The Annals of Statistics
2019, Vol. 47, No. 2, 758–782
https://doi.org/10.1214/17-AOS1634

RANK VERIFICATION FOR EXPONENTIAL FAMILIES

BY KENNETH HUNG AND WILLIAM FITHIAN[1]

University of California, Berkeley

Many statistical experiments involve comparing multiple population groups. For example, a public opinion poll may ask which of several political candidates commands the most support; a social scientific survey may report the most common of several responses to a question; or, a clinical trial may compare binary patient outcomes under several treatment conditions to determine the most effective treatment. Having observed the "winner" (largest observed response) in a noisy experiment, it is natural to ask whether that candidate, survey response or treatment is actually the "best" (stochastically largest response). This article concerns the problem of *rank verification*—post hoc significance tests of whether the orderings discovered in the data reflect the population ranks. For exponential family models, we show under mild conditions that an unadjusted two-tailed pairwise test comparing the first two-order statistics (i.e., comparing the "winner" to the "runner-up") is a valid test of whether the winner is truly the best. We extend our analysis to provide equally simple procedures to obtain lower confidence bounds on the gap between the winning population and the others, and to verify ranks beyond the first.

1. Introduction.

1.1. *Motivating example*: *Iowa Republican caucus poll*. Table 1 shows the result of a Quinnipiac University poll asking 890 Iowa Republicans their preferred candidate for the Republican presidential nomination [Quinnipiac University Poll Institute (2016)]. Donald Trump led with 31% of the vote, Ted Cruz came second with 24%, Marco Rubio third with 17% and ten other candidates including "Don't know" trailed behind.

Seeing that Trump leads this poll, several salient questions may occur to us: Is Trump really winning, and if so by how much? Furthermore, is Cruz really in second, is Rubio really in third and so on? Note that there is implicitly a problem of multiple comparisons here, because if Cruz had led the poll instead, we would be asking a different set of questions ("Is Cruz really winning," etc.). Indeed, the selection issue appears especially pernicious due to the so-called "winner's curse": given that Trump leads the poll, it more likely than not overestimates his support.

Received February 2017; revised June 2017.

[1]Supported in part by the Gerald J. Lieberman Fellowship.

MSC2010 subject classifications. Primary 62F07; secondary 62F03.

Key words and phrases. Ranking, selective inference, exponential family, multiple comparison, sample best.

TABLE 1

Results from a February 1, 2016, *Quinnipiac University poll of* 890 *Iowa Republicans. To compute the last column* (*Votes*), *we make the simplifying assumption that the reported percentages in the third column* (*Result*) *are raw vote shares among survey respondents. The asterisks indicate that the rank is verified at level* 0.05 *by a stepwise procedure*

Rank	**Candidate**	**Result**	**Votes**
1*	Trump	31%	276
2*	Cruz	24%	214
3*	Rubio	17%	151
4*	Carson	8%	71
5	Paul	4%	36
6	Bush	4%	36
7	Huckabee	3%	27
⋮	⋮	⋮	⋮

Nevertheless, if we blithely ignore the selection issue, we might carry out the following analyses to answer the questions we posed before at significance level $\alpha = 0.05$. We assume for simplicity that the poll represents a simple random sample of Iowa Republicans; that is, the data are a multinomial sample of size 890 and underlying probabilities $(\pi_{\text{Trump}}, \pi_{\text{Cruz}}, \ldots)$. (The reality is a bit more complicated: before releasing the data, Quinnipiac has post-processed it to make the reported result more representative of likely caucus-goers. The raw data is proprietary.)

1. *Is Trump really winning?* If Trump and Cruz were in fact tied, then Trump's share of their combined 490 votes would be distributed as Binomial$(490, 0.5)$. Because the (two-tailed) p-value for this pairwise test is $p = 0.006$, we reject the null and conclude that Trump is really winning.

2. *By how much?* Using an exact 95% interval for the same binomial model, we conclude Trump has at least 7.5% more support than Cruz (i.e., $\pi_{\text{Trump}} \geq 1.075\pi_{\text{Cruz}}$) and also leads the other candidates by at least as much.

3. *Is Cruz in second, Rubio in third, etc.?* We can next compare Cruz to Rubio just as we compared Trump to Cruz (again rejecting because 214 is significantly more than half of 365), then Rubio to Carson, and so on, continuing until we fail to reject. The first four comparisons are all significant at level 0.05, but Paul and Bush are tied so we stop.

Perhaps surprisingly, all of the three procedures described above are statistically valid despite their ostensibly ignoring the implicit multiple-comparisons issue. In other words, Procedures 1 and 2 control the Type I error rate at level α and Procedure 3 controls the familywise error rate (FWER) at level α. The remainder

of this article is devoted to justifying these procedures for the multinomial family, and extending to analogous procedures in other exponential family settings. While methods analogous to Procedures 1 and 2 have been justified previously for balanced independent samples from log-concave location families [Gutmann and Maymin (1987), Stefansson, Kim and Hsu (1988)], they have not been justified in exponential families before now.

1.2. *Generic problem setting and main result.* Generically, we will consider data drawn from an exponential family model with density

$$X \sim \exp\bigl(\theta' x - \psi(\theta)\bigr) g(x), \tag{1}$$

with respect to either the Lebesgue measure on $\mathbb{R}^n$ or counting measure on $\mathbb{Z}^n$. We assume further that $g(x)$ is symmetric with respect to permutation, and Schur concave, a mild technical condition defined in Section 2. In addition to the multinomial family, model (1) also encompasses settings such as comparing independent binomial treatment outcomes in a clinical trial, competing sports teams under a Bradley–Terry model, entries of a Dirichlet distribution, and many more; see Section 2 for these and other examples.

