

Consistency of Choice in Nonparametric Multiple Comparisons

Mark Fey[†] and Kevin A. Clarke[‡]
University of Rochester

July 12, 2011

Abstract

We are interested in the inconsistencies that can arise in the context of rank-based multiple comparisons. It is well known that these inconsistencies exist, but we prove that every possible distribution-free, rank-based multiple comparison procedure with certain reasonable properties is susceptible to these phenomena. The proof is based on a generalization of Arrow's Theorem, a fundamental result in social choice theory which states that when faced with three or more alternatives, it is impossible to rationally aggregate preference rankings subject to certain desirable properties. Applying this theorem to treatment rankings, we generalize a number of existing results in the literature and demonstrate that procedures that use rank sums cannot be improved. Finally, we show that best possible procedures are based on the Friedman rank statistic and the k -sample sign statistic, in that these statistics minimize the potential for paradoxical results.

Key words: Distribution-free; Arrow's Theorem; Cycling; Irrelevant alternatives.

[†]Associate Professor, Department of Political Science, University of Rochester, Rochester, NY 14627-0146. Tel.: (585) 275-5810; Fax: (585) 271-1616; Email: mark.fey@rochester.edu.

[‡]Corresponding author. Associate Professor, Department of Political Science, University of Rochester, Rochester, NY 14627-0146. Tel.: (585) 275-5217; Fax: (585) 271-1616; Email: kevin.clarke@rochester.edu.

1 Introduction

We are interested in distribution-free rank-based multiple comparisons, which is an idea that goes back to Kramer (1956) and which was first surveyed by Nemenyi (1963). In these procedures, objects (treatments) are ranked, and the sum of the ranks is determined for each object. Those with significantly large rank sum differences are declared to be different (McDonald & Thompson 1967).

Distribution-free rank sum multiple comparisons have been developed for both one- and two-way layouts. Ranks in the one-way layout can be applied either in a pairwise fashion (only the observations from the i th and j th treatments are ranked, and the rankings must be redone when the comparison moves to the next pair) or jointly (observations from all k treatments are ranked from smallest to largest). Similarly, ranks in the two-way layout can be applied either in a pairwise fashion or within the blocks. Some well-known examples of the kinds of procedures we consider are listed in Table 7 in Appendix 1.

Both methods of assigning ranks—pairwise and jointly—have well-known drawbacks (Lehmann 2006, Miller 1981). When observations are ranked in a pairwise fashion, an inconsistency known as cycling can arise where treatment j is declared superior to treatment i and treatment k superior to j , but without k being superior to i . When observations are ranked jointly or within blocks, the significance of a comparison between a pair of treatments depends upon the observations from treatments not involved in the comparison. Thus, results may change depending upon the number of treatments being considered. This type of inconsistency is known as the problem of irrelevant alternatives.

The point of this paper is to demonstrate that these inconsistencies can be explained as a generalization of a fundamental result in social choice theory due to Arrow (1963). In doing so, we generalize earlier findings by Haunsperger (1992), Haunsperger (1996), and Taplin (1997), who focus

mainly on characterizing the kinds of inconsistencies that can occur for a particular choice of statistical test. Specifically, our result shows that no distribution-free rank sum procedure (hereafter, we use the shorthand “test”) with certain reasonable properties exists that always avoids both types of inconsistency. Thus, the choice of a statistical test in these situations involves a tradeoff between the possibility of dependence on irrelevant alternatives and the possibility of cycling. We also show that the tests that minimize these inconsistencies are based on either the Friedman rank statistic or the k -sample sign statistic.

2 Examples

Examples of the paradoxes that can arise through use of rank-based procedures have been identified by several authors in the statistical literature; our review is correspondingly brief.

As an example of the inconsistency that can result when treatments are ranked jointly or within blocks, consider the data on the effectiveness of hypnosis in Table 1. The emotions of fear, happiness, depression, and calmness were requested (in random order) from each of eight subjects during hypnosis (Lehmann 2006, 264).

Table 1: Effectiveness of hypnosis (Lehmann 2006, 264). *The data are skin potential (adjusted for initial level) measured in millivolts.*

Subject	1	2	3	4	5	6	7	8
Fear	23.1	57.6	10.5	23.6	11.9	54.6	21.0	20.3
Happiness	22.7	53.2	9.7	19.6	13.8	47.1	13.6	23.6
Depression	22.5	53.7	10.8	21.1	13.7	39.2	13.7	16.3
Calmness	22.6	53.1	8.3	21.6	13.3	37.0	14.8	14.8

Given that the data are blocked, a Friedman-type multiple comparison is appropriate (Nemenyi 1963). The results of the multiple comparison proce-

dures are in Table 2. When all four treatments are compared (columns 2 and 3), none of the contrasts are statistically significant.

