RESEARCH ARTICLE

The aggregation paradox for statistical rankings and nonparametric tests

Haikady N. Nagaraja¹, Shane Sanders^{2*}

1 The Ohio State University, Division of Biostatistics, College of Public Health, Columbus, OH, United States of America, **2** Syracuse University, David B. Falk College of Sport & Human Dynamics, Syracuse, NY, United States of America

* sdsander@syr.edu

Abstract

The relationship between social choice aggregation rules and non-parametric statistical tests has been established for several cases. An outstanding, general question at this intersection is whether there exists a non-parametric test that is *consistent upon aggregation* of data sets (not subject to *Yule-Simpson Aggregation Paradox* reversals for any ordinal data). *Inconsistency* has been shown for several non-parametric tests, where the property bears fundamentally upon robustness (ambiguity) of non-parametric test (social choice) results. Using the *binomial*($n, p = 0.5$) random variable CDF, we prove that aggregation of $r(\geq 2)$ constituent data sets—each rendering a qualitatively-equivalent *sign test for matched pairs* result—reinforces and strengthens constituent results (*sign test consistency*). Further, we prove that magnitude of *sign test consistency* strengthens in significance-level of constituent results (*strong-form consistency*). We then find preliminary evidence that *sign test consistency* is preserved for a generalized form of aggregation. Application data illustrate (in)consistency in non-parametric settings, and links with information aggregation mechanisms (as well as paradoxes thereof) are discussed.

OPEN ACCESS

Citation: Nagaraja HN, Sanders S (2020) The aggregation paradox for statistical rankings and nonparametric tests. PLoS ONE 15(3): e0228627. <https://doi.org/10.1371/journal.pone.0228627>

Editor: Alan D Hutson, Roswell Park Cancer Institute, UNITED STATES

Received: June 19, 2019

Accepted: January 20, 2020

Published: March 12, 2020

Copyright: © 2020 Nagaraja, Sanders. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data is publicly source, and the URLs that we used are provided within the paper.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

The relationship between social choice aggregation rules and non-parametric statistical analysis is well-established [see, e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9]. In a related literature, fundamental relationships between discrete choice and non-parametric statistical analysis have been developed [see 10, 11, 12, 13, 14, 15, 16, 17]. While violations of social choice principles can inform us as to the possible paradoxes present in non-parametric tests, the mapping is imperfect due to the issue of statistical significance. Consider a comparison of three groups, where the term “group” refers to sampled elements of the same population. If, for example, a violation of transitivity occurs in a set of raw, ordinal data when comparing three groups pairwise, this does not imply that any sort of paradox is evident in the results of a corresponding set of pairwise significance tests. This distinction is studied prominently in Bargagliotti and Greenwell [9].

In a seminal work, Haunsperger and Saari [4] find that “. . . two or more data sets each may individually support a certain ordering of the samples under [the] *Kruskal-Wallis* [test], yet their union, or aggregate, yields a different outcome” [18, p. 261]. In other words, the *Kruskal-Wallis* (KW) test [19] is not necessarily *consistent upon aggregation* of data. The works of

Haunsperger and Saari [4] and later of Haunsperger [18] provide characterizations as to aggregation paradoxes in non-parametric statistics. These works extend a prominent literature concerning the classic *Yule-Simpson* (association or statistical aggregation) *Paradox* [see 20 or 21 and, e.g., 22, 23, 24, 25, 26, 27, 28, or 29 for later applications] to the non-parametric realm. The *Yule-Simpson Paradox* has far-reaching implications within statistics and econometrics. Wardrop [25] demonstrates that the Paradox has an important bearing upon the existence (non-existence) of the hot-hand fallacy and related *law of small numbers*. Specifically, Wardrop notes that researchers studying whether a hot hand fallacy exists should analyze data at the level of aggregation that (potentially fallacy-prone) fans view the data [see, e.g., Miller and Sanjurjo [30] for recent, related work on the the hot hand fallacy].

Bickel et al. [31] find evidence that the Paradox can influence evaluation of gender bias in graduate admissions data. Specifically, statistically significant results at the program level need not hold when evaluating graduate admissions at the university level. Albers [32] finds a related result in evaluating gender bias in research funding. In light of the *Yule-Simpson Paradox*, these studies establish the importance of analyzing data at the appropriate level of aggregation, whatever it may be. Charig et al. [33] examine treatment effects for small and large kidney stones, both separately and pooled. Consistent with the *Paradox*, they find evidence that treatment efficacy reverses from separate to pooled tests.

Further developing the *Paradox* for non-parametric settings, Haunsperger [18] discovers a computable criterion by which the *consistency (upon aggregation)* of an ordinal data set (when *KW*-tested) can be characterized. Bargagliotti [34] finds that *Bhappkar's V test* [35] and the *Wilcoxon-Mann-Whitney (WMW)* test [Wilcoxon 1945, Mann and Whitney 1947]—two other central non-parametric tests—are also not generally *consistent upon aggregation*. Hao and Houser [38] state, “Both robust and simple to implement, [the WMW test] has gained exceptional popularity among empirical scientists, even social scientists. For example, in 2009 almost half of papers in *Experimental Economics* used the WMW test. Hao and Houser [38] demonstrate that adaptive procedures for the WMW test can lead to “substantial improvements in the ability to detect differences in locations” [p. 1940]. Specifically, Gastwirth [39] and Hogg et al. [40] develop adaptive ranks tests with such a quality. In some respects, the *Yule-Simpson Paradox* relates to the power and sensitivity of given statistical tests, and how these qualities (may) relate to sample size. The *Paradox* also relates to a broader literature on information aggregation and paradoxes in which more information is not always revealed to result in better (expected) outcomes [see, e.g., Kaufmann and Weber [41], Koessler et al. [42], Bennouri et al. [43], Axelrod et al. [44], Hanson et al. [45]]. Of course, information aggregation paradoxes may be symptomatic of the decision-maker, of the information aggregation mechanism, or of both. Bennouri et al. [43] demonstrate that accepted information aggregation mechanisms can vary substantially in terms of effectiveness (i.e., in supporting optimal decisions).

The *KW test* [see, e.g., 46; 47; 48 for more recent descriptions and generalizations of the test] represents a non-parametric statistical analog of the $n(\geq 2)$ group *rank sum* aggregation rule in social choice theory. The *WMW test* represents a related two-sample rank sum test. The *Wilcoxon rank sum test* and the *Mann-Whitney U test* are nominally distinct (rank sum) tests; their respective test statistics are linearly dependent upon one another such that the two tests lead to the same end (in terms of significance results). *Bhappkar's V test*, on the other hand, compares categorically-assigned data across groups. That these non-parametric tests fail to attain general *consistency upon aggregation* is both surprising and problematic. Philosophically, these results are akin to a social aggregation rule that takes as inputs two *Yes* votes (e.g., from an election governed by a *Yes-No* voting rule), aggregates them, and returns *No* as the social choice (outcome). We might rather expect two such primitive data sets to strengthen the case

for the original result upon aggregation. That such an expectation is not met for this set of tests suggests not only that aggregated statistical rankings are potentially arbitrary but that the primitive rankings—or any statistical rankings from the test—are themselves potentially arbitrary. That is, even a primitive data set is an aggregation of data subsets from the primitive set's power set.

While Bargagliotti [34] shows a necessary condition (of the data) to ensure *consistent (upon aggregation)* results for the *KW* and *Bhaskar's V* tests, respectively, the literature has not determined whether *inconsistency upon aggregation* is a general paradox among non-parametric statistical tests. Both the title of Haunsperger and Saari ("The Lack of Consistency for Statistical Decision Procedures") and that of Haunsperger ("Aggregated Statistical Rankings are Arbitrary") certainly leave open the intriguing possibility of a general result. Selvitella [49] further considers the "ubiquity" of the *Yule-Simpson Paradox*. Based upon early results within the literature, Haunsperger [18] writes, "Haunsperger and Saari [4] show that this Simpson-like paradox occurs with many, if not most, statistical decision procedures. Because of the results [therein], we must expect *KW* to be the non-parametric system most immune to such difficulties" [p. 263]. We consider, then, whether general *consistency under aggregation* is an *impossibility* among non-parametric tests. Alternative tests, such as the *sign test for matched pairs*, apply distinct aggregation rules to non-parametric data in order to assign aggregated statistical rankings of groups. Herein, we examine whether the *sign test for matched pairs* (heretofore, "*sign test*") is generally *consistent upon aggregation*, where the *sign test* represents another central non-parametric test with a social choice aggregation rule analog. It is important to note that the *McNemar* [50] test is closely related to the *sign test for matched pairs*. Specifically, the latter test represents an exact test of the former.

The *sign test* uses a *binomial*($n, p = 0.5$) test statistic to pairwise rank two groups from inter-group (inter-sample) comparisons of matched-pair elements. The *sign test* statistic is calculated in a manner that is procedurally-equivalent to the two-candidate case of *Borda rule* aggregation and to the two-candidate case of *majority rule* aggregation (the latter two of which are, then, themselves equivalent to one another in the two-group case). That is to say, each paired-element comparison of the *sign test* yields a mutually exclusive "vote" in favor of one of the two groups. This is equivalent to two-candidate *Borda* or two-candidate *majority rule* voting, whereby each voter makes a binary preference comparison of two candidates and (assuming voting is sincere) casts a mutually exclusive vote in favor of her preferred candidate. The two candidates are then ranked according to *majority rule* type aggregation.