We will generically use the term *population* to refer to the treatment group, sports team, political candidate, etc. represented by a given random variable X_j. As we will see, $\theta_j \geq \theta_k$ if and only if X_j is stochastically larger than X_k; thus, there is a well-defined stochastic ordering of the populations that matches the ordering of the entries of θ. We will refer to the population with maximal θ_j as the *best*, the population with second largest θ_j as the *second best*, the one with maximal X_j as the *winner* and the one with the second-largest X_j as the *runner-up*, where ties between observations are broken randomly to obtain a full ordering. Following the convention in the ranking and selection literature, we assume that if there are multiple largest θ_j, then one is arbitrarily marked as the best. Note that in cases where it is more interesting to ask which is the smallest population (e.g., if X_j is the number of patients on treatment j who suffer a heart attack during a trial) we can change the variables to $-X$ and the parameters to $-\theta$; this does not affect the Schur concavity assumption.

Write the order statistics of X as

$$X_{[1]} \geq X_{[2]} \geq \cdots \geq X_{[n]},$$

where $[j]$ will denote the random index for the jth order statistic. Thus, $\theta_{[j]}$ is the entry of θ corresponding to the jth order statistic of X (so $\theta_{[1]}$ might *not* equal $\max_j \theta_j$, for example).

In each of the above examples, there is a natural exact test we could apply to test $\theta_j = \theta_k$ for any two *fixed* populations j and k. In the multinomial case, we would apply the conditional binomial test based on the combined total $X_j + X_k$ as discussed in the previous section. For the case of independent binomials, we

would apply Fisher's exact test, again conditioning on $X_j + X_k$. These are both examples of a generic UMPU pairwise test in which we condition on the other $n-2$ indices (notated $X_{\setminus\{j,k\}}$) and $X_j + X_k$, and reject the null if X_j is outside the $\alpha/2$ and $1-\alpha/2$ quantiles of the conditional law $\mathcal{L}_{\theta_j=\theta_k}(X_j \mid X_j + X_k, X_{\setminus\{j,k\}})$. Crucially, this null distribution does not depend on the value of θ provided that $\theta_j = \theta_k$. We call this test the (two-tailed) *unadjusted pairwise test* since it makes no explicit adjustment for selection. Similarly, inverting this test for other values of $\theta_j - \theta_k$ yields an *unadjusted pairwise confidence interval*. (To avoid trivialities in the discrete case, we assume these procedures are appropriately randomized at the rejection thresholds to give exact level-α control.)

Generalizing the procedures described in Section 1.1, we obtain the following:

1. *Is the winner really the best?* To test the hypothesis $H : \theta_{[1]} \le \max_{j\neq[1]} \theta_j$: Carry out the unadjusted pairwise test comparing the winner to the runner-up. If the test rejects at level α, reject H and declare that the winner is really the best.

2. *By how much?* To construct a lower confidence bound for $\theta_{[1]} - \max_{j\neq[1]} \theta_j$: Construct the unadjusted pairwise confidence interval comparing the winner to the runner-up, and report the lower confidence bound obtained for $\theta_{[1]} - \theta_{[2]}$ if it is nonnegative, report $-\infty$ otherwise.

3. *Is the runner-up really the second best, etc.?* Continue by comparing the runner-up to the second runner-up, again using the unadjusted pairwise test, and so on down the list comparing adjacent values. Stop the first time the test does not reject; if there are j rejections, declare that

$$\theta_{[1]} > \theta_{[2]} > \cdots > \theta_{[j]} > \max_{k>j} \theta_{[j]}.$$

Procedures 2 and 3 are conservative stand-ins for exact, but slightly more involved, conditional inference procedures. In particular, as we will see, reporting $-\infty$ in Procedure 2 is typically much more conservative than is necessary.

We now state our main theorem: under a mild technical assumption, Procedures 1–3 described above are statistically valid, even accounting for the selection.

THEOREM 1. *Assume the model* (1) *holds and* $g(x)$ *is a Schur-concave function. Then*:

1. *Procedure* 1 *has exact level* α *conditional on* H *being true* (*conditional on the best population not winning*), *and marginally has level* $\alpha \cdot \mathbb{P}(H \text{ is true}) \le \alpha(1 - \frac{1}{n})$.

2. *Procedure* 2 *gives a conservative* $1-\alpha$ *lower confidence bound for* $\theta_{[1]} - \max_{j\neq[1]} \theta_j$.

3. *Procedure* 3 *is a conservative stepwise procedure with FWER no larger than* α.

Note that Theorem 1 implies that we could actually replace α with $\frac{n}{n-1}\alpha$ to obtain a more powerful version of Procedure 1 when n is not too large.

We define Schur concavity and discuss its properties in Section 2. Because any log-concave and symmetric function is Schur concave, Theorem 1 applies to all of the cases discussed above. The proof combines the conditional selective-inference framework of Fithian, Sun and Taylor (2014) with classical multiple-testing methods, as well as new technical tools involving majorization and Schur concavity.

Note that these procedures make an implicit adjustment for selection because they use two-tailed, rather than one-tailed, unadjusted tests. If we instead based our tests on an independent realization $X^* = (X_1^*, \ldots, X_n^*)$ then, for example, Procedure 1 could use a right-tailed version of the unadjusted pairwise test. In the case $n = 2$, Procedure 1 amounts to a simple two-tailed test of the null hypothesis $\theta_1 = \theta_2$, and it is intuitively clear that a one-tailed test would be too liberal. More surprising is that, no matter how large n is, Procedures 1–3 require no further adjustment beyond what is required when $n = 2$.

1.3. *Related work.* Rank verification has been studied extensively in the ranking and selection literature. See Gupta and Panchapakesan (1971, 1985) for surveys of the subset selection literature. The two main formulations of ranking and selection are closely related to procedures for multiple comparisons with the best treatment [Edwards and Hsu (1983), Hsu (1984)], but more powerful methods are available in some cases for procedures involving only the first sample rank, the problem of comparisons with the sample best; see Hsu (1996) for an overview and discussion of the relationships between these problems.