Table 2: Multiple Comparisons for the data in Table 1. *When four treatments are compared (left side of the table), none of the contrasts are statistically significant. When one treatment is dropped (the right side of the table), the difference between treatments 1 and 3 is now statistically significant.*

Treatments	1,2,3,4 ($r_{0.05} = 13.62$)		1,2,3 ($r_{0.05} = 9.58$)	
Comparison	Obs. Diff.	Sig.	Obs. Diff.	Sig.
1 against 2	7	No	5	No
1 against 3	8	No	10	Yes
1 against 4	13	No	-	-
2 against 3	1	No	5	No
2 against 4	6	No	-	-
3 against 4	5	No	-	-

When the fourth treatment (calmness) is dropped (columns 4 and 5), however, the difference between treatment 1 (fear) and treatment 3 (depression) is now statistically significant. That is, the statistical finding that the fear outcome differs from the depression outcome depends on whether or not the analyst includes the calmness treatment in his or her analysis.

It is important to understand that it is not just the outcome of the test that may be inconsistent; the ranks sums, themselves, can demonstrate this same kind of inconsistency. Consider the data in Table 3. The ranks are assigned within the 7 blocks, and in the first four columns, the ranks are synonymous with the data. The treatments are ranked $M_3 \prec M_2 \prec M_1 \prec M_4$. In columns 5 through 7, treatment 4 is dropped, and the data are ranked again. This time, the treatments are ranked $M_1 \prec M_2 \prec M_3$. Notice that in this instance the entire ordering of treatments 1, 2, and 3 are reversed by dropping the treatment with the highest rank sum, M_4 , from the analysis.

These kinds of inconsistencies happen not only with Friedman-type ranks, but also with joint ranks (Kruskal-Wallis). Consider the data in Table 4 taken from Haunsperger (1992). The $9 \times 4 = 36$ observations are jointly ranked

Table 3: Friedman Ranks. *Columns 2-5 contain the data, which are synonymous with their with-in block ranks. M_4 is dropped in columns 6-8, and the data are reranked. The rank sums in the last row demonstrate the reversal: $M_3 \prec M_2 \prec M_1 \prec M_4$ versus $M_1 \prec M_2 \prec M_3$.*

Subject	M_1	M_2	M_3	M_4	M_1	M_2	M_3
1	1	2	3	4	1	2	3
2	4	1	2	3	3	1	2
3	3	4	1	2	2	3	1
4	1	2	3	4	1	2	3
5	4	1	2	3	3	1	2
6	3	4	1	2	2	3	1
7	1	2	3	4	1	2	3
Rank Sum	17	16	15	22	13	14	15

Table 4: Kruskal-Wallis data (Haunsperger 1992, 151).

Subject	C_1	C_2	C_3	C_4
1	4.20	4.38	4.12	4.04
2	4.32	4.23	4.10	4.42
3	4.07	4.14	4.16	4.44
4	4.11	4.08	4.40	4.02
5	4.22	4.18	4.19	4.00
6	4.46	4.30	4.24	4.41
7	4.29	4.25	4.21	4.45
8	4.33	4.13	4.34	4.06
9	4.15	4.43	4.37	4.03

in the first four columns of Table 5. Based on the sum of these joint ranks, the four treatments have the following ordering: $C_3 \prec C_2 \prec C_1 \prec C_4$. When treatment 4 is dropped, however, the ordering based on joint ranks reverses for the remaining three treatments: $C_1 \prec C_2 \prec C_3$.

One way to avoid this phenomenon is to compare treatments two at a time, rather than jointly. That is, instead of ranking all treatments and using these rankings to compare treatments, an alternative approach is to simply

Table 5: . The Kruskal-Wallis data in Table 4 are jointly ranked in columns 1-4. C_4 is dropped, and the data are reranked in columns 5-7. The rank sums in the last row demonstrate the reversal: $C_3 \prec C_2 \prec C_1 \prec C_4$ versus $C_1 \prec C_2 \prec C_3$.