In Section 2 of this paper, we construct a theoretical framework by which to consider whether the *sign test* is *consistent upon aggregation*. In Section 3, we utilize this framework to explore the properties of aggregated statistical rankings from the *sign test*. We ultimately find that statistical rankings from the *sign test* are *consistent upon aggregation*, and we present a general proof of the result. That is, we utilize the cumulative distribution function (cdf) of a *binomial*($n, p = 0.5$) random variable to show generally that the aggregation of two or more data sets—each rendering a *sign test* result of the same quality—can only strengthen the original result (i.e., can only decrease the *p-value* of the *sign test* statistic). This result is shown for any number of aggregated (primitive) data sets. We prove an additional result as to the nature of *sign test consistency*. Namely, the *sign test p-value* exhibits a greater proportional decrease (in aggregation) as the original (primitive data *sign test*) *p-value* itself becomes smaller. In aggregation, then, we expect a greater proportional *p-value* decline given significant *sign test* results for the constituent, primitive data (than for insignificant *sign test* results). This represents something of a *strong form consistency* result, whereby data aggregation reinforces significant *sign test* results to a greater (proportional) degree. Lastly, we construct and define a generalized notion of aggregation to consider the aggregation of data sets of different sample

sizes. We find preliminary evidence that the *sign test* is also *consistent upon generalized aggregation*. Analysis of the generalized form is of great practical importance, as data sets of different sample sizes are often aggregated (e.g., in the case of many longitudinal and unbalanced panel data sets).

The matched pairs nature of the *sign test* is crucial to the result of *consistency upon aggregation*. In the aggregation of ordinal or numerical data with pre-assigned matched pairs, data are not aggregated into an overall outcome sequence but, rather, into a larger set of matched pairs. As such, *sign test* aggregation is reinforcing or *consistent* rather than potentially *inconsistent*. We find that it is possible for a non-parametric test to exhibit *consistency*.

In Section 4, we develop a generic application in which we present a set of ordinal data for two distinct groups. Elements of the data can be paired across group (based, e.g., on some characteristic such as “twin-ness” in a control-treatment study) and analyzed via the *sign test* (for matched pairs). Alternatively, elements for the respective groups may remain unpaired, and elements can then be aggregated into an outcome or rank-sequence such that overall group rankings are generated from a *WMW rank sum test*. Within the application, we demonstrate a case of *inconsistency upon aggregation* for the *WMW test*. We do so to show readers possible characteristics of an ordinal data that is susceptible to *inconsistency*. For the same ordinal data, we demonstrate an example of the (proved, general) *consistency upon aggregation* of the *sign test*. As such, the procedural differences between the two tests are also highlighted (with respect to the *consistency* property). For the application, we aggregate sets of ordinal data twice, thrice, and four times. It is shown that the *sign test p-value* shrinks monotonically in the number of data sets aggregated. This follows from the general results of Section 3. For the same ordinal data, we observe a case in which the *WMW test p-value* rises monotonically in the number of data sets aggregated (as was shown to be possible by Bargagliotti [34]).

Section 5 identifies and develops an empirical application from the sport of high school team cross country running, a sport that uses rank sum scoring to map from an individual outcome sequence (of individual finishing positions) to team scores. The empirical application uncovers evidence of *rank sum “inconsistency”* and *sign test consistency* for high school team cross country meet data. Section 6 concludes.

2 Theoretical set-up and preliminary Lemmas

2.1 Preliminary definitions

Let us begin our exposition with some definitions concerning data aggregation.

Definition of *n*-aggregation of data: Consider a primitive, *n*-element data set, *A*. Let *n*-aggregation involving *A* be an aggregation of *A* and $r - 1$ ($r \in \mathbb{Z}^+$) other data set(s) each of sample size *n* to form an aggregated data set.

Haunsperger and Saari [4] and Haunsperger [18] construct a powerful methodology to assess *consistency upon aggregation* for KW. Namely, they find a condition on a matrix of ordinal data rankings that is equivalent to *consistency* (mutually exclusive with *inconsistency*). As they utilize a matrix approach, their methodology considers balanced sample size data aggregation (i.e., *n*-aggregation). Note that *n*-aggregation is a restricted form of data aggregation, as it does not consider the aggregation of two or more data sets of different sample sizes. The non-parametric aggregation paradox has focused on this restricted form of data aggregation to date (see, e.g., any of the seminal papers mentioned in the introduction). This is perhaps due to the tractability of examining this form of aggregation, as we will observe in the present analysis. In this study, our primary theorems will concern *n*-aggregation of ordinal data. However, we also present preliminary results for a more general version of data aggregation of the following form.

Definition of n_i -aggregation of data: Consider a primitive, n_0 -element data set, A . Let n_i -aggregation involving A be an aggregation of A and $r - 1$ ($r \in \mathbb{Z}^+$) other data set(s) of respective sample sizes $n_1, n_2, \dots, n_{r-1}$ to form an aggregated data set.

For tractability, n -aggregation has been incorporated as a standard approach in examining the *Yule-Simpson Paradox* for ordinal data. However, analysis of the generalized n_i -aggregation form is an important consideration. Often, data sets of different sample sizes are pooled (e.g., in the case of many longitudinal and unbalanced panel data sets).

Let us now define *consistency upon aggregation* largely as in Haunsperger [18]. She states, “A statistical procedure that endows a set of data with an ordinal ranking of the candidates is consistent [upon] aggregation if the aggregate of any $[r]$ sets of data, each of which corresponds to a given ordering of the candidates, gives rise to the same ordering of the candidates for any positive integer $[r]$ ” [p. 264]. We refine the definition slightly to incorporate the notion of statistical significance explicitly as follows.

Definition of *Consistency upon Aggregation*: A statistical procedure that endows a set of data with an ordinal ranking of the candidates is *consistent upon aggregation* if the aggregate of any r sets of data, each of which corresponds to a given statistically-significant ordering of the candidates, gives rise to the same statistically-significant ordering (i.e., at the same α -level) of the candidates for any positive integer r and for all possible α levels.

Thinking of *consistency* from the perspective of significance testing, we can evaluate the *consistency* of a procedure by simply verifying whether it is possible for the p -value of a test result to rise following aggregation. If so, statistically-significant strict orderings can be lost in aggregation such that the procedure is not generally *consistent*.

2.2 Theoretical setup

Let X_n be a *Binomial* random variable with parameters n and $p = 1/2$, and $N = rn$ for some integer $r > 1$. Then, the *sign test* for matched pairs is *consistent upon n -aggregation* if and only if (iff) for any k between $n/2$ and n ,

$$P(X_N \geq rk) < P(X_n \geq k). \quad (1)$$

From (1), it may appear that we are aggregating primitive data sets each of which renders the very same test statistic value, k . In fact, (1) provides a general condition for *consistency upon aggregation*. Let us think of k as the lowest *binomial*($n, p = 0.5$) test statistic value (from tests of given sets of primitive data) such that a given statistically-significant ordering emerges from the primitive data (i.e., at a stipulated α -level). As such, the term rk represents a lower bound on the aggregated data test statistic value, and the left-hand side p -value in (1) represents an upper-bound for the p -value from a test of the aggregated data. Then, (1) states that the p -value for an associated test of the aggregated data will always be less than the highest (but still significant) p -value among the associated primitive data sets.

To establish (1), we first define

$$a_j = \sum_{i=0}^{r-1} a_j(i), \quad b_j = \binom{n}{j}, \quad c_j = \frac{a_j}{b_j}, \quad \text{and} \quad d_j = \frac{\sum_{i=j}^n a_i}{\sum_{i=j}^n b_i}, \quad (2)$$

where

$$a_j(i) = \binom{rn}{rj+i}, \quad i = 0, \dots, (r-1), \quad (3)$$

and $(n-1)/2 < k \leq j \leq n$. Here, a_j counts the distinct number of ways of obtaining between rj and $rj+r-1$ successes in rn Bernoulli trials, and b_j counts the number of ways of obtaining j successes in n Bernoulli trials. The coefficients c_j and d_j are useful for studying the relative variation in a_j and b_j and their upper tail partial sums. We now show that c_j and consequently d_j are strictly increasing in the range of our interest.

Lemma 1 For sequences a_j , b_j , c_j given in (2), and $a_j(i)$ given in (3), the following properties hold.

- (a). For a fixed $i > 0$, $a_j(i)/b_j$ strictly decreases as j increases in the interval $[(n-1)/2 - (i/r), n]$.
- (b). $a_j(0)/b_j$ strictly decreases as j increases in the interval $((n-1)/2, n]$, and when $j = (n-1)/2$ is an integer, $a_j(0)/b_j = a_{j+1}(0)/b_{j+1}$.
- (c). $c_j = a_j/b_j$ strictly decreases for j in the interval $[(n-1)/2, n]$.

Proof. From the expressions for $a_j(i)$ and b_j given in (3) and (2) respectively, we note that

$$\frac{a_j(i)}{b_j} = \frac{(rn)!j!(n-j)!}{(rj+i)!(rn-rj-i)!n!} \quad \text{and} \quad \frac{a_{j+1}(i)}{b_{j+1}} = \frac{(rn)!(j+1)!(n-j-1)!}{(rj+r+i)!(rn-rj-r-i)!n!}.$$

Canceling out the common factors in the above two expressions, we conclude that the condition $a_j(i)/b_j > a_{j+1}(i)/b_{j+1}$ holds iff

$$\frac{(n-j)(r(j+1)+i)}{(j+1)(r(n-j)-i)} \prod_{t=1}^{r-1} \frac{r(j+1)+i-t}{r(n-j)-i-t} > 1, \quad (4)$$

for any $j \leq (n-1)$ in the specified range. The first factor on the left side of (4) above exceeds 1 iff $(n-j)i > -i(j+1)$. For any j , this is true for all $i > 0$, and when $i = 0$, the first factor equals 1. The ratio corresponding to the t th factor in the product exceeds 1 iff $r(j+1)+i-t > r(n-j)-i-t$, or equivalently $j > (n-1)/2 - i/r$. When $j = (n-1)/2 - i/r$ is an integer, the product equals 1, and for $i > 0$ the first factor exceeds 1 and so the strict monotonicity of $a_j(i)/b_j$ holds in $[(n-1)/2 - i/r, n]$; thus the conclusion in part (a) is established.