Comparisons with the sample best have been especially well studied and the validity of Procedures 1 and 2 have been established in a different setting: balanced independent samples from log-concave location families. Gutmann and Maymin (1987) prove the validity of Procedure 1 in this setting, and Bofinger (1991), Kannan and Panchapakesan (2009), Maymin and Gutmann (1992) give similar results for other models including scale and location-scale families. Stefansson, Kim and Hsu (1988) provide an alternative proof for the validity of Procedure 1 in the same setting, leading to a lower confidence bound analogous to that of Procedure 2; interestingly, the proof involves a very early application of the partitioning principle, later developed into fundamental technique in multiple comparisons [Finner and Strassburger (2002)]. These results use very different technical tools than the ones we use here, require independence between the different groups (ruling out, e.g., the multinomial family), and do not address the exponential family case. Because most exponential families are not location-scale families (the Gaussian being a notable exception), and because our results involve more general dependence structures, both our proof techniques and our technical results are complementary to the techniques and results in the above works.

For the multinomial case, Gupta and Nagel (1967), discussed in Section 3.1, remain the state of the art in finite-sample tests; Gupta and Wong (1976) discuss related approaches for Poisson models. Berger (1980) mentions an alternative, simpler rule which performs a binomial test on each population, but its

power does not necessarily increase as the size m of observations increases in cases like Multinomial$(m; 2/3, 1/3, 0, \ldots, 0)$. Nettleton (2009) proves validity for an asymptotic version of the winner-versus-runner-up test, and Gupta and Liang (1989) consider an empirical Bayes approach for selecting the best binomial population wherein a parametric prior distribution is assumed for the success probabilities for the different populations. Ng and Panchapakesan (2007) discuss an exact test for a modified problem in which the maximum count is fixed instead of the total count; that is, we sample until the leading candidate has at least m votes. As Section 3.1 shows, our test can be much more powerful than the one in Gupta and Nagel (1967), especially if there are many candidates, because of the way our critical rejection threshold for $X_{[1]} - X_{[2]}$ adapts to the data. Thus, our work closes a significant gap in the ranking and selection literature, extending the result of Gutmann and Maymin (1987) and others to new families like the multinomial, independent binomials and many others.

1.4. *Outline*. Section 2 defines Schur concavity, and gives several examples satisfying this condition. Section 3 justifies Procedure 1 and compares its power to that of Gupta and Nagel (1967). Sections 4 and 5 justify Procedures 2 and 3, respectively, and Section 6 concludes.

2. Majorization and Schur concavity.

2.1. *Definitions and basic properties*. We start by reviewing the notion of *majorization*, defined on both $\mathbb{R}^n$ and $\mathbb{Z}^n$.

DEFINITION 1. For two vectors a and b in $\mathbb{R}^n$ (or $\mathbb{Z}^n$), suppose sorting the two vectors in descending order gives $a_{(1)} \geq \cdots \geq a_{(n)}$ and $b_{(1)} \geq \cdots \geq b_{(n)}$. We say that $a \succeq b$ (a majorizes b) if for $1 \leq i < n$,

$$a_{(1)} + \cdots + a_{(i)} \geq b_{(1)} + \cdots + b_{(i)} \quad \text{and}$$
$$a_{(1)} + \cdots + a_{(n)} = b_{(1)} + \cdots + b_{(n)}.$$

This forms a partial order in $\mathbb{R}^n$ (or $\mathbb{Z}^n$).

Intuitively, majorization is a partial order that monitors the evenness of a vector: the more even a vector is, the "smaller" it is. There are two properties of majorization that we will use in the proofs.

LEMMA 2. 1. *Suppose* $(x_1, x_2, x_3, \ldots)$ *and* $(x_1, y_2, y_3, \ldots)$ *are two vectors in* $\mathbb{R}^n$. *Then*

$$(x_1, x_2, x_3, \ldots) \succeq (x_1, y_2, y_3, \ldots) \quad \textit{if and only if} \quad (x_2, x_3, \ldots) \succeq (y_2, y_3, \ldots).$$

2. (*Principle of transfer*) *If* $x_1 > x_2$ *and* $t \geq 0$, *then*

$$(x_1 + t, x_2, x_3, \ldots) \succeq (x_1, x_2 + t, x_3, \ldots).$$

If $t \leq 0$, *the majorization is reversed.*

PROOF. 1. The property follows from an equivalent formulation of majorization listed in Marshall, Olkin and Arnold (2011), where $x \succeq y$ if and only if

$$\sum_{j=1}^{n} x_n = \sum_{j=1}^{n} y_n \quad \text{and} \quad \sum_{j=1}^{n} (x_j - a)_+ \geq \sum_{j=1}^{n} (y_j - a)_+ \qquad \text{for all } a \in \mathbb{R}.$$

2. Proved in Marshall, Olkin and Arnold (2011). □

DEFINITION 2. A function g is Schur concave if $x \succeq y$ implies $g(x) \leq g(y)$.

A Schur-concave function is symmetric by default since $a \succeq b$ and $b \succeq a$ if and only if b is a permutation of the coordinates of a. Conversely, a symmetric and log-concave function is Schur concave [Marshall, Olkin and Arnold (2011)]. Interestingly, Gupta, Huang and Panchapakesan (1984) also show that, in the context of independent location families, Schur concavity of the probability density is equivalent to monotone likelihood ratio.

2.2. *Examples*. Many common exponential family models have Schur-concave carrier densities. Below we give a few examples:

EXAMPLE 1 (Independent binomial treatment outcomes in a clinical trial). If each of n different treatments are applied to m patients independently, the number of positive outcomes X_j for treatment j is Binomial(m, p_j). The best treatment would be the treatment with the highest success probability p_j. The joint distribution of X is given by

$$p(x) \propto \exp\left(\sum_j x_j \log \frac{p_j}{1-p_j}\right) \frac{1}{x_1!(m-x_1)! \cdots x_n!(m-x_n)!}$$

The carrier measure above is Schur concave. The unadjusted pairwise test in this family is Fisher's exact test.