Subject	C_1	C_2	C_3	C_4	C_1	C_2	C_3
1	17	29	10	4	12	24	5
2	25	20	8	32	20	15	3
3	6	12	14	34	1	7	9
4	9	7	30	2	4	2	25
5	19	15	16	1	14	10	11
6	36	24	21	31	27	19	16
7	23	22	18	35	18	17	13
8	26	11	27	5	21	6	22
9	13	33	28	3	8	26	23
Rank Sum	174	173	172	147	125	126	127

compare each pair of treatments separately. Unfortunately, this method, known as pairwise ranking, gives rise to a different kind of inconsistency. Consider the data in the first three columns of Table 6 taken from Lehmann (2006, 245). When the rank sums are computed, $X \prec Y$, $Y \prec Z$, but $X \succ Z$. Thus, we conclude that X is significantly better than Y and Y is significantly better than Z , but also that Z is significantly better than X . Thus, the pairwise ranking approach fails to generate a sensible ordering of the treatments.

3 Desirable Properties of Tests

Our approach is to first define some intuitively appealing properties that any reasonable test should possess and then show that, in fact, there is no distribution-free test that possesses all of these properties.

Suppose that we have n units of observation that each have been exposed to k treatments. This setup comprises the one-way layout with k general

Table 6: Pairwise Ranks. *The data in columns 1-3 (Lehmann 2006, 245) are ranked in pairs in columns 4-9. The rank sums in the last row demonstrate the inconsistency: $X \prec Y$, $Y \prec Z$, and $X \succ Z$.*

Subject	Data			Paired Rankings					
	X	Y	Z	X	Y	Y	Z	X	Z
1	2	4	1	1	3	2	1	2	1
2	3	5	7	2	4	3	5	3	4
3	9	6	8	6	5	4	6	6	5
Rank Sum				9	12	9	12	11	10

alternatives as well as the two-way layout with k general alternatives known as a randomized complete block design. See Hollander & Wolfe (1999) for a complete overview.

Formally, the data consist of nk observations, where X_{im} is the observation of the i th unit under the m th treatment, for $i = 1, \dots, n$ and $m = 1, \dots, k$. Thus, the observations are the rows and the treatments are the columns. For simplicity, we assume that the value of each observation X_{im} is distinct. (This is equivalent to assuming that the joint distribution generating the distribution is non-atomic.) A statistical test operates on the data X and generates a test statistic which is compared to a critical value. Of course, any number of hypotheses may be the subject of such a test. In this paper, we are interested in generating an ordering of the treatments according to their effect on the data (Lehmann 2006). That is, we consider a test $\phi(X)$ which, for each treatment m and l , chooses whether treatment m is significantly better than treatment l , treatment l is significantly better than treatment m , or neither treatment is significantly better than the other. We represent the first conclusion by the notation $t_m \succ t_l$, the second conclusion by $t_l \succ t_m$, and the last conclusion by $t_m \sim t_l$. To emphasize that these conclusion depend on the data X , we sometimes write \succ_X for the ordering relation of the treatments. Naturally, the outcome of the test also could depend on the number of observations. The statistical literature describes

a number of such tests and in this section we consider *all* tests that satisfy certain desirable properties.

The class of tests that we consider are those that are **distribution-free**. This term (as well as the term “nonparametric”) has been used in a number of different ways in the statistical literature. We use the term “distribution-free” to refer to statistical tests that have the same distribution of the test statistic regardless of the underlying distribution. For a precise definition, see Appendix 2. In particular, the level of a distribution-free test never exceeds the stated significance level regardless of the underlying distribution.

It can be shown that, in our setting, distribution-free tests are exactly those tests that are invariant to continuous, positive transformations of the data and that these tests are exactly those based on rank statistics (Birnbaum & Rubin 1954, Bell 1960, Bell 1964, Bell & Smith 1972). Intuitively, invariance to continuous, positive transformation of the data is equivalent to “distribution-freeness” because any two continuous distributions can be linked by such a transformation that preserves the property that one distribution is stochastically larger than the other, and this invariance is equivalent to rank statistic tests because the set of ranks is a maximal invariant (Lehmann 1986, 315).

We now specify the properties that any reasonable distribution-free test should possess. The first property that is natural to consider is that the test not depend on how the units of observation are numbered. We state this symmetry property in terms of invariance to permutation.

Symmetry For every data matrix X , the test is invariant to any permutation of the rows of X .

The second property that we require is that the comparisons between treatments not be inconsistent. That is, if the test indicates that treatment t_m is significantly better than treatment t_l and treatment t_l is significantly better than treatment t_r , then the test should yield that treatment t_m is

significantly better than treatment t_r . Formally, we require that the test generate a strict partial order of the treatments.