When $j = (n-1)/2$ is an integer, the first factor as well as the product term equal 1, leading us to the conclusion in part (b), and strict monotonicity holds for $j > (n-1)/2$.

Since a_j is the sum of the $a_j(i)$, and as long as at least one of the $a_j(i)/b_j$ shows strict monotonicity in j , the strict monotonicity of c_j follows. This claim holds for an integer $j = (n-1)/2$ also, completing the proof of part (c).

Lemma 2 For the d_j defined in (2),

$$d_n < d_{n-1} < \cdots < d_{\frac{n}{2}} \text{ for even } n;$$

$$d_n < d_{n-1} < \cdots < d_{\frac{n+1}{2}} < d_{\frac{n-1}{2}} \text{ for odd } n.$$

Proof. We use an induction argument. For this, first we prove that $d_n < d_{n-1}$. Recall that $d_n = a_n/b_n$ and $d_{n-1} = (a_n + a_{n-1})/(b_n + b_{n-1})$, where the a_j and b_j are defined in (2). Hence, the condition: $d_n < d_{n-1}$ is equivalent to the condition:

$$\frac{a_n}{b_n} < \frac{a_n + a_{n-1}}{b_n + b_{n-1}}. \quad (5)$$

Since denominators above are positive, (5) holds iff $a_n b_n + a_n b_{n-1} < a_n b_n + a_{n-1} b_n$, or equivalently,

$$\frac{a_n}{b_n} < \frac{a_{n-1}}{b_{n-1}}.$$

This condition follows from part (c) of Lemma 1 since $c_j = a_j/b_j$.

Next, assuming $d_{j+1} < d_j$, we will prove that $d_j < d_{j-1}$. From the definition of d_j given in (2), we conclude that

$$\begin{aligned} d_{j+1} < d_j &\Leftrightarrow \frac{\sum_{i=j+1}^n a_i}{\sum_{i=j+1}^n b_i} < \frac{\sum_{i=j}^n a_i}{\sum_{i=j}^n b_i} \\ &\Leftrightarrow b_j \sum_{i=j+1}^n a_i < a_j \sum_{i=j+1}^n b_i, \end{aligned} \quad (6)$$

upon cross-multiplication and cancellation of common terms. Similarly, the condition: $d_j < d_{j-1}$ is equivalent to the condition:

$$b_{j-1} \sum_{i=j}^n a_i < a_{j-1} \sum_{i=j}^n b_i. \quad (7)$$

The left side sum in (7) is

$$b_{j-1} a_j + b_{j-1} \sum_{i=j+1}^n a_i = b_{j-1} a_j + \frac{b_{j-1}}{b_j} b_j \sum_{i=j+1}^n a_i < b_{j-1} a_j + \frac{b_{j-1}}{b_j} a_j \sum_{i=j+1}^n b_i, \quad (8)$$

where the inequality above follows from (6). From part (c) of Lemma 1, it follows that $c_j < c_{j-1}$; that is, $(a_j/b_j) < (a_{j-1}/b_{j-1})$ or $a_j b_{j-1} < a_{j-1} b_j$. Hence, the upper bound in (8),

$$b_{j-1} a_j + \frac{b_{j-1}}{b_j} a_j \sum_{i=j+1}^n b_i < b_j a_{j-1} + \frac{b_j a_{j-1}}{b_j} \sum_{i=j+1}^n b_i = a_{j-1} \sum_{i=j}^n b_i;$$

this is the right side of the (strict) inequality in (7). This strict inequality (in (7)) also holds for $j-1 = (n-1)/2$ for odd n , and $j-1 = n/2$ for even n . This establishes our claim.

Remark 1. This monotonicity property of the ratio of the partial sums is perhaps a well-known result in the “Inequalities” literature. We have not found a convenient reference, and have chosen to prove it with elementary principles; we have also fine-tuned the result for the even n case.

3 Main results

Theorem 1 Consistency upon Aggregation: Let X_n be a Binomial random variable with parameters n and $p = 1/2$. Then, with $N = rn$, the inequality in (1) holds for all integers $k \geq n/2$, and integers $r \geq 2$.

Proof. In terms of the coefficients introduced in (2) and (3),

$$P(X_N \geq rk) = 2^N \sum_{j=k}^n a_j$$

and

$$P(X_n \geq k) = 2^n \sum_{j=k}^n b_j,$$

where $N = rn$. The inequality in (1) can be expressed as

$$\frac{P(X_N \geq rk)}{P(X_n \geq k)} \equiv \left(\frac{1}{2}\right)^{(r-1)n} d_k < 1. \quad (9)$$

We have seen in Lemma 2 that d_k strictly increases as k decreases toward $n/2$. Thus, the maximum value for the ratio of the two binomial probabilities in (9) is attained with $k = (n + 1)/2$ for odd n and $k = n/2$ for even n . We now show that this maximum value is less than 1 for both odd and even n .

When n is odd and $n = 2m - 1$, we take $k = m = (n + 1)/2$ in (9). By the symmetry of the binomial probabilities, $P(X_n \geq m) = 1/2$.

Now $N(= rn)$ may be even or odd. When N is even, $r(\geq 2)$ is also even, and $rn/2$ is the median of X_N . Further, $P(X_N \geq rn/2) > 1/2$ and $P(X_N \geq rn/2 + 1) < 1/2$. Thus,

$$P(X_N \geq rm) = P\left(X_N \geq \frac{r(n+1)}{2}\right) \leq P\left(X_N \geq \frac{rn}{2} + 1\right) < \frac{1}{2},$$

where the first inequality holds since $r/2 \geq 1$. When $N(= rn)$ is odd, $r(\geq 3)$ is also odd, and $P(X_N \geq (rn + 1)/2) = 1/2$. Thus,

$$P(X_N \geq rm) = P\left(X_N \geq \frac{r(n+1)}{2}\right) < P\left(X_N \geq \frac{rn+1}{2}\right) = \frac{1}{2},$$

where the strict inequality holds since $(rn + r)/2$ and $(rn + 1)/2$ are both integers and the difference $(r - 1)/2 \geq 1$. Thus, the strict inequality in (9) holds for all $k \geq n/2$ when n is odd.

When n is even and $n = 2m$, we take $k = n/2 = m$, and note that N is always even. By the symmetry of the binomial probabilities,

$$P(X_n \geq m) = \frac{1}{2} + \frac{1}{2}P(X_n = m);$$

$$P(X_N \geq rm) = \frac{1}{2} + \frac{1}{2}P(X_N = rm).$$

Thus, the strict inequality in (9) holds iff

$$P(X_N = rm) < P(X_n = m). \quad (10)$$

We will now establish (10). Note that X_N has the same distribution as $X_n + Y$ where Y is Binomial with parameters $N - n$ and $p = 1/2$, and X_n and Y are independent.

$$P(X_N = rm) = \sum_{i=0}^n P(X_n = i)P(Y = rm - i).$$

Since $\binom{n}{i}$ is largest when $i = n/2 = m$, so is $P(X_n = i)$. Hence,

$$P(X_N = rm) < P(X_n = m) \sum_{i=0}^n P(Y = rm - i) \leq P(X_n = m),$$

which establishes (10). That is, (9) holds for all $k \geq n/2$ when n is even as well.

A similar result on the lower tail probabilities can be established by symmetry. As such, we have shown that any significant sign test result is *consistent upon n-aggregation*. This result has important implications for classic results (e.g., by [31] and [33]). If we apply the sign test to the underlying data of these studies, while normalizing the sample size to equality across group and across treatment, it is not possible for an aggregation paradox to occur. Later in the paper, we will consider whether the sign test is susceptible to instances of aggregation paradox given unequal sample sizes across group. We can refine Theorem 1 by observing that we have proved a great deal more. In fact, we have established the following.

Theorem 2 Strong Form Consistency upon Aggregation: Let X_N be a Binomial random variable with parameters $N = rn$ and $p = 1/2$ where $n \geq 1$ and $r \geq 2$ is an integer. Then, as the integer k decreases in $[n/2, n]$, $P(X_N \geq rk)/P(X_n \geq k)$ monotonically increases with a strict upper bound of 1.

That the ratio is bounded by 1 follows from Theorem 1. Theorem 2 tells us that the *n-aggregated* data associated *p-value* rises toward the primitive data associated *p-value* as the primitive data *p-value* rises. In other words, *n-aggregation* creates a greater proportional decrease in the *sign test p-value* as the original (primitive) *p-value* itself becomes smaller. In aggregation, then, we expect a greater proportional *p-value* decline given significant *sign test* results for the constituent, primitive data (than for insignificant *sign test* results). This represents something of a *strong form consistency* result, whereby data aggregation reinforces significant *sign test* results to a greater (proportional) degree. A more general result of the following form is highly desirable.