EXAMPLE 2 (Competitive sports under the Bradley–Terry model). Suppose n players compete in a round robin tournament, where player j has ability θ_j, and the probability of player j winning against player k is

$$\frac{e^{\theta_j - \theta_k}}{1 + e^{\theta_j - \theta_k}} = \frac{e^{(\theta_j - \theta_k)/2}}{e^{(\theta_j - \theta_k)/2} + e^{(\theta_k - \theta_j)/2}}.$$

Let Y_{jk} be an indicator for the match between player j and k, where we take $Y_{jk} = 1$ if j beats k and $Y_{jk} = 0$ if k beats j. For symmetry, we will also adopt the convention that $Y_{jk} + Y_{kj} = 1$. Thus the joint distribution of $Y = (Y_{jk})_{j \neq k}$ is

$$p(y) \propto \exp\left(\sum_j 2\theta_j \sum_{k \neq j} y_{jk}\right) = \exp(2\theta' x),$$

where $x_j = \sum_{k \neq j} y_{jk}$. In other words, if X_j is the number of wins by player j, then $X = (X_1, \ldots, X_n)$ is a sufficient statistic with distribution

$$p(x) = \exp(2\theta' x) g(x),$$

where $g(x)$ is a function that counts the number of possible tournament results giving the net win vector x. A bijection proof shows that x is indeed Schur concave. Therefore, we can use Procedures 1–3 to compare player qualities.

After conditioning on $U(X) = (X_1 + X_2, X_3, \ldots, X_n)$, and under the assumption $\theta_1 = \theta_2$, every feasible configuration of Y is equally likely. If n is not too large (say, no more than 40 players), we can find the conditional distribution of X_1 by enumerating over the configurations; for larger n, computation might pose a more serious problem, requiring us, for example, to compute the p-value using Markov Chain Monte Carlo techniques [Besag and Clifford (1989)].

EXAMPLE 3 (Comparing the variances of different normal populations). Suppose there are n normal populations with laws $N(\mu_j, \sigma_j^2)$ and m independent observations from each of them. The sample variance for population j can be denoted as R_j. By Cochran's theorem, $(m-1)R_j \sim \sigma_j^2 \chi_{m-1}^2$, and thus the joint distribution of R is

$$\begin{aligned} r &\sim \prod_{j=1}^{n} \left(\frac{(m-1)r_j}{\sigma_j^2} \right)^{(m-3)/2} e^{-(m-1)r_j/2\sigma_j^2} 1_{\{r_j > 0\}} \\ &\propto \exp\left(-\frac{m-1}{2\sigma_1^2} r_1 - \cdots - \frac{m-1}{2\sigma_n^2} r_n \right) \prod_{j=1}^{n} r_j^{(m-3)/2} 1_{\{r>0\}}. \end{aligned}$$

The carrier measure is $\prod_{j=1}^{n} r_j^{(m-3)/2} 1_{\{r>0\}}$, which is Schur concave. Thus, we can use Procedures 1–3 to find populations with the smallest or largest variances. In this example, the distribution of $X_1/(X_1 + X_2)$ conditional on $(X_1 + X_2, X_3, \ldots, X_4)$ is distributed as Beta$(m/2, m/2)$ under the null, or equivalently X_1/X_2 is conditionally distributed as $F_{m,m}$; hence a (two-tailed) F-test is valid for comparing the top two populations.

3. Verifying the winner: Is the winner really the best? First, we justify the notion that the population with largest θ_j is also the largest population in stochastic order.

THEOREM 3. *For a multivariate exponential family with a symmetric carrier distribution, $X_1 \geq X_2$ in stochastic order if and only if $\theta_1 \geq \theta_2$.*

PROOF. It suffices to prove the "if" part, as the "only if" part can be follows from swapping the role of θ_1 and θ_2. For any fixed a, and $x_1 \geq a$ and $x_2 < a$, we

have $x_1 > x_2$ and

$$\exp\big(\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n - \psi(\theta)\big) g(x)$$
$$\geq \exp\big(\theta_1 x_2 + \theta_2 x_1 + \cdots + \theta_n x_n - \psi(\theta)\big) g(x).$$

Integrating both sides over the region $\{x : x_1 \geq a, x_2 < a\}$ gives

$$\mathbb{P}[X_1 \geq a, X_2 < a] \geq \mathbb{P}[X_1 < a, X_2 \geq a].$$

Now adding $\mathbb{P}[X_1 \geq a, X_2 \geq a]$ to both probabilities gives

$$\mathbb{P}[X_1 \geq a] \geq \mathbb{P}[X_2 \geq a],$$

meaning that X_1 is greater than X_2 in stochastic order. □

Before proving our main result for Procedure 1, we give the following lemmas, the first of which clarifies a key idea in the proof, and the second is needed for a sharper bound in (2).

LEMMA 4 [Berger (1982)]. *If p_j are valid p-values for testing null hypothesis H_{0j}, then $p_* = \max_j p_j$ is a valid p-value for the union null (i.e., disjunction null) hypothesis $H_0 = \bigcup_j H_{0j}$.*

PROOF. Under H_0, one of the H_{0j} is true; without loss of generality assume it is H_{01}. Then

$$\mathbb{P}[p_* \leq \alpha] \leq \mathbb{P}[p_1 \leq \alpha] \leq \alpha.$$

Therefore, p_* is a valid p-value for the union null hypothesis. □

LEMMA 5. *If $\theta_1 \geq \max_{j \neq 1} \theta_j$, then $\mathbb{P}[1 \text{ wins}] \geq \frac{1}{n}$.*

PROOF. We can prove so with a coupling argument: for any sequence $x_1, x_2, \ldots, x_n$, define $\tau(x) = \{\tau(x_j)\}_{j=1,\ldots,n}$, obtained by swapping x_1 with the largest value in the sequence x. Hence

$$\exp\big(\theta_1 \tau(x_1) + \cdots + \theta_n \tau(x_n) - \psi(\theta)\big) g(X)$$
$$\geq \exp\big(\theta_1 x_1 + \cdots + \theta_n x_n - \psi(\theta)\big) g(X).$$

If we integrating both sides over $\mathbb{R}^n$ (or $\mathbb{Z}^n$ in the case of counting measure), the right-hand side gives 1. Since τ is an n-to-1 mapping, the left-hand side is n times the integral over $\{x_1 \geq \max_{j>1} x_j\}$. In other words,

$$n\mathbb{P}[1 \text{ wins}] \geq 1$$

as desired.