Ordering of Treatments For every data matrix X , the ordering relation of the treatments, \succ_X is a strict partial order. That is, \succ is

1. irreflexive, so that for all m , $t_m \not\succ t_m$,
2. asymmetric, so that $t_m \succ t_l$ implies $t_l \not\succ t_m$, and
3. transitive, so that $t_m \succ t_l$ and $t_l \succ t_r$ implies $t_m \succ t_r$.

It is important to emphasize that the ordering relation of the treatments is only a partial order. Specifically, the transitivity condition only applies to significantly different treatments; it is certainly possible that the test cannot determine a significant difference between either treatments t_m and t_l or treatments t_l and t_r , but it can determine a significant difference between t_m and t_r .

The next property rules out uninteresting or pathological tests by requiring that the test have power.

Nonzero Power For each pair of treatments t_m and t_l , there exists an integer N such that for all data sizes $n \geq N$, if at least $n - 1$ units of observation have $X_{im} > X_{il}$, then the test yields $t_m \succ t_l$.

This property says that for sufficiently large data sets, if all or all but one of the units of observation are better under treatment t_m than under treatment t_l , the test should choose treatment t_m over t_l . In particular, this property rules out the trivial test which never rejects the null of equality.

The final property is that results of the test should not change if we remove one or more treatments from consideration. Let $M = \{1, \dots, k\}$ be the set of treatments and for any non-empty $R \subset M$, let $X(R)$ be the data set consisting of the data for the treatments in the set R . That is, $X(R) = \{X_{im}\}_{\substack{i=1, \dots, n \\ m \in R}}$.

Independence of Treatments For every data matrix X and every non-empty $R \subset M$ with $m, l \in R$, $t_m \succ_X t_l$ implies $t_m \succ_{X(R)} t_l$.

This property is essentially an independence condition that requires the outcome of the test be independent of other, unrelated, treatments. In other words, the comparison of treatments t_m and t_l should not depend on the presence or absence of some third treatment in the analysis.

4 Results

Having established several reasonable properties for statistical test, we now address the question of whether there is a test that possesses all of these desirable properties. The answer to this question is our main result.

Theorem 1 *There is no symmetric, distribution-free test with nonzero power that orders treatments and that satisfies independence of treatments.*

Proof: Suppose for a proof by contradiction that there exists a symmetric, distribution-free test with nonzero power that orders treatments and that satisfies independence of treatments. As the test satisfies the Nonzero Power property, fix the data size n to be such that for every pair of treatments m and l , if at least $n - 1$ units of observation have $X_{im} > X_{il}$, then the test yields $t_m \succ t_l$. We use a version of Arrow's Theorem known as the Positional Dictatorship Theorem (Gevers 1979, Roberts 1980) to reach a contradiction. The details of this theorem are presented in Appendix 3.

We begin by defining a binary relation R on X by letting $X_{im} R X_{jl}$ if and only if $X_{im} \geq X_{jl}$. This relation is clearly an ordering and it is a linear order because of our assumption that all data points are distinct. This ordering is invariant to any increasing transformation of the data, so it is a sufficient summary of the data for a distribution-free test. Likewise, we use the outcome of our test to define a binary relation on the set of

treatments. Formally, we let mP^*l if and only if $t_m \succ_X t_l$. By the Ordering of Treatments property, this relation is a partial order. Therefore, the test defines a generalized quasi-transitive social preference function (hereafter, GQTSPF), as described in Appendix 3.

Next, we verify that each of the conditions described in Appendix 3 are satisfied. First, the Symmetry property of the test immediately implies that the corresponding GQTSPF satisfies Anonymity. Second, from our choice of n the Nonzero Power property of the test insures that P^* satisfies the Pareto condition. Third, the Independence of Treatments property of the test implies that P^* satisfies Independence of Irrelevant Alternatives (IIA).