Theorem 3 (Conjecture) Let X_N denote a Binomial random variable with parameters N and $p = 1/2$. Then for positive integers $n_1 < n_2$ and constant $c \in [1/2, 1]$ such that cn_1 and cn_2 are integers, the following inequality holds.

$$P\left(\frac{X_{n_2}}{n_2} \geq c\right) < P\left(\frac{X_{n_1}}{n_1} \geq c\right). \quad (11)$$

We are able to establish this claim for the boundary values of c (that is, $c = 1/2$ or 1) using the arguments presented earlier, and conjecture that the result is true for other c values in $[1/2, 1]$. Limited computational work supports this conjecture. Further, a large-sample approximation, described below, leads us to believe in the conjecture, especially when n_1 and n_2 are large and far apart.

From central limit theorem, as $n \rightarrow \infty$, we know that $\sqrt{n}(X_n/n - 1/2) \xrightarrow{d} N(0, 1/4)$. Further, the binomial distribution, being already unimodal and symmetric (as $p = 1/2$), the convergence is fast. So, if n is large we have the following approximation:

$$P\left(\frac{X_n}{n} \geq c\right) = P\left(\sqrt{n}\left(\frac{X_n}{n} - \frac{1}{2}\right) \geq \sqrt{n}\left(c - \frac{1}{2}\right)\right) \approx 1 - \Phi\left(\sqrt{4n}\left(c - \frac{1}{2}\right)\right), \quad (12)$$

where Φ is the cdf of a standard normal random variable. This “close” approximation to the tail probability strictly decreases as n increases as long as $c > 1/2$ and hence we believe (11) holds for large $n_1 < n_2$.

Remark 2. When $c = 1/2$ and n_1 and n_2 are odd integers, the strict inequality claimed in (11) does not hold and each of the probabilities equals 1/2 as for an odd integer n ,

$$P(X_n \geq n/2) = P(X_n \geq (n+1)/2) = 1/2.$$

Also, it follows from (12) that, irrespective of the odd or even nature of n , the limiting value of this probability is 1/2.

4 Application 1: Demonstrating aggregation of statistical rankings for *Sign* and *WMW* tests

4.1 Tests of a primitive, ordinal data set

Consider two groups, A and B . Each group consists of nine elements or data points (i.e., $\frac{n}{2} = 9$ or $n = 18$). That is, $A = \{a_1, a_2, \dots, a_9\}$ and $B = \{b_1, b_2, \dots, b_9\}$ such that the primitive (unaggregated) data for the *WMW rank sum test* set-up is an outcome sequence of 18 ordinal (rank-ordered) data points. For the application, let

$$F_{AB} = a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9, a_9$$

represent the ordinal data (sequence). For the *sign test* set-up, each $a_i \in A$ is matched with a counterpart $b_j \in B$ according to the matching criterion. As we are *a priori* unsure (notationally) which b_j will be matched with each given a_i , we say that each a_i is matched with some $b_{\pi(i)}$, where $\pi: D \rightarrow D$ is defined as a bijection between $D = \{1, 2, 3, \dots, 8, 9\}$ and itself such that element $a_i \in A$ and element $b_{\pi(i)} \in B$ uniquely pair with one another for all $i \in D$. As such, the following matched pairs data set is generated (uniquely in the present example) from the outcome sequence F_{AB} :

$$M_{AB} = \{(a_1, b_{\pi(1)}), (a_2, b_{\pi(2)}), (a_3, b_{\pi(3)}), \dots, (b_{\pi(9)}, a_9)\}.$$

The set M_{AB} is a set of rank-ordered (matched) pairs of elementary data points. That is to say, M_{AB} is a set whose elements are outcome sequences on matched pairs. If an element of M_{AB} is represented as (a_i, b_j) , this is equivalent to finding that $a_i \succ b_j$ (" a_i ranks higher than b_j ") in outcome sequence F . If an element of M_{AB} is represented as (b_j, a_i) , this is equivalent to saying that $b_j \succ a_i$ in outcome sequence F_{AB} . As such, F_{AB} and M_{AB} are different representations of the same ordinal data set. Whereas F_{AB} is a sequence representation of the data, M_{AB} is a matched-pairs representation. For the application data—but not generally— M_{AB} is uniquely derived from F_{AB} . That is to say, the elements of F_{AB} are sequenced such that the ordering of any given matched pair in F_{AB} is invariant to the specific match formed. That is, when $i \in \{1, 2, 3, \dots, 8\}$, we know that $a_i \succ b_j$ regardless of which j is paired with i . When $i = 9$, we know that $b_j \succ a_i$ regardless of which j is paired with i . For the application data (but not generally), one can uniquely derive M_{AB} from F_{AB} . However, it is never possible to derive F_{AB} uniquely from M_{AB} . That is to say, F_{AB} sometimes maps to a unique M_{AB} . Given only M_{AB} , however, one can never map to a unique F_{AB} . In such a case, one is always missing some rank comparison information needed to support a super-ordering or outcome sequence of the overall data. Specifically, one cannot obtain rank comparisons of unmatched elements (a_i and $b_{\pi(j)}$, $j \neq i$) from M_{AB} .

From F_{AB} , we can calculate rank sum scores for A and B , $R_{F_{AB}}(A)$ and $R_{F_{AB}}(B)$. Generally, we define $R_{F_{XY}}(\cdot)$ as follows. Consider two groups, X and Y , and let F_{XY} represent an outcome sequence between elements of the two groups. For each element $x_i \in X$, let $x_i^+ = \{a \in F_{XY} | a \succ x_i\}$. Then, the rank of x_i in the sequence F_{XY} , $r(x_i | F_{XY})$, equals $|x_i^+| + 1$, and the rank sum score for group X given F_{XY} is represented as $R_{F_{XY}}(X) = \sum_{x_j \in X} r(x_j | F_{XY})$. Given the outcome sequence F_{AB} specified earlier in this application, then, we obtain

$$R_{F_{AB}}(A) = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 18 = 54; R_{F_{AB}}(B) = (18 \times 19/2) - 54 = 117,$$

and apply the *WMW rank sum test*. Let $R_{F_{AB}} = \max\{R_{F_{AB}}(A), R_{F_{AB}}(B)\}$. Then, the upper-tail *WMW rank sum p-value* for this sample difference, $P(R_{F_{AB}} \geq 117 | n = 18)$, is equal to 0.0027. Then, we conclude from a *WMW test* of the primitive data that $A \succ_{\alpha = 0.01} B$, where the

notation “ $\succ_{\alpha = 0.01}$ ” reads “ranks significantly higher than at the $\alpha = 0.01$ significance level.” In other words, A has a significantly lower rank sum score than B such that we conclude from F_{AB} that group A ranks significantly higher than group B at the assigned significance level.

Let us consider the same data in matched pairs form, M_{AB} . Of the 9 matched pairs, an element of A outranks the corresponding element of B for 8 of 9 matched pairs. Formally, let us define the *sign test* statistic value for A , $S_{M_{AB}}(A)$, as the number of elements $a_i \in A$ such that $a_i \succ_{\pi_i} b_{\pi_i}$: $S_{M_{AB}}(A) = |\{a_i \in A : a_i \succ_{\pi_i} b_{\pi_i}\}|$. Therefore, we obtain the *sign test* values for the primitive data as $S_{M_{AB}}(A) = 8$ and $S_{M_{AB}}(B) = 1$ and let $S_{M_{AB}} = \max\{S_{M_{AB}}(A), S_{M_{AB}}(B)\}$.

The upper-tail *sign test* for matched pairs *p-value* for this sample difference, $P(S_{M_{AB}} \geq 8 \mid \frac{n}{2} = 9)$ is equal to 0.0098. Then, we conclude from a *sign test* for matched pairs of the primitive data that $A \succ_{\alpha = 0.01} B$. For this primitive application data, the *WMW test* and the *sign test* each provide corresponding results at all standard significance levels.

4.2 Tests of aggregated, ordinal data

Now, let us replicate the (ordinality of the) primitive data once, combine the primitive data and its ordinal replicate, and re-test the aggregated data. One should note that an ordinally-replicated data set may be quite different numerically from the primitive (source) data from which it arises such that two ordinally-identical data sets can combine to form various, feasible aggregated sequences. The aggregated outcome sequence consists of 36 rank-ordered elements. One possible aggregated outcome sequence is:

$$FF'_{AB} = a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9, F'_{AB}, a_9$$

where $F'_{AB} = a'_1, a'_2, a'_3, a'_4, a'_5, a'_6, a'_7, a'_8, b'_1, b'_2, b'_3, b'_4, b'_5, b'_6, b'_7, b'_8, b'_9, a'_9$ is ordinally equivalent to F_{AB} . In this aggregated sequence, we have simply spliced the replicate sequence, F'_{AB} , into F_{AB} such that F'_{AB} is strung (consecutively) between the penultimate and ultimate elements of the original sequence. The implication of this aggregated outcome sequence is that—for the most part—the numerical values of the original data outrank those of the ordinal-replicate data (despite the identical ordinal quality of the two data sets). From FF'_{AB} , we find that

$$R_{FF'_{AB}}(A) = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 18 + 19 + 20 + 21 + 22 + 23 + 24 + 25 + 35 + 36 = 279;$$

$$R_{FF'_{AB}}(B) = (36 \times 37/2) - 279 = 387.$$

The upper-tail *WMW rank sum p-value* for this sample difference, $P(R_{FF'_{AB}} \geq 387 \mid N = 2n = 36)$, is equal to 0.0438. At the $\alpha = 0.01$ significance level, then, we conclude from a *WMW test* of the aggregated data that $A \sim_{\alpha = 0.01} B$ (i.e., that A and B are rank-indistinguishable at the $\alpha = 0.01$ significance level), where “ $\sim_{\alpha = 0.01}$ ” reads “is rank-indistinguishable to at the $\alpha = 0.01$ significance level.”