In the case of counting measure, the above argument follows if a subscript is attached to identical observations uniformly to ensure strict ordering. □

We are now ready to prove our result for Procedure 1, restated here for reference.

PART 1 OF THEOREM 1. *Assume the model* (1) *holds and* $g(x)$ *is a Schur-concave function. Procedure* 1 (*the unadjusted pairwise test*) *has level* α *conditional on the best population not winning.*

PROOF. Let j^* denote the (fixed) index of the best population, so $\theta_{j^*} \geq \max_{j \neq j^*} \theta_j$. The type I error—the probability of incorrectly declaring any other j to be the best—is

$$\mathbb{P}\Big[\bigcup_{j \neq j^*} \text{declare } j \text{ best}\Big] \leq \sum_{j \neq j^*} \mathbb{P}[\text{declare } j \text{ best} \mid j \text{ wins}]\mathbb{P}[j \text{ wins}],$$

recalling that ties are broken randomly, so there is only one winner in any realization. Thus, it is enough to bound $\mathbb{P}_\theta[\text{declare } j \text{ best} \mid j \text{ wins}] \leq \alpha$, for each $j \neq j^*$, and for all θ with $j^* \in \arg\max_j \theta_j$. Then we will have

$$\begin{aligned}\mathbb{P}\Big[\bigcup_{j \neq j^*} \text{declare } j \text{ best}\Big] &\leq \sum_{j \neq j^*} \alpha \cdot \mathbb{P}[j \text{ wins}] \\ &= \alpha \mathbb{P}\big[j^* \text{ does not win}\big] \leq \frac{n-1}{n}\alpha,\end{aligned} \tag{2}$$

where the last inequality follows from Lemma 5.

We start by assuming that we are working with the Lebesgue measure rather than the counting measure (eliminating the possibility of ties). The necessary modification of the proof for the counting measure case is provided at the end of this proof.

To minimize notational clutter, we consider only the case where the winner is 1, that is, $X_1 \geq \max_{j>1} X_j$. Furthermore, we will denote the runner-up with 2. This is not necessarily true, but we will use it as a shorthand to simplify our notation. For other cases, the following proof remains valid under relabeling and can thus be applied. In this case, we will test the null hypothesis $H_{01} : \theta_1 \leq \max_{j>1} \theta_j$, which is the union of the null hypotheses $H_{01j} : \theta_1 \leq \theta_j$ for $j \geq 2$. For each of these, we can construct an exact p-value p_{1j}, which is valid under H_{01j} conditional on A_1, the event that X_1 is the winner. Hence by Lemma 4, a test that rejects when $p_{1*} = \max_j p_{1j} \leq \alpha$ is valid for H_{01} conditional on A_1. Procedure 1 performs an unadjusted pairwise test comparing X_1 to X_2. Hence it is sufficient to show that $p_{12} = p_{1*}$ and that rejecting when $p_{12} \leq \alpha$ coincides with the unadjusted pairwise test.

Our proof has three main parts: (1) deriving p_{1j} for each $j \geq 2$, (2) showing that $p_{12} \geq p_{1j}$ for each $j \geq 2$, and (3) showing that p_{12} is an unadjusted pairwise p-value.

Derivation of p_{1j}. Following the framework in Fithian, Sun and Taylor (2014), we first construct the p-values by conditioning on the selection event where the

winner is 1:

$$A_1 = \Big\{ X_1 \geq \max_{j>1} X_j \Big\}.$$

For convenience, we let

$$D_{jk} = \frac{X_j - X_k}{2} \quad \text{and} \quad M_{jk} = \frac{X_j + X_k}{2}.$$

We then reparametrize to replace X_1 and X_j with D_{1j} and M_{1j}. The distribution is now an exponential family with sufficient statistics $D_{1j}, M_{1j}, X_{\setminus\{1,j\}}$ and corresponding natural parameters $\theta_1 - \theta_j, \theta_1 + \theta_j, \theta_{\setminus\{1,j\}}$. We now consider

$$\mathcal{L}_{\theta_1 - \theta_j = 0}(D_{1j} \mid M_{1j}, X_{\setminus\{1,j\}}, A_1). \tag{3}$$

We can rewrite the selection event in terms of our new parameterization as

$$\begin{aligned} A_1 &= \{X_1 \geq X_j\} \cap \Big\{ X_1 \geq \max_{k \neq 1,j} X_k \Big\} \\ &= \{D_{1j} \geq 0\} \cap \Big\{ D_{1j} \geq \max_{k \neq 1,j} X_k - M_{1j} \Big\}. \end{aligned}$$

The conditional law of D_{1j} in (3), in particular, is a truncated distribution

$$\begin{aligned} &p(d_{1j} \mid M_{1j}, X_{\setminus\{1,j\}}, A_1) \\ &\quad \propto \exp\big((\theta_1 - \theta_j) d_{1j} + \theta_2 X_2 + \cdots + (\theta_1 + \theta_j) M_{1j} + \cdots + \theta_n X_n\big) \\ &\qquad \times g(M_{1j} + d_{ij}, X_2, \ldots, M_{ij} - d_{ij}, \ldots X_n) 1_{A_1} \\ &\quad \overset{(a)}{\propto} g(M_{1j} + d_{1j}, X_2, \ldots, M_{1j} - d_{1j}, \ldots X_n) 1_{A_1}, \end{aligned}$$

where at step (a), conditioning on $X_{\setminus\{1,j\}}$ and M_{1j} removes dependence on $\theta_{\setminus\{1,j\}}$ and $\theta_1 + \theta_j$, respectively, while $\theta_1 - \theta_j$ is taken to be 0 under our null hypothesis. Note that we consider this as a one-dimensional distribution of D_{1j} on $\mathbb{R}$, where M_{1j} and $X_{\setminus\{1,j\}}$ are treated as fixed.