Therefore, by the version of the Positional Dictatorship Theorem given in Appendix 3, the test must correspond to a positional oligarchy, as defined in Appendix 3. We conclude the proof by showing no such test can satisfy the Nonzero Power condition. Let d be position in the positional oligarchy. As a positional oligarchy is always non-empty, such a position exists. Define the values of X for treatment t_m by $X_{im} = 2i$ for $i \neq d$ and $X_{dk} = 2d - 1$ and for treatment t_l by $X_{il} = 2i - 1$ for $i \neq d$ and $X_{dl} = 2d$. All but one observation have $X_{im} > X_{il}$ so by the Nonzero Power condition, we must have $t_m \succ t_l$. But since d is a position in the positional oligarchy, this implies that $(d, m)R(d, l)$ which is equivalent to $X_{dm} \geq X_{dl}$. This last statement contradicts the definition of X_{dm} and X_{dl} , which establishes that there is no test that satisfies all of the conditions of the theorem. ■

This result is important because it establishes that there is no possible test that always orders choices in a consistent way. It is worth emphasizing that this theorem applies to both one-way and two-way layouts and to tests that rank treatments by observation as well as tests that use overall rankings. This theorem means that any given test must violate at least one of these properties. Indeed, it is easy to give examples of tests that satisfy four of the five properties. In a one-way layout, Dunn's (1964) procedure satisfies all

of the properties except for Independence of Treatments, while the Dwass-Steel-Critchlow-Fligner procedure (Hollander & Wolfe 1999, 240) satisfies all of the properties except for Ordering of Treatments. In a two-way layout, the Nemenyi-McDonald-Thompson procedure (Hollander & Wolfe 1999, 295) satisfies all of the properties except for Independence of Treatments, while Steel's (1959*b*) satisfies all of the properties except for Ordering of Treatments.

Certainly, this result raises several related questions. One obvious question is that if there is no test that satisfies the two desirable properties of ordering treatments and independence of treatments, can we pick a test that will satisfy one of these properties and also in some sense do well on the other? The answer to this question is positive.

First, suppose that we restrict ourselves to tests that always order choices. In fact, we restrict ourselves further to “positional methods” that assign point values to each rank and sum up the points. The Friedman multiple comparison procedure is one example. As proved by Saari (1990) and applied to statistical tests by Haunsperger (1992), among positional voting rules, the Borda rule is the one that has the fewest violations of the Independence of Treatments property. In our context, this fact implies the following result.

Theorem 2 *Among positional method tests (which are all symmetric, distribution-free tests with nonzero power that order treatments), the Friedman statistic multiple comparison test has the fewest violations of independence of treatments.*

Alternatively, we can consider statistical tests which always satisfy independence of treatments and look for a test that has the fewest violations of ordering of treatments. As proved by Maskin (1995) and Campbell & Kelly (2000), among voting rules, the majority rule is the one that is transitive on the largest domain of preference orders. In our context, this fact implies the following result.

Theorem 3 *Among all symmetric, distribution-free tests with nonzero power that satisfy independence of treatments, the k -sample sign statistic multiple comparison test has the fewest violations of ordering of treatments.*

Thus, although no test satisfies all of the desirable properties that we have identified, the multiple comparison tests based on the Friedman rank statistic and the k -sample sign statistic offer the best tradeoffs between these properties.

But what can be said about the comparison between these two tests? Is one better than the other? The answer to this question is given in the next theorem.

Theorem 4 *Suppose the Friedman rank statistic satisfies independence of treatments for some data matrix X . Then the k -sample sign statistic satisfies the ordering of treatments property on X .*

Proof: Suppose the Friedman rank statistic satisfies independence of treatments for some data matrix X and chooses treatment t_m over treatment t_l . Because it satisfies independence of treatments on X , the test chooses t_m over t_l no matter how many other treatments are included with treatments t_m and t_l . In particular, the test chooses t_m over t_l when these are the only two treatments included. But the Friedman rank statistic with two treatments is identical to the k -sample sign statistic and therefore the k -sample sign statistic must choose t_m over t_l . This same point applies to all of the pairwise comparisons and therefore it must be the case that the Friedman rank statistic and the k -sample sign statistic give the same answer for every pair of treatments. It follows that since the Friedman rank statistic satisfies the ordering of treatments property, for the data matrix X the k -sample sign statistic also satisfies this property. ■

Thus, whenever the Friedman rank statistic is “well-behaved” in the sense that it satisfies all of the properties we have described, the k -sample sign

statistic is also be “well-behaved.” However, the converse of this theorem is false. That is, it is possible that the k -sample sign statistic satisfies the ordering of treatments property on a data matrix X , but the Friedman rank statistic does not satisfy the independence of treatments property for the data X . As a general matter, then, the k -sample sign statistic is strictly better than the Friedman rank statistic when minimizing the number of inconsistencies. (This result says nothing about the relative power of the two procedures.) Put another way, for *any* data matrix X , either the k -sample sign statistic will be “well-behaved” and agree with the Friedman test or the the Friedman rank statistic will violate independence of treatments.