We now consider a *sign test* for matched pairs upon the same aggregated data. Specifically, we combine M_{AB} and its replicate into the same data set to obtain the set MM'_{AB} ,

$$\{(a_1, b_{\pi(1)}), (a_2, b_{\pi(2)}), (a_3, b_{\pi(3)}), \dots, (b_{\pi(9)}, a_9), (a'_1, b'_{\pi(1)}), (a'_2, b'_{\pi(2)}), (a'_3, b'_{\pi(3)}), \dots, (b'_{\pi(9)}, a'_9)\}.$$

As M_{AB} is an unordered set of matched pairs (i.e., the set of elements is unordered, whereas each element is itself an ordered pair), the aggregation of M_{AB} and its ordinal replicate, M'_{AB} , is also an unordered set of matched pairs. As such, MM'_{AB} features the same matched pair ordinal comparisons as does M_{AB} , while representing each such pair twice. Matches for the *sign test* are pre-assigned at the primitive data level according to the (also pre-assigned) matching

criterion (e.g., “twinness”). Therefore, matched pairings are preserved in aggregation. This can be said in a more straightforward manner: The identity of one’s twin (match), once assigned, is invariant to the level of data aggregation. Matched pairing preservation in data aggregation is an important property that causes the *sign test* to be distinct from independent sample non-parametric tests. As the profile of matched pairings is itself a feature of the primitive data in the case of the *sign test* (in much the same way that the primitive data outcome sequence is a feature of the primitive data in the case of a *WMW test*), one would not be strictly aggregating the primitive data if one were to permute the matched pairings in aggregation. Rather, one would thereby change a feature of the primitive data at the same time (in the same manner that a re-ordering of primitive data would go beyond aggregation in the case of a *WMW test*). Given this feature of the *sign test*, the matched pair elements of M_{AB} and M'_{AB} , respectively, are not rank-compared with one another (across primitive group) within a sign test analysis of MM'_{AB} .

Of the eighteen matched pairs in the once ordinaly-replicated data, an element of A out-ranks the corresponding element of B for sixteen matched pairs. Therefore, we have the following sign-test statistic for the aggregated data:

$$S_{MM'_{AB}}(A) = 16 \quad S_{MM'_{AB}}(B) = 2.$$

Then, the upper-tailed *sign test* for matched pairs *p-value* for this test statistic value, $P(S_{MM'_{AB}} \geq 16)$ is equal to 0.000484. Then, we conclude from a *sign test* for matched pairs of the aggregated data that $A \succ_{\alpha=0.01} B$. For the once ordinaly-replicated data, the *sign test* retains significance at the $\alpha = 0.01$ level. Consistent with our general results, the *sign test p-value* diminishes with data aggregation in this case. On the other hand, the *WMW test* loses significance at the $\alpha = 0.01$ level when applied to the once ordinaly-replicated data.

To consider the result of Theorem 2, we alter M_{AB} as follows.

$$\tilde{M}_{AB} = \{(\tilde{a}_1, \tilde{b}_{\pi(1)}), (\tilde{a}_2, \tilde{b}_{\pi(2)}), (\tilde{a}_3, \tilde{b}_{\pi(3)}), \dots, (\tilde{b}_{\pi(8)}, \tilde{a}_8), (\tilde{b}_{\pi(9)}, \tilde{a}_9)\}.$$

We also alter M'_{AB} as follows.

$$\tilde{M}'_{AB} = \{(\tilde{a}'_1, \tilde{b}'_{\pi(1)}), (\tilde{a}'_2, \tilde{b}'_{\pi(2)}), (\tilde{a}'_3, \tilde{b}'_{\pi(3)}), \dots, (\tilde{b}'_{\pi(8)}, \tilde{a}'_8), (\tilde{b}'_{\pi(9)}, \tilde{a}'_9)\}.$$

Aggregating $\tilde{M}_{AB}$ and $\tilde{M}'_{AB}$, we obtain $\tilde{M}\tilde{M}'_{AB}$.

$$\tilde{M}\tilde{M}'_{AB} = \{(\tilde{a}_1, \tilde{b}_{\pi(1)}), (\tilde{a}_2, \tilde{b}_{\pi(2)}), (\tilde{a}_3, \tilde{b}_{\pi(3)}), \dots, (\tilde{b}_{\pi(8)}, \tilde{a}_8), (\tilde{b}_{\pi(9)}, \tilde{a}_9), (\tilde{a}'_1, \tilde{b}'_{\pi(1)}), (\tilde{a}'_2, \tilde{b}'_{\pi(2)}), (\tilde{a}'_3, \tilde{b}'_{\pi(3)}), \dots, (\tilde{b}'_{\pi(8)}, \tilde{a}'_8), (\tilde{b}'_{\pi(9)}, \tilde{a}'_9)\}.$$

Given $\tilde{M}_{AB}$ and $\tilde{M}'_{AB}$, we have the following values for the *sign test* statistic:

$$\begin{aligned} S_{\tilde{M}_{AB}}(A) &= S_{\tilde{M}'_{AB}}(A) = 7, & S_{\tilde{M}_{AB}}(B) &= S_{\tilde{M}'_{AB}}(B) = 2; \\ S_{\tilde{M}\tilde{M}'_{AB}}(A) &= 14 & S_{\tilde{M}\tilde{M}'_{AB}}(B) &= 4. \end{aligned}$$

From our previous analysis of M_{AB} and MM'_{AB} , we have that:

$$P(S_{M_{AB}} \geq 8) = 0.0098; P(S_{MM'_{AB}} \geq 16) = 0.000484.$$

With respect to Theorem 2, then, we have that

$$\frac{P(S_{MM'_{AB}} \geq 16)}{P(S_{M_{AB}} \geq 8)} = \frac{0.000484}{0.0098} \approx 0.0494.$$

We also have that

$$\frac{P(S_{MM'_{AB}} \geq 14)}{P(S_{M_{AB}} \geq 7)} = \frac{0.009211}{0.04779} \approx 0.193.$$

This result illustrates the (general) property shown in Theorem 2. Namely, the *sign test p-value* exhibits a greater proportional decrease when replicated as the *p-value* for the *sign test* of the primitive data decreases (i.e., for more significant *sign test* results of the primitive data). Not only does the *sign test* exhibit *consistency upon aggregation*. *Sign test* results are more strongly reinforced in aggregation—in terms of proportional *p-value* decrease—as the original *sign test p-value* (of the primitive data) decreases.

Returning to F_{AB} , M_{AB} , FF'_{AB} , and MM'_{AB} , we iterate this aggregation to combine three and then four instances of the primitive ordinal data as follows.

The case of 3 Ordinal Replicates Aggregated. We now take

$$FF'F''_{AB} = a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9, a'_1, a'_2, a'_3, a'_4, a'_5, a'_6, a'_7, a'_8, b'_1, b'_2, b'_3, b'_4, b'_5, b'_6, b'_7, b'_8, b'_9, F''_{AB}, a'_9, a_9$$

where $F''_{AB} = a''_1, a''_2, a''_3, a''_4, a''_5, a''_6, a''_7, a''_8, b''_1, b''_2, b''_3, b''_4, b''_5, b''_6, b''_7, b''_8, b''_9, a''_9$ is ordinally equivalent to F_{AB} and therefore to F'_{AB} , and

$$MM'M''_{AB} = \{(a_1, b_{\pi(1)}), (a_2, b_{\pi(2)}), (a_3, b_{\pi(3)}), \dots, (b_{\pi(9)}, a_9), (a'_1, b'_{\pi(1)}), (a'_2, b'_{\pi(2)}), (a'_3, b'_{\pi(3)}), \dots, (b'_{\pi(9)}, a'_9), (a''_1, b''_{\pi(1)}), (a''_2, b''_{\pi(2)}), (a''_3, b''_{\pi(3)}), \dots, (b''_{\pi(9)}, a''_9)\}.$$

For this case, the *WMW test p-value* is 0.122, and the *sign test p-value* is 0.000027. The *WMW test p-value* has risen once again in level of aggregation, whereas the *sign test p-value* has once again fallen.

The case of 4 Ordinal Replicates Aggregated. Now take

$$FF'F''F'''_{AB} = a_1, a_2, \dots, a_7, a_8, b_1, b_2, \dots, b_8, b_9, a'_1, a'_2, \dots, a'_7, a'_8, b'_1, b'_2, \dots, b'_8, b'_9, a''_1, a''_2, \dots, a''_7, a''_8, b''_1, b''_2, \dots, b''_8, b''_9, F'''_{AB}, a'_9, a_9$$

where $F'''_{AB} = a'''_1, a'''_2, a'''_3, a'''_4, a'''_5, a'''_6, a'''_7, a'''_8, b'''_1, b'''_2, b'''_3, b'''_4, b'''_5, b'''_6, b'''_7, b'''_8, b'''_9, a'''_9$ is ordinally equivalent to F_{AB} , F'_{AB} , and therefore to F''_{AB} , and

$$MM'M''M'''_{AB} = \{(a_1, b_{\pi(1)}), (a_2, b_{\pi(2)}), (a_3, b_{\pi(3)}), \dots, (b_{\pi(9)}, a_9), (a'_1, b'_{\pi(1)}), (a'_2, b'_{\pi(2)}), (a'_3, b'_{\pi(3)}), \dots, (b'_{\pi(9)}, a'_9), (a''_1, b''_{\pi(1)}), (a''_2, b''_{\pi(2)}), (a''_3, b''_{\pi(3)}), \dots, (b''_{\pi(9)}, a''_9), (a'''_1, b'''_{\pi(1)}), \dots, (b'''_{\pi(9)}, a'''_9)\}.$$

For this case, the *WMW test p-value* is 0.209, and the *sign test p-value* is less than 0.00001. The *WMW test p-value* has risen once again in level of aggregation, whereas the *sign test p-value* has once again fallen. The overall significance testing results for the primitive and aggregated data are represented in Table 1.