The p-value for H_{01j} is thus

$$p_{1j} = \frac{\int_{D_{1j}}^{\infty} g(M_{1j} + z, X_2, \ldots, M_{1j} - z, \ldots, X_n)\, dz}{\int_{\max\{X_2 - M_{1j}, 0\}}^{\infty} g(M_{1j} + z, X_2, \ldots, M_{1j} - z, \ldots, X_n)\, dz}. \tag{4}$$

Finally, by construction, p_{1j} satisfies

$$\mathbb{P}_{H_{01j}}[p_{1j} < \alpha \mid M_{1j}, X_{\setminus\{1,j\}}, A_1] \leq \alpha \qquad \text{a.s.}$$

Marginalizing over $M_{1j}, X_{\setminus\{1,j\}}$,

$$\mathbb{P}_{H_{01j}}[p_{1j} < \alpha \mid A_1] \leq \alpha.$$

Therefore, these p_{1j} are indeed valid p-values.

Demonstration that $p_{1*} = p_{12}$. We now proceed to show that p_{12}, the p-value comparing the winner to the runner-up, is the largest of all p_{1j}. Without loss of generality, it is sufficient to show that $p_{12} \geq p_{13}$.

From the first part of this proof, both p-values are constructed by conditioning on $X_{\setminus\{1,2,3\}}$. Upon conditioning these, (X_1, X_2, X_3) follows an exponential family distribution, with carrier distribution

$$g_{X_4,\ldots,X_n}(X_1, X_2, X_3) = g(X_1, \ldots, X_n),$$

here, $X_4, \ldots, X_n$ are used in the subscript as they are conditioned on and no longer considered as variables. The first point in Lemma 2 says that the function $g_{X_4,\ldots,X_n}$ is Schur concave as well. We have reduced the problem to the case when $n = 3$: we can apply the result for $n = 3$ to $g_{X_4,\ldots,X_n}$ to yield $p_{12} \geq p_{13}$ for $n > 3$.

We have reduced to the case when $n = 3$. The p-values thus are

$$p_{12} = \frac{\int_{D_{12}}^{\infty} g(M_{12} + z, M_{12} - z, X_3)\, dz}{\int_{0}^{\infty} g(M_{12} + z, M_{12} - z, X_3)\, dz},$$

$$p_{13} = \frac{\int_{D_{13}}^{\infty} g(M_{13} + z, X_2, M_{13} - z)\, dz}{\int_{\max\{X_2 - M_{13}, 0\}}^{\infty} g(M_{13} + z, X_2, M_{13} - z)\, dz}.$$

The maximum in the denominator of p_{13} prompts us to consider two separate cases. First, we suppose $X_2 < M_{13}$. Changing variables such that the lower limits of both integrals in the numerator are 0, we can reparametrize the integrals above to give

$$\begin{aligned} p_{12} &= \frac{\int_{0}^{\infty} g(X_1 + z, X_2 - z, X_3)\, dz}{\int_{0}^{\infty} g(M_{12} + z, M_{12} - z, X_3)\, dz} \\ &= \frac{\int_{0}^{\infty} g(X_1 + z, X_2 - z, X_3)\, dz}{\int_{-D_{12}}^{\infty} g(X_1 + z, X_2 - z, X_3)\, dz}, \\ p_{13} &= \frac{\int_{0}^{\infty} g(X_1 + z, X_2, X_3 - z)\, dz}{\int_{0}^{\infty} g(M_{13} + z, X_2, M_{13} - z)\, dz} \\ &= \frac{\int_{0}^{\infty} g(X_1 + z, X_2, X_3 - z)\, dz}{\int_{-D_{13}}^{\infty} g(X_1 + z, X_2, X_3 - z)\, dz}. \end{aligned}$$

To help see the re-parametrization, each of these integrals can be thought of in terms of integrals along segments and rays. For example, p_{12} can be represented in terms of integrals A and B in Figure 1. Specifically,

$$p_{12} = \frac{B}{A + B}.$$

Figure 2 has both the p-values shown on the same diagram. Proving $p_{12} \geq p_{13}$ is the same as proving

$$\frac{B}{A + B} \geq \frac{D}{C + D} \quad \Longleftrightarrow \quad \frac{B}{A} \geq \frac{D}{C}.$$

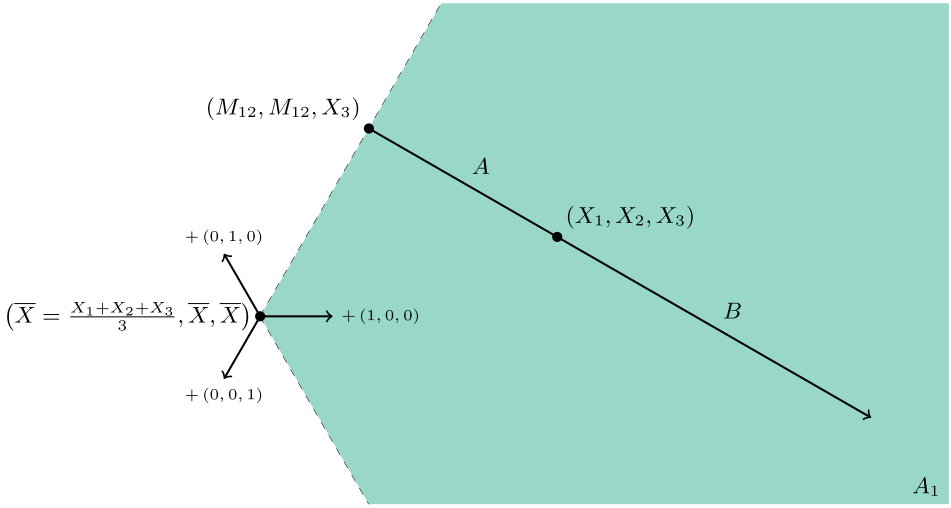

FIG. 1. *The p-value p_{12} can be written in terms of integral A along the segment and B along the ray. The diagram is drawn a level set of $x_1 + x_2 + x_3$. The green region represents the selection event A_1.*

We will prove so by extending A to include $\tilde{A}$ on the diagram. We denote the sum $A + \tilde{A}$ as A'. Formally,

$$\begin{aligned} A' &= \int_{-D_{13}}^{0} g(X_1 + z, X_2 - z, X_3)\, dz \\ &\geq \int_{-D_{12}}^{0} g(X_1 + z, X_2 - z, X_3)\, dz = A. \end{aligned} \tag{5}$$

It is thus sufficient to show that $B \geq D$ and $C \geq A'$.