5 Conclusion

We have examined the inconsistencies that can arise in the context of ranked-based multiple comparisons. We have shown by way of examples that these inconsistencies can occur. Moving beyond these examples, we have shown that every possible distribution-free rank-based multiple comparison with certain reasonable properties is susceptible to these phenomena. In doing so, we generalized a number of existing results in the literature. Finally, we argue that the best possible tests are the multiple comparison test based on the Friedman rank statistic and the k -sample sign statistic, in that these tests minimize the potential for paradoxical results.

Appendix 1: Rank-Based MCPs and Papers

Table 7: This categorization is based on Miller (1981), Hochberg & Tamhane (1987), and Hollander & Wolfe (1999).

	One-Way Layout	
	Paired	Joint
General Configuration	k -sample ranks Steel (1959 <i>a</i>), Dwass (1960), Steel (1960), Steel (1961), Fligner (1884), Critchlow & Fligner (1991)	Kruskal-Wallis ranks Nemenyi (1963), Dunn (1964), McDonald & Thompson (1967)
Control v. Treatment	Steel (1959 <i>b</i>)	Nemenyi (1963), Dunn (1964), Damico & Wolfe (1987)
Ordered Treatments	k -sample ranks Hayter & Stone (1991)	
	Two-Way Layout	
	Paired	Joint
General Configuration	k -sample ranks Nemenyi (1963), McDonald & Thompson (1967)	Friedman ranks Nemenyi (1963), McDonald & Thompson (1967), Odeh (1977), Skillings & Mack (1981)
Control v. Treatment	k -sample sign Steel (1959 <i>b</i>), Nemenyi (1963), Rhyne & Steel (1965)	Nemenyi (1963), Wilcoxon & Wilcox (1964), Miller (1981)

Appendix 2: Distribution-free and Invariant Tests

We begin with the definitions of distribution-free and invariance that we use. Both definitions involve the use of a group \mathcal{G} of transformations of the sample space. That is, each transformation $g \in \mathcal{G}$ maps the sample space onto itself. Let S be the class of strictly increasing continuous distributions on \mathbb{R} . Formally, the group \mathcal{G} is said to be an invariant class of transformations if

1. \mathcal{G} is a group, so that
 - (a) if $g_1, g_2 \in \mathcal{G}$, then the product transformation $g_1 \circ g_2 \in \mathcal{G}$, and
 - (b) if $g \in \mathcal{G}$, then the inverse transformation $g^{-1} \in \mathcal{G}$, and
2. for all $g \in \mathcal{G}$, if $F \in S$, then $F_g(x) = F[g^{-1}(x)] \in S$.

In the k -treatment case we consider here, the appropriate group of invariant transformations is the group of strictly increasing continuous transformations of \mathbb{R} onto itself.

The definition of a distribution-free statistic is stated in relation to a class of distributions and a group of transformations.

Definition 1 *A statistic T is strongly distribution-free with respect to the class of distribution S and a group of transformations \mathcal{G} if for all real t , all $F \in S$, and all $g \in \mathcal{G}$*

$$P_F[T(x) \leq t] = P_G[T(x) \leq t],$$

where $G(x) = F_g(x) = F[g^{-1}(x)]$.

A statistic is invariant to a transformation if its value does not change under the transformation. It is almost invariant if its value almost never changes.

Definition 2 *A statistic T is invariant with respect to a group of transformations \mathcal{G} if for all $g \in \mathcal{G}$*

$$T(x) = T(gx).$$

A statistic T is almost invariant with respect to the class of distribution S and a group of transformations \mathcal{G} if all $F \in S$ and for all $g \in \mathcal{G}$

$$P_F[T(x) \neq T(gx)] = 0.$$

As we only consider non-sequential tests, it is known that the classes of strongly distribution-free, almost invariant, and rank statistic tests coincide.

Theorem 5 (Theorem 5.2, Bell & Smith (1972)) *If T is non-sequential, then the following are equivalent:*

1. *T is strongly distribution-free,*
2. *T is almost invariant,*
3. *T is equivalent to a rank statistic.*

Note that any test satisfying the Symmetry Property must be non-sequential.