Why does the WMW test display inconsistency upon aggregation? Boudreau et al. [52] explore the social choice properties of rank sum scoring. They find that rank sum scoring approaches may invite various, related (social choice) paradoxes when a group with a superior

Table 1. Application 1 statistical test *p*-values.

Test	Data Set	<i>p</i> -value
WMW Rank Sum	F_{AB}	0.0027
WMW Rank Sum	FF'_{AB}	0.0438
WMW Rank Sum	$FF'F''_{AB}$	0.122
WMW Rank Sum	$FF'F''F'''_{AB}$	0.209
Sign Test for Matched Pairs	M_{AB}	0.0098
Sign Test for Matched Pairs	MM'_{AB}	0.000484
Sign Test for Matched Pairs	$MM'M''_{AB}$	0.000027
Sign Test for Matched Pairs	$MM'M''M'''_{AB}$	<0.00001

<https://doi.org/10.1371/journal.pone.0228627.t001>

rank sum “falls behind” (i.e., has fewer elements represented) at some point in the outcome sequence. In the application, group *A* has a superior rank sum in the primitive outcome sequence. However, group *A* also possesses the weakest overall element in the primitive outcome sequence. This makes group *A*’s significant rank superiority in the primitive outcome sequence vulnerable upon aggregation. In the aggregated data, if sufficiently many elements position after element b_9 but before element a_9 (as did occur), this particular interposition disproportionately inflates group *A*’s rank sum score. Such an outcome is not possible for matched pairs ordinal data aggregation. In that case, pairs remain assigned such that the primitive ordinal data and its replicate(s) reinforce one another.

5 (Empirical) Application 2: The aggregation paradox in ranking cross-country running team performance

As in Hammond [51] and Boudreau et al. [52], we use the setting of team cross country races to understand aggregation properties of ordinal data. In a cross country running meet, rank sum scoring is used as the official scoring methodology to assign team rankings. Boudreau et al. [52] describe the scoring methodology in detail:

In cross country running, for example, teams are compared on the performances of individual runners in a race; the outcome sequence is the overall ranking as to how runners finish. Rank sum scoring means that each group (team) receives a number of points equal to an element-wise ranking in the outcome sequence (1 for 1st place, 2 for 2nd, and so on), and comparison of groups is based on the sum of their element scores: groups with lower scores rank above those with higher scores [p. 220].

In the present application, we have identified two high school cross country programs for comparison: Carmel High School (Carmel, Indiana) and Fishers High School (Fishers, Indiana). We consider these two schools because their respective Varsity (Boys) teams competed against one another during the 2016 season, as did their respective Junior Varsity (Boys) teams. The two meets—the Varsity Indiana High School Athletic Association Noblesville Sectional meet and the Junior Varsity Hamilton County meet—occurred on the same running course (White River Elementary School Course; Noblesville, Indiana) such that the aggregation of the two data sets is valid (i.e., runners across meets faced the same basic performance elements). Moreover, Carmel’s Varsity and Junior Varsity teams won each respective meet by the very same convincing margin when rank sum scored against Fishers without the consideration of third-party teams (as in the *WMW test*). That is to say, respective pairwise rank sum scores were the same for each school in each race. Lastly, this particular setting was chosen

because both the Varsity and Junior Varsity meets have been publicly recorded on *Athletic.net* (from which we obtained the data). It is typically difficult to find publicly-available Junior Varsity meet results. The results of these two cross country meets can be found at <https://www.athletic.net/CrossCountry/meet/119851/results/481308> and at <https://www.athletic.net/CrossCountry/Results/Meet.aspx?Meet=119845&show=all> respectively.

This setting provides something of a natural experiment by which we can compare the two Cross Country programs at different levels of aggregation (i.e., a comparison of Varsity teams, a comparison of Junior Varsity teams, and an aggregated comparison of the two programs across the levels of competition) under the same basic competitive conditions. From the finishing time results, we obtain the observed primitive outcome sequence for the Varsity meet as

$$F_V(C, S) = c_1, c_2, s_1, c_3, c_4, c_5, s_2, c_6, s_3, c_7, s_4, s_5, s_6, s_7,$$

where $F_V(C, S)$ symbolizes the observed Varsity race outcome sequence for our two Varsity teams, $C = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7\}$ is a set representing the 7-element (7-runner) Carmel Varsity team, and $S = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$ is a set representing the 7-element (7-runner) Fishers Varsity team. The resulting rank sum scores are

$$R_{F_V}(C) = 36 \quad \text{and} \quad R_{F_V}(S) = 69,$$

and the upper-tail *WMW rank sum p-value* for this sample difference, $P(R_{F_V} \geq 69 | n = 14)$, is equal to 0.019. In a *WMW rank sum test* of the Varsity data, then, we conclude that $C \succ_{\alpha = 0.025} S$. At the $\alpha = 0.025$ significance level or any higher α -level, that is, we conclude that Carmel's 2016 Varsity team was faster than Fisher's 2016 Varsity team.

For the Junior Varsity meets, the observed race outcome sequence is given by

$$F'_{JV}(C', S') = c'_1, c'_2, c'_3, s'_1, c'_4, s'_2, c'_5, c'_6, s'_3, c'_7, s'_4, s'_5, s'_6, s'_7,$$

where $F'_{JV}(C', S')$ symbolizes the observed race outcome sequence for our two Junior Varsity teams in the Junior Varsity race, $C' = \{c'_1, c'_2, c'_3, c'_4, c'_5, c'_6, c'_7\}$ is a set representing the 7-element (7-runner) Carmel Junior Varsity team, and $S' = \{s'_1, s'_2, s'_3, s'_4, s'_5, s'_6, s'_7\}$ is a set representing the 7-element (7-runner) Fishers Junior Varsity team. This leads to rank sums

$$R'_{F'_{JV}}(C) = 36 \quad \text{and} \quad R'_{F'_{JV}}(S) = 69.$$

We see that the rank sum scores for the two teams are the same for the Varsity and Junior Varsity meets. Then, the upper-tail *WMW rank sum p-value* for the Junior Varsity sample difference, $P(R_{F'_{JV}} \geq 69 | n = 14)$, is also equal to 0.019. In a *WMW rank sum test* of the Junior Varsity data, then, we conclude that $C' \succ_{\alpha = 0.025} S'$. At the $\alpha = 0.025$ significance level or at any higher α -level, that is, we conclude that Carmel's 2016 Junior Varsity team was significantly higher in quality than Fisher's 2016 Junior Varsity team.

One might ponder the utility of ranking teams at the Junior Varsity level. Junior Varsity competition often provides a leading indicator of a program's future strength. Further, the quality of a Junior Varsity team can indicate program depth and robustness against injury. We might therefore expect a strong overall program to be strong at each level.

Using finishing times, we pool the two outcome sequences to create the aggregated sequence

$$FF'_{JV} = c_1, c_2, f_1, c_3, c_4, c_5, f_2, c_6, f_3, c_7, f_4, f_5, f_6, f_7, c'_1, c'_2, c'_3, f'_1, c'_4, f'_2, c'_5, c'_6, f'_3, c'_7, f'_4, f'_5, f'_6, f'_7$$

and obtain

$$R'_{JV}(C) = 170 \quad \text{and} \quad R'_{JV}(F) = 236.$$

As with its constituent outcome sequences, FF'_{JV} represents the true (observed) outcome sequence when the two meets (races) are pooled and runners are sequenced in ascending order of finishing time. Given the observed primitive race sequences, the pooled (aggregated) outcome sequence is simply the 14 Varsity elements (runners) in their primitive order followed by the 14 Junior Varsity elements. In other words, the first element in outcome sequence F_{JV} follows the final element in outcome sequence F_V . Each meet represents a distinct level of competition by which these two cross country programs are ranked against one another.

Given that $R'_{JV}(C) = 170$ and $R'_{JV}(F) = 236$, the *WMW rank sum p-value* for this sample difference, $P(R'_{JV} \geq 236 | n = 28)$, is equal to 0.069. In a *WMW rank sum test* of the aggregated data, then, we conclude that $(C \sim_{\alpha=0.025} S)$, where “ $\sim_{\alpha=0.025}$ ” reads “ranks indifferently to at the $\alpha = 0.025$ significance level.” In fact, we also find that $(C \sim_{\alpha=0.05} S)$. At the $\alpha = 0.05$ significance level or any smaller α -level, that is, we conclude from the aggregated data that Carmel’s 2016 Boys Cross Country program was not significantly different from that of Fishers. We obtain this result despite finding significant evidence that each of Carmel’s Boys Cross Country (primitive or constituent) teams (i.e., the Varsity and Junior Varsity teams) separately ranks significantly higher than its respective counterpart team at Fishers High School. In other words, we find an empirical example of an aggregation paradox when comparing these two Cross Country programs. This aggregation paradox is fairly extreme. Not only does the test of aggregated data lose significance at the $\alpha = 0.025$ level but also at the $\alpha = 0.05$ level. Further, the *p-value* for the aggregated data is more than the sum of the *p-values* for the two constituent data sets!