Indeed from the second point in Lemma 2, we have

$$(X_1 + z, X_2 - z, X_3) \succeq (X_1 + z, X_2, X_3 - z)$$

for $z \leq 0$ and the majorization reversed for $z \geq 0$. This majorization relation is indicated as the dotted line in Figure 2. So Schur concavity shows that

$$g(X_1 + z, X_2 - z, X_3) \leq g(X_1 + z, X_2, X_3 - z)$$

for $z \leq 0$, and the inequality reversed for $z \geq 0$. Taking integrals on both sides yields the desired inequality.

For the second case where $X_2 \geq M_{13}$, the segment C will reach the line $x_1 = x_2$ first before it reaches $x_1 = x_3$, ending at $(X_2, X_2, X_1 - X_2 + X_3)$ instead. But we can still extend A by $\tilde{A}$ to (X_2, X_1, X_3). The rest of the proof follows. In either cases, $p_{12} \geq p_{13}$, or in generality, $p_{12} \geq p_{1j}$ for $j > 1$. In other words, $p_{12} = p_{1*}$.

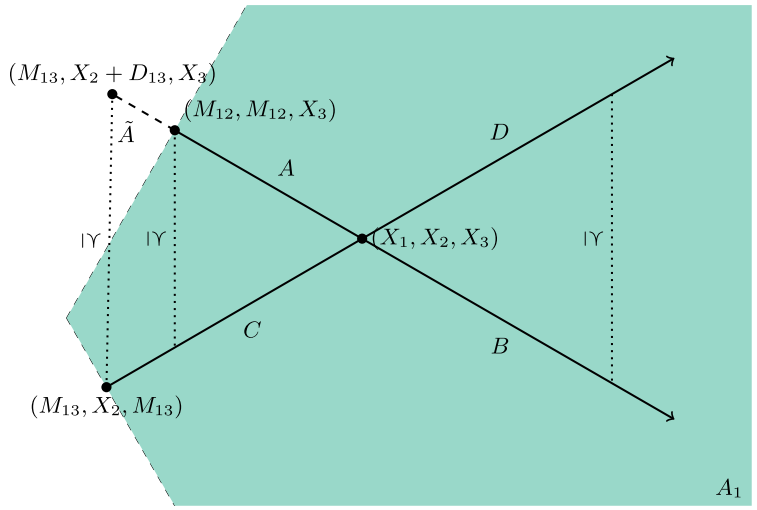

FIG. 2. *The p-value p_{12} can be written in terms of integral A along the segment and B along the ray; and p_{13} in terms of C and D. A' would refer to the sum of A with the dashed line portion labeled as $\tilde{A}$, formally explained in equation (5). The majorization relation is indicated by the dotted line.*

p_{12} is an unadjusted pairwise p-value. Before conditioning on A_1, the distribution in (3) is symmetric around 0 under $\theta_1 = \theta_j$. Since the denominator of p_{12} integrates over half of this symmetric distribution, it is always equal to $1/2$. Thus, the one-sided conditional test at level α is equivalent to the one-sided unadjusted test at level $\alpha/2$, or equivalently the two-sided unadjusted pairwise test at level α.

Modification for counting measure. Now suppose the exponential family is defined on the counting measure instead. If ties are broken independently and randomly, the end points on the rays can be considered as "half an atom" if the coordinates are integers (or a smaller fraction of an atom in case of a multi-way tie). The number of atoms on each ray is the same (after the extension $\tilde{A}$) and the atoms on each ray can be paired up in exactly the same way as illustrated in Figure 2, with the inequalities above still holding for each pair of the atoms. Summing these inequalities yields our desired result. □

3.1. *Power comparison in the multinomial case.* As the construction of this test follows Fithian, Sun and Taylor (2014), it uses UMPU selective level-α tests for the pairwise p-values. This section compares the power of our procedure to the best previously known method for verifying multinomial ranks, by Gupta and Nagel (1967). They devise a rule to select a subset that includes the maximum π_j. In other words, if the selected subset is $J(X)$, it guarantees

$$\mathbb{P}\big[\arg\max_j \pi_j \in J(X)\big] \geq 1 - \alpha. \tag{6}$$

This is achieved by finding an integer d, as a function on m, n and α, and selecting the subset

$$J(X) = \left\{ j : X_j \geq \max_k X_k - d \right\}.$$

We take $d(m, n, \alpha)$ to be the smallest integer such that (6) holds for any π; Gupta and Nagel (1967) provide an algorithm for determining d.

Subset selection is closely related to testing whether the winner is the best. In particular, we can define a test that declares j the best whenever $J(X) = \{j\}$. If $J(X)$ satisfies (6), this test is valid at level α. We next compare the power of the resulting test against the power of our Procedure 1 in a multinomial example with $\pi \propto (e^\delta, 1, \ldots, 1)$, for several combinations of m and n.