Appendix 3: The Positional Dictatorship Theorem

We begin with some definitions. Consider a binary relation R on a finite set Z . The relation R is *complete* if xRy or yRx for all $x, y \in Z, x \neq y$, *reflexive* if xRx for all $x \in Z$, and *transitive* if xRy and yRz imply xRz for all $x, y, z \in Z$. The relation R is called an *ordering* if it is complete, reflexive, and transitive, and a *linear order* if it is an ordering such that xRy and yRx imply $x = y$ for all $x, y \in Z$. Next, for a given binary relation R , let P denote its asymmetric part. That is, xPy if and only if xRy and not yRx . We say R is *quasi-transitive* if P is transitive and R is a *quasi-transitive order* if it is complete, reflexive, and quasi-transitive. Note that R is a quasi-transitive order if and only if P is irreflexive, asymmetric, and transitive.

Suppose there is a set N of individuals and a set K of alternatives, with $|K| \geq 3$. In the standard version of Arrow's Theorem, each individual has a linear order over the set of alternatives and we are interested in aggregating these $|N|$ incomparable orderings into a social ordering over the alternatives. Our problem, however, permits interpersonal comparisons of ranks and only requires quasi-transitive social orderings, so we must adjust the aggregation problem accordingly. Let R be a linear order on the set $N \times K$ and let \mathcal{R} be the set of all such orderings. Then a *generalized quasi-transitive social preference function (GQTSPF)* is a function that assigns a quasi-transitive order R^* on Z to each $R \in \mathcal{R}$. We denote such a GQTSPF by f .

We now define the conditions we will impose on a GQTSPF.

Pareto For $k, l \in K$, if $(i, k)P(i, l)$ for all $i \in N$, then kP^*l .

Independence of Irrelevant Alternatives (IIA) If $R, R' \in \mathcal{R}$ are two orderings such that for $x, y \in Z$, $(i, x)R(j, y)$ iff $(i, x)R'(j, y)$ and $(j, y)R(i, x)$ iff $(j, y)R'(i, x)$ for all $i, j \in N$, then $xf(R)y$ iff $xf(R')y$ and $yf(R)x$ iff $yf(R')x$.

Anonymity If $R, R' \in \mathcal{R}$ are two orderings such that for all $x, y \in Z$ and all $i, j \in N$, $(i, x)R(j, y)$ iff $(\pi(i), x)R'(\pi(j), y)$ for some permutation π on N , then $f(R) = f(R')$.

Finally, we must define a particular type of quasi-transitive social ordering R^* on Z . For an ordering $R \in \mathcal{R}$ and $x \in Z$, let R_x be the ordering on K defined by $iR_x j$ iff $(i, x)R(j, x)$. Using this ordering, let $i(d, R_x)$ be the individual that is ranked d th from the bottom. We say $R^* = f(R)$ is a *positional oligarchy* if there exists some non-empty subset $M \subseteq K$ such that for all $R \in \mathcal{R}$ and for all $x, y \in Z$, xP^*y implies $(i(d, R_x), x)R(i(d, R_y), y)$ for all $d \in M$ and $(i(d, R_x), x)P(i(d, R_y), y)$ for all $d \in M$ implies xP^*y .

These definitions allow us to state the version of the Positional Dictatorship Theorem due to (Roberts 1980) that we make use of.

Theorem: If a GQTSPF satisfies IIA, Pareto, and Anonymity, then it is a positional oligarchy.

References

- Arrow, Kenneth J. 1963. *Social Choice and Individual Values*. 2nd ed. New York: Wiley.
- Bell, Charles B. 1960. "On the Structure of Distribution-Free Statistics." *The Annals of Mathematical Statistics* 31(3):703–709.
- Bell, Charles B. 1964. "A Characterization of Multisample Distribution-Free Statistics." *The Annals of Mathematical Statistics* 35(2):735–738.
- Bell, Charles B. & Paul J. Smith. 1972. "Completeness theorems for characterizing distribution-free statistics." *Annals of the Institute of Statistical Mathematics* 24(1):435–453.
- Birnbaum, Zygmund W. & Herman Rubin. 1954. "On Distribution-Free Statistics." *The Annals of Mathematical Statistics* 25(3):593–598.
- Campbell, Donald E. & Jerry S. Kelly. 2000. "A simple characterization of majority rule." *Economic Theory* 15(3):689–700.
- Critchlow, Douglas E. & Michael A. Flinger. 1991. "On distribution-free multiple comparisons in the one-way analysis of variance." *Communications in Statistics: Theory and Methods* 20(1):127–139.
- Damico, J.A. & D.A. Wolfe. 1987. "Extended tables of the exact distribution of a rank statistic for all treatments multiple comparisons in a one-way layout." *Communications in Statistics: Theory and Methods* 16(7):2343–2360.
- Dunn, Olive Jean. 1964. "Multiple comparisons using rank sums." *Technometrics* 6(3):241–252.
- Dwass, Meyer. 1960. Some k-sample rank-order tests. In *Contributions to Probability and Statistics*, ed. I. Olkin, S.G. Ghurye, H. Hoeffding, W.G.