Of course, there are alternatives in sport to rank sum scoring. For example, team tennis matches typically employ a methodology that is consistent with the *sign test* methodology. For a dual team tennis match, each team assigns an intra-squad rank to each player or set of players, and each player for a given team is match paired to play against the player on the opposing squad who possesses the same intra-squad rank. Matches then take place and a team winner is assigned based on majority rule aggregation of individual match results. We can apply such a scoring rule to the case of team cross country. More specifically, we can apply this *sign test* type team tennis scoring to the application data from the two high school cross country meets considered.

$$M_V(C, S) = \{(c_1, s_1), (c_2, s_2), (c_3, s_3), \dots, (c_7, s_7)\}$$

$$M'_{JV}(C, S) = \{(c'_1, s'_1), (c'_2, s'_2), (c'_3, s'_3), \dots, (c'_7, s'_7)\}$$

In *sign test* type team tennis scoring, each Carmel runner defeated his assigned pair on the Fishers team (for each of the two competition levels). As such,

$$S_{M_V}(C) = 7 \quad S_{M_V}(S) = 0$$

$$S_{M'_{JV}}(C) = 7 \quad S_{M'_{JV}}(S) = 0$$

Then, we have that $P(S_{M_V}(\cdot) \geq 7 | \frac{n}{2} = 7) = P(S_{M'_{JV}}(\cdot) \geq 7 | \frac{n}{2} = 7) = 0.004075$. By aggregating the two data sets, we obtain

$$MM'_{JV}(C, S) = \{(c_1, s_1), (c_2, s_2), (c_3, s_3), \dots, (c_7, s_7), (c'_1, s'_1), (c'_2, s'_2), (c'_3, s'_3), \dots, (c'_7, s'_7)\}$$

and

$$S_{MM'_{VJV}}(C) = 14 \quad S_{MM'_{VJV}}(S) = 0$$

Then, we have that $P(S_{MM'_{VJV}}(\cdot) \geq 14 | \frac{n}{2} = 7) = P(S_{MM'_{VJV}}(\cdot) \geq 14 | \frac{n}{2} = 14) = 0.000091$. The *sign test p-value* diminishes in level of aggregation for the empirical application. Whereas the results of the *WMW test* are *inconsistent upon aggregation* in this empirical application, we again observe an instance of the general *consistency upon aggregation* quality of the *sign test*.

6 Application 3: Aggregation properties of the binomial test for equality of dependent proportions

McNemar's test for the equality of proportions of matched pairs considers the discordant pairs with responses (Yes, No) and (No, Yes). Let n be the total number of discordant pairs, and X be the number of discordant pairs with response (Yes, No). Then the test statistic is given by $T = \{|X - (n - X)| - 1\} / \sqrt{n}$, and the null hypothesis is rejected if T^2 is too large (see for example, Rosner ([53], p. 375)). Under the null hypothesis, conditioned on n X has a Binomial distribution with parameters n and $p = 1/2$. When n is small (say under 20) the exact Binomial distribution of X is used ([53], p. 377) to compute the p value and the null hypothesis of equality of proportions is rejected when X is too small or large. When n is large, chi-square approximation with 1 degree of freedom is used and the upper tail probability associated with observed T^2 is used as the p value. Thus, when number of discordant pairs in two data sets are equal, say n each, we can use Theorem 2 to establish strong form consistency upon aggregation of the test. When the number of discordant pairs vary, at least when the n_i are large, (11) can be used to establish consistency upon aggregation.

7 Application 4: Computation of Yule-Simpson Paradox for a case of rank sum scoring

Here, we consider a case of rank sum scoring in which two groups, each with two elements, are compared. The outcome of the comparison is a 4-element sequence (e.g., a, b, b, a). There are $4!/(2!2!) = 6$ such sequences. For each sequence, we replicate that same sequence and pool the two ordinally-identical data sets to create every possible pooled sequence of eight elements for the two groups that preserves the within-sample ordering for each original sequence. For example a, b, b, a can be replicated and pooled with its replicate, a', b', b', a' to create a', b', b', a' , a', a, b, b, a or, alternatively, $a, b, b, a', a, b', b', a'$. For each of the six sequences in this case, there are 70 possible poolings. This can be easily verified via a decision tree. Essentially, there are up to 5 "bins" in which to place elements of the replicate data set into the original data set, but the number of bins for a given element is constrained because one cannot re-order the original, constituent sequences. A decision tree shows that there are $[\sum_{k=1}^5 \sum_{i=1}^k i] = 70$ possible poolings for each original sequence. Across all 6 original sequences, there are $70 \cdot 6 = 420$ possible poolings. We find that 72 of these poolings (17.14 percent) result in an instance of *Yule-Simpson Paradox* of one of the following type: $R_c(A) \leq R_c(B)$ but $R_p(A) > R_p(B)$ or $R_c(A) \geq R_c(B)$ but $R_p(A) < R_p(B)$. That is, one group has at least as low a rank sum score as the other in the constituent data set, but this is not the case in the pooled data set. This result suggests that instances of *Yule-Simpson Paradox* are fairly regularly occurring for small sample cases. Qualitatively, the result is in line with previous small sample results on *Yule-Simpson Paradox* incidence for other tests (see [28]; [54]).

8 Conclusion

In this work, we have used social choice theory and non-parametric statistical theory to examine the *Yule-Simpson Aggregation Paradox* as it applies to non-parametric statistical tests. As discussed, the *Paradox* relates to a broader literature concerning the effectiveness of information aggregation mechanisms. Herein, we have shown that the *sign test* for matched pairs exhibits general *consistency upon aggregation*. This is the first non-parametric test for which this property has been demonstrated generally, while several non-parametric tests have been shown in the prior literature to *not* possess this property. This result illustrates that there exists a non-parametric statistical test for which *consistency upon aggregation* is *possible*. In the present work, moreover, we find paired matching of ordinal data (across groups) to be important toward the result of general *consistency*. We prove an additional result as to the nature of *sign test consistency*. Namely, the *sign test p-value* exhibits a greater proportional decrease in aggregation as the original (primitive data *sign test*) *p-value* itself becomes smaller. In aggregation, then, we expect a greater proportional *p-value* decline given significant *sign test* results for the constituent, primitive data (than for insignificant *sign test* results). This represents something of a *strong form consistency* result, whereby data aggregation reinforces significant *sign test* results to a greater (proportional) degree.

Incorporating a generalized form of data aggregation, we also generate preliminary evidence that the *sign test* possesses a generalized form of *consistency upon aggregation*, in which the primitive data sets have varying sample sizes. For tractability, *n-aggregation* has been incorporated as a standard approach in examining the *Yule-Simpson Paradox* for ordinal data. However, analysis of the generalized, n_i – aggregation form is an important consideration. Often, data of different sample sizes are pooled (e.g., in the case of many longitudinal and unbalanced panel data sets). While we believe that the result holds for the n_i – aggregation problem as well, our proof is incomplete; we can formally claim the consistency only for large samples (see Theorem/Conjecture 3). With the n – aggregation form, we proved the consistency for any n by directly comparing $P(X_n \geq k)$ and $P(X_N \geq rk)$ where $N = rn$. This approach does not seem to work when the n_i are unequal. We focused on the $p = 1/2$ case, as our interest was on the p – value, or on the property of the test under the typical null hypothesis. If one is interested in the power properties, we need similar results for $p \neq 1/2$. Using the normal approximation, we can conclude that the inequality stated in (11) holds for $c > p$ when n_1 and n_2 are large and sufficiently apart.

We further incorporate a generic application that demonstrates tests of the same (aggregated) data by the *sign test* and by the *WMW rank sum test*, respectively. In the example, the *WMW test* results exhibit *inconsistency upon aggregation*. *Sign test* results demonstrate (the predicted) *consistency upon aggregation*, however. In the application, we further verify that the *sign test* results exhibit “strong form consistency” in aggregation.

An empirical application obtained from Indiana Boys High School Cross Country running data demonstrates a real-world application exhibiting both *rank sum inconsistency* and *sign test consistency*. Future work might evaluate alternative matched pairs style non-parametric tests to determine if there is a family of such tests possessing *consistency upon aggregation*. Indeed, the property is of central importance. As all data sets can be viewed as a potential aggregation of primitive data sets (i.e., from the power set of the aggregated set of data), *consistency* tells us generally whether a given statistical test evaluates data in an unambiguous manner.

Author Contributions

Conceptualization: Haikady N. Nagaraja, Shane Sanders.

Formal analysis: Haikady N. Nagaraja, Shane Sanders.

Writing – original draft: Haikady N. Nagaraja, Shane Sanders.

