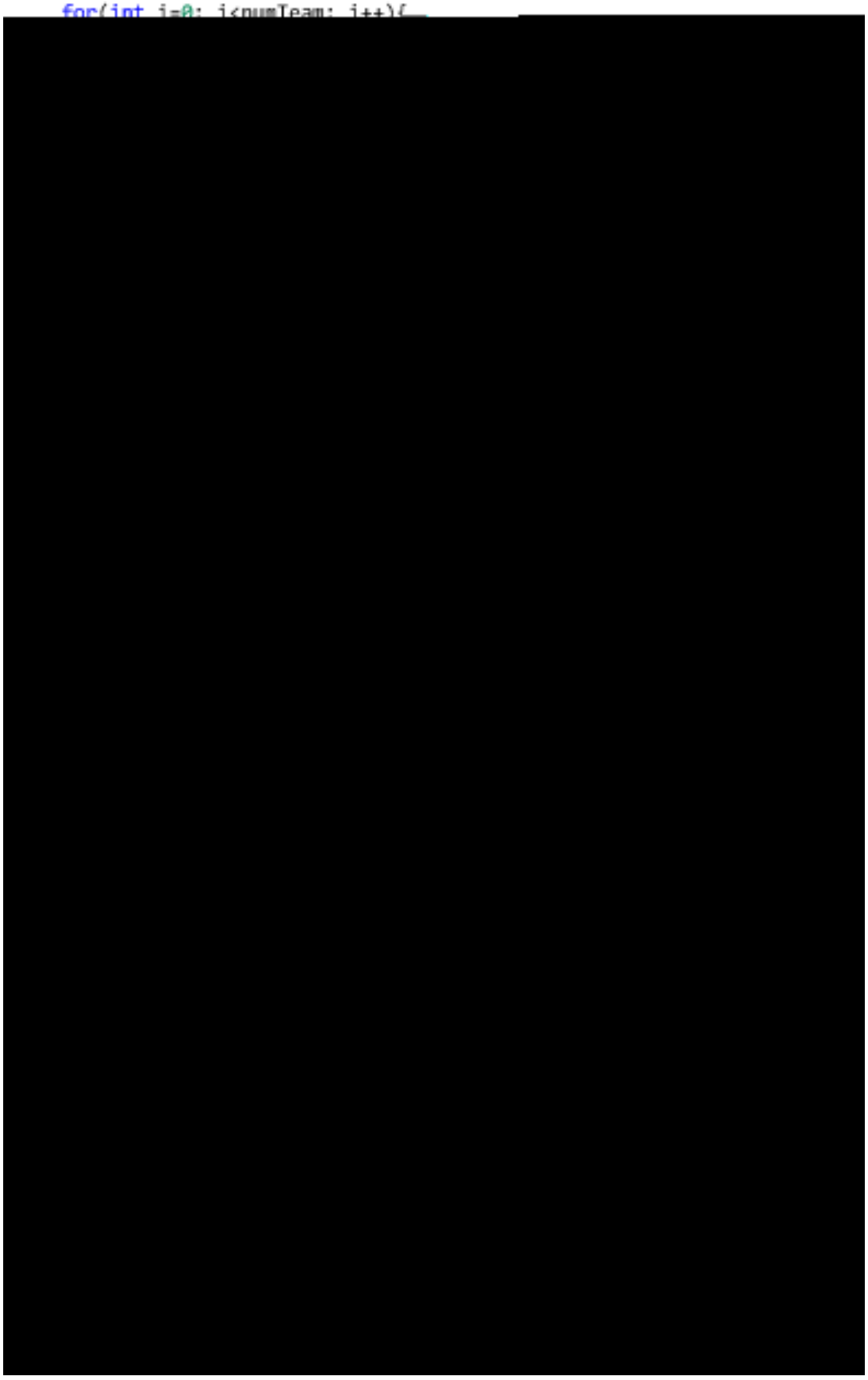
[22] Winner, L. (2021). Cell Phone Radiation Ratings by Model/Brand: users.stat.u .edu/ winner/datasets.html

[23] Yule, G. U. (1903). NOTES ON THE THEORY OF ASSOCIATION OF ATTRIBUTES IN STATISTICS. Biometrika, 2(2), 121{134. https://doi.org/10.1093/biomet/2.2.121

5 Appendix: Computational Code

```
import java.io.BufferedWriter;
import java.io.FileWriter;
import java.io.IOException;
import java.util.Arrays;
```





```
int yScore = 0;
int counter = 0;
for (int i = 0; i < dataPoints.length(); i++) {
    counter++;
    for(int j=0; j<numTeam; j++){
        if(dataPoints.charAt(i) == groups[j]){
```

```
int min=groupsScores[0];
int minIndex = 0;
for (int i = 0; i < numTeam; i++){
//detect tie
}
}
return(groupsScores[minIndex]);
```

```
radox(String dataPoints, int numGroups){
ed = numWinnerChanged;
edPossible = numWinnerChangedPossible;
dentWinner(dataPoints,numGroups);
```

```
private void findSimpsonsPa
    long prevNumWinnerChang
    long prevNumWinnerChang
    char winner = findDepe
```

```
        }

        long topScore = findDependentWinnerScore(dataPoints,numGroups);

        if (numSimponsParadoxViolationHighScore.containsKey(topScore)){
            numSimponsParadoxViolationHighScore.put(topScore, numSimponsParadoxVi
olationHighScore.get(topScore)+deltaNumWinnerChanged);
                                                                    }else{
```

findSimpsonsParadox(dataPoints, numGroups);