- Madow & H.B. Mann. Stanford, CA: Stanford University Press pp. 198–202.
- Fligner, Michael A. 1884. “A Note on Two-Sided Distribution-Free Treatment Versus Control Multiple Comparisons.” *Journal of the American Statistical Association* 79(385):208–211.
- Gevers, Louis. 1979. “On interpersonal comparability and social welfare orderings.” *Econometrica* 47(1):75–89.
- Haunsperger, Deanna B. 1992. “Dictionaries of Paradoxes for Statistical Tests on k samples.” *Journal of the American Statistical Association* 87(417):149–155.
- Haunsperger, Deanna B. 1996. “Paradoxes in nonparametric tests.” *The Canadian Journal of Statistics* 24(1):95–104.
- Hayter, Anthony J. & Glenn Stone. 1991. “Distribution Free Multiple Comparisons for Monotonically Ordered Treatment Effects.” *Australian Journal of Statistics* 33(3):335–346.
- Hochberg, Yosef & Ajit Tamhane. 1987. *Multiple Comparison Procedures*. New York: John Wiley and Sons.
- Hollander, Myles & Douglas A. Wolfe. 1999. *Nonparametric Statistical Methods*. 2 ed. New York: John Wiley and Sons.
- Kramer, Amihud. 1956. “A quick, rank test for significance of differences in multiple comparisons.” *Food Technology* 10(8):391–392.
- Lehmann, E. L. 1986. *Testing Statistical Hypotheses*. 2 ed. New York: John Wiley.
- Lehmann, Erich L. 2006. *Nonparametrics: Statistical Methods Based on Ranks*. New York: Springer Science.

- Maskin, Eric. 1995. Majority Rule, Social Welfare Functions, and Game Forms. In *Choice, Welfare, and Development: A Festschrift in Honour of Amartya K. Sen*. Oxford University Press, USA.
- McDonald, B.J. & W.A. Thompson. 1967. "Rank sum multiple comparisons in one- and two-way classifications." *Biometrika* 54(3-4):487–497.
- Miller, Rupert G. 1981. *Simultaneous Statistical Inference*. New York: Springer-Verlag.
- Nemenyi, Peter. 1963. Distribution-free multiple comparisons PhD thesis Princeton University Princeton, NJ: .
- Odeh, R.E. 1977. "Extended tables of the distributions of rank statistics for treatment versus control in randomized block designs." *Communications in Statistics: Simulation and Computation* 7(2):101–113.
- Rhyme, A. L. & R. G. D. Steel. 1965. "Tables for a treatments versus control multiple comparisons sign test." *Technometrics* 7:293–306.
- Roberts, Kevin W. S. 1980. "Possibility theorems with interpersonally comparable welfare levels." *Review of Economic Studies* 47(2):409–420.
- Saari, Donald. 1990. "The Borda dictionary." *Social Choice and Welfare* 7(4):279–317.
- Skillings, J.H. & G.A. Mack. 1981. "On the use of a Friedman-type statistic in balanced and unbalanced block designs." *Technometrics* 23:171–177.
- Steel, Robert G.D. 1959a. "A Multiple Comparison Rank Sum Test: Treatments Versus Control." *Biometrics* 15(4):560–572.
- Steel, Robert G.D. 1959b. "A Multiple Comparison Sign Test: Treatments Versus Control." *Journal of the American Statistical Association* 54(288):767–775.

- Steel, Robert G.D. 1960. "A Rank Sum Test for Comparing All Pairs of Treatments." *Technometrics* 2(2):197–207.
- Steel, Robert G.D. 1961. "Some Rank Sum Multiple Comparisons Tests." *Biometrics* 17(4):539–552.
- Taplin, Ross H. 1997. "The Statistical Analysis of Preference Data." *Applied Statistics* 46(4):493–512.
- Wilcoxon, Frank & Roberta A. Wilcox. 1964. *Some Rapid Approximate Statistical Procedures*. 2 ed. Pearl River, NJ: Lederle Laboratories.