References

1. Tideman T. N., & Plassmann F. (2013). Developing the aggregate empirical side of computational social choice. *Annals of mathematics and artificial intelligence*, 68(1-3), 31–64. <https://doi.org/10.1007/s10472-013-9360-z>
2. 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>
3. Yaari G., & Eisenmann S. (2011). The hot (invisible?) hand: can time sequence patterns of success/failure in sports be modeled as repeated random independent trials?. *PloS One*, 6(10), e24532. <https://doi.org/10.1371/journal.pone.0024532> PMID: 21998630
4. Gehrlein W. V., & Plassmann F. (2014). A comparison of theoretical and empirical evaluations of the Borda Compromise. *Social Choice and Welfare*, 43(3), 747–772. <https://doi.org/10.1007/s00355-014-0798-7>
5. Kock N. (2015). How Likely is Simpson's Paradox in Path Models?. *International Journal of e-Collaboration (IJEC)*, 11(1), 1–7.
6. Haunsperger D. B. (1992). Dictionaries of paradoxes for statistical tests on k samples. *Journal of the American Statistical Association*, 87(417), 149–155. <https://doi.org/10.2307/2290463>
7. Matzkin R. L. (1992). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica*, 239–270. <https://doi.org/10.2307/2951596>
8. Wilcoxon F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83. <https://doi.org/10.2307/3001968>
9. Pavlides M. G., & Perlman M. D. (2009). How likely is Simpson black's paradox? *The American Statistician*, 63(3), 226–233. <https://doi.org/10.1198/tast.2009.09007>
10. Haunsperger D. B. (1996). Paradoxes in nonparametric tests. *Canadian Journal of Statistics*, 24(1), 95–104. <https://doi.org/10.2307/3315692>
11. Kruskal W. H., & Wallis W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621. <https://doi.org/10.1080/01621459.1952.10483441>
12. Pearl J. (2014). Comment: understanding Simpson black's paradox. *The American Statistician*, 68(1), 8–13. <https://doi.org/10.1080/00031305.2014.876829>
13. Bhapkar V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. *Journal of the American Statistical Association*, 61(313), 228–235. <https://doi.org/10.1080/01621459.1966.10502021>
14. Hanson R., Oprea R., & Porter D. (2006). Information aggregation and manipulation in an experimental market. *Journal of Economic Behavior & Organization*, 60(4), 449–459. <https://doi.org/10.1016/j.jebo.2004.09.011>
15. Hammond T. H. (2007). Rank injustice?: How the scoring method for cross-country running competitions violates major social choice principles. *Public Choice*, 133(3-4), 359–375. <https://doi.org/10.1007/s11127-007-9193-6>
16. Datta S., & Satten G. A. (2005). Rank-sum tests for clustered data. *Journal of the American Statistical Association*, 100(471), 908–915. <https://doi.org/10.1198/016214504000001583>
17. Charig C. R., Webb D. R., Payne S. R., & Wickham J. E. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br Med J (Clin Res Ed)*, 292(6524), 879–882.
18. Gautier E., & Kitamura Y. (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica*, 81(2), 581–607. <https://doi.org/10.3982/ECTA8675>
19. Simpson E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13, 238–241. <https://doi.org/10.1111/j.2517-6161.1951.tb00088.x>
20. Bargagliotti A. E. (2009). Aggregation and decision making using ranked data. *Mathematical Social Sciences*, 58(3), 354–366. <https://doi.org/10.1016/j.mathsocsci.2009.07.006>
21. Bhattacharya D. (2015). Nonparametric welfare analysis for discrete choice. *Econometrica*, 83(2), 617–649. <https://doi.org/10.3982/ECTA12574>

22. De Neve J., & Thas O. (2015). A regression framework for rank tests based on the probabilistic index model. *Journal of the American Statistical Association*, 110(511), 1276–1283. <https://doi.org/10.1080/01621459.2015.1016226>
23. Gastwirth J. L. (1965). Percentile modifications of two sample rank tests. *Journal of the American Statistical Association*, 60(312), 1127–1141. <https://doi.org/10.1080/01621459.1965.10480856>
24. Wardrop R. L. (1995). Simpson's paradox and the hot hand in basketball. *The American Statistician*, 49(1), 24–28. <https://doi.org/10.2307/2684806>
25. Rosner B. (2011). *Fundamentals of Biostatistics*, Seventh Edition. Brooks/Cole, Boston.
26. Bickel P. J., Hammel E. A., & O'Connell J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175), 398–404. <https://doi.org/10.1126/science.187.4175.398> PMID: 17835295
27. Mann H. B., & Whitney D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1), 50–60. <https://doi.org/10.1214/aoms/1177730491>
28. McNemar Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157. <https://doi.org/10.1007/bf02295996> PMID: 20254758
29. Vartia Y. O. (1983). Efficient methods of measuring welfare change and compensated income in terms of ordinary demand functions. *Econometrica*, 51(1), 79–98. <https://doi.org/10.2307/1912249>
30. Kaufmann C., & Weber M. (2013). Sometimes less is more—The influence of information aggregation on investment decisions. *Journal of Economic Behavior & Organization*, 95, 20–33. <https://doi.org/10.1016/j.jebo.2013.08.005>
31. Gou J., & Zhang F. (2017) Experience Simpson's Paradox in the Classroom. *The American Statistician*, 71(1), 61–66. <https://doi.org/10.1080/00031305.2016.1200485>
32. Hausman J. A., & Newey W. K. (1995). Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica*, 63(6), 1445–1476.
33. Akritas M. G., Antoniou E. S., & Kuha J. (2006). Nonparametric analysis of factorial designs with random missingness: bivariate data. *Journal of the American Statistical Association*, 101(476), 1513–1526. <https://doi.org/10.1198/016214506000000537>
34. Evans T. A., Seaton S. E., & Manktelow B. N. (2013). Quantifying the potential bias when directly comparing standardised mortality ratios for in-unit neonatal mortality. *PLOS One*, 8(4), e61237. <https://doi.org/10.1371/journal.pone.0061237> PMID: 23577213
35. Saari D. G. (1999). Explaining all three-alternative voting outcomes. *Journal of Economic Theory*, 87(2), 313–355. <https://doi.org/10.1006/jeth.1999.2541>
36. Haunsperger D. B. (2003). Aggregated statistical rankings are arbitrary. *Social Choice and Welfare*, 20(2), 261–272. <https://doi.org/10.1007/s003550200179>
37. Stringer S., Wray N. R., Kahn R. S., & Derks E. M. (2011). Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. *PloS One*, 6(11), e27964. <https://doi.org/10.1371/journal.pone.0027964> PMID: 22140493
38. Selvitella A. (2017). The ubiquity of the Simpson's Paradox. *Journal of Statistical Distributions and Applications*, 4(1), 2. <https://doi.org/10.1186/s40488-017-0056-5>
39. Albers C. J. (2015). Dutch research funding, gender bias, and Simpson's paradox. *Proceedings of the National Academy of Sciences*, 112(50), E6828–E6829. <https://doi.org/10.1073/pnas.1518936112>
40. Miller J. B., & Sanjurjo A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica*, 86(6), 20192047. <https://doi.org/10.3982/ECTA14943>
41. Hao L., & Houser D. (2015). Adaptive procedures for the Wilcoxon-Mann-Whitney test: Seven decades of advances. *Communications in Statistics-Theory and Methods*, 44(9), 1939–1957. <https://doi.org/10.1080/03610926.2012.762394>
42. Bargagliotti A. E., & Greenwell R. N. (2011). Statistical significance of ranking paradoxes. *Communications in Statistics: Theory and Methods*, 40(5), 916–928. <https://doi.org/10.1080/03610920903506546>
43. Matzkin R. L. (1994). Restrictions of economic theory in nonparametric methods. *Handbook of Econometrics*, 4, 2523–2558. [https://doi.org/10.1016/S1573-4412\(05\)80011-X](https://doi.org/10.1016/S1573-4412(05)80011-X)
44. Boudreau J., Ehrlich J., Raza M. F., & Sanders S. (2018). The likelihood of social choice violations in rank sum scoring: algorithms and evidence from NCAA cross country running. *Public Choice*, 174(3–4), 219–238. <https://doi.org/10.1007/s11127-017-0494-0>
45. Mattei N. (2011, October). Empirical evaluation of voting rules with strictly ordered preference data. In *International Conference on Algorithmic Decision Theory* (pp. 165–177). Springer, Berlin, Heidelberg.
46. Hauert C., De Monte S., Hofbauer J., & Sigmund K. (2002). Volunteering as red queen mechanism for cooperation in public goods games. *Science*, 296(5570), 1129–1132. <https://doi.org/10.1126/science.1070582> PMID: 12004134

47. Saari D. G. (1995). A chaotic exploration of aggregation paradoxes. *SIAM Review*, 37(1), 37–52. <https://doi.org/10.1137/1037002>
48. Hogg R. V., Fisher D. M., & Randles R. H. (1975). A two-sample adaptive distribution-free test. *Journal of the American Statistical Association*, 70(351a), 656–661. <https://doi.org/10.1080/01621459.1975.10482489>
49. Koessler F., Noussair C., & Ziegelmeyer A. (2012). Information aggregation and belief elicitation in experimental parimutuel betting markets. *Journal of Economic Behavior & Organization*, 83(2), 195–208. <https://doi.org/10.1016/j.jebo.2012.02.017>
50. Briesch R. A., Chintagunta P. K., & Matzkin R. L. (2010). Nonparametric discrete choice models with unobserved heterogeneity. *Journal of Business & Economic Statistics*, 28(2), 291–307. <https://doi.org/10.1198/jbes.2009.07219>
51. Bennouri M., Gimpel H., & Robert J. (2011). Measuring the impact of information aggregation mechanisms: An experimental investigation. *Journal of Economic Behavior & Organization*, 78(3), 302–318. <https://doi.org/10.1016/j.jebo.2011.01.016>
52. Axelrod B. S., Kulick B. J., Plott C. R., & Roust K. A. (2009). The design of improved parimutuel-type information aggregation mechanisms: Inaccuracies and the long-shot bias as disequilibrium phenomena. *Journal of Economic Behavior & Organization*, 69(2), 170–181. <https://doi.org/10.1016/j.jebo.2007.09.010>
53. McFadden D., & Train K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5), 447–470. [https://doi.org/10.1002/1099-1255\(200009/10\)15:5%3C447::AID-JAE570%3E3.0.CO;2-1](https://doi.org/10.1002/1099-1255(200009/10)15:5%3C447::AID-JAE570%3E3.0.CO;2-1)
54. 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/10.2307/2684305>