Figure 3 gives the power curves for Multinomial(m, π) and

$$\pi \propto \left(e^\delta, 1, \ldots, 1\right),$$

for various combinations of m and n. For their method, we use $\alpha = 0.05$; but in light of the extra factor of $\frac{n-1}{n}$ in (2), we will apply the selective procedure with $\frac{n}{n-1}\alpha$ such that the marginal type I error rate of both procedures are controlled

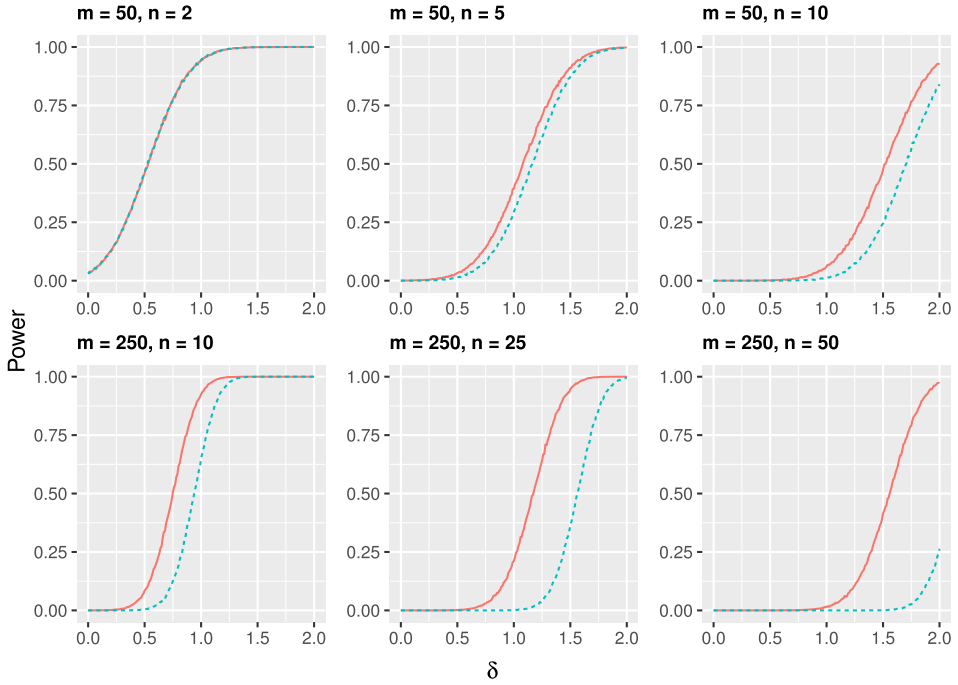

FIG. 3. *Power curves as a function of δ. The plots in the first row all have $m = 50$ and the second row $m = 250$. The solid line and the dashed line are the power for the selective test and Gupta and Nagel's test, respectively.*

at α. Their test coincides with our test at $n = 2$; however, as n grows, the selective test shows significantly more power than Gupta and Nagel's test.

To interpret, for example, the upper right panel of Figure 3, suppose that in a poll of $m = 50$ respondents, one candidate enjoys 30% support and the other $n - 1 = 9$ split the remainder ($\delta = \log \frac{0.3}{0.7/9} \approx 1.35$). Then our procedure has power approximately 0.3 to detect the best candidate, while Gupta and Nagel's procedure has power around 0.1.

To understand why our method is more powerful, note that both procedures operate by comparing $X_{[1]} - X_{[2]}$ to some threshold, but the two methods differ in how that threshold is determined. The threshold from Gupta and Nagel (1967) is fixed and depends on m and n alone, whereas in our procedure the threshold depends on $X_{[1]} + X_{[2]}$, a data-adaptive choice.

The difference between the two methods is amplified when n is large and $\pi_{(1)} \ll 1/2$. In that case, d from Gupta and Nagel is usually computed based on the worst-case scenario $\pi = (\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0)$; that is, d is the upper α quantile of

$$X_1 - X_2 \sim m - 2 \cdot \text{Binomial}\left(m, \frac{1}{2}\right) \approx \text{Normal}(0, m).$$

Thus $d \approx \sqrt{m} z_\alpha$, where z_α is the upper α quantile of a standard Gaussian. On the other hand, our method defines a threshold based on the upper $\frac{n}{n-1} \cdot \frac{\alpha}{2}$ quantile of

$$X_1 - X_2 \mid X_1 + X_2 \sim X_1 + X_2 - 2 \cdot \text{Binomial}\left(X_1 + X_2, \frac{1}{2}\right),$$

which is approximately $\sqrt{X_1 + X_2} z_{\alpha/2}$. If $\pi_{(1)} \ll 1/2$, then with high probability $X_1 + X_2 \ll m$, making our test much more liberal.

4. Confidence bounds on differences: By how much? By generalizing the above, we can construct a lower confidence bound for $\theta_{[1]} - \max_{j \neq [1]} \theta_j$. Here, we provide a more powerful Procedure 2′ first. We will proceed by inverting a statistical test of the hypothesis $H^\delta_{0[1]} : \theta_{[1]} - \max_{j \neq [1]} \theta_j \leq \delta$, which can be written as a union of null hypotheses:

$$H^\delta_{0[1]} = \bigcup_{j \neq [1]} H_{0[1]j} : \theta_{[1]} - \theta_j \leq \delta.$$

By Lemma 4, we can construct selective exact one-tailed p-values $p^\delta_{[1]j}$ for each of these by conditioning on $A_{[1]}$, $M_{[1]j}$ and $X_{\setminus\{[1],j\}}$, giving us an exact test for $H_{0[1]}$ by rejecting whenever $\max_{j \neq [1]} p^\delta_{[1]j} < \alpha$.

THEOREM 6. *The p-values constructed above satisfy* $p^\delta_{[1][2]} \geq p^\delta_{[1]j}$ *for any* $j \neq [1]$.

PROOF. Again we start with assuming $X_1 \geq X_2 \geq \max_{j>2} X_j$ for convenience. The p-values in question are derived from the conditional law

$$\mathcal{L}_{\theta_1-\theta_j=\delta}(D_{1j} \mid M_{1j}, X_2, \ldots, X_n, A),$$

which is the truncated distribution

$$\begin{aligned} p(d_{1j}) &\propto \exp\big((\theta_1-\theta_j)d_{1j} + \theta_2 X_2 + \cdots + (\theta_1+\theta_j)M_{1j} + \cdots + \theta_n X_n\big) \\ &\quad \times g(M_{1j}+d_{1j}, X_2, \ldots, M_{1j}-d_{1j}, \ldots X_n)1_{A_1} \\ &\propto \exp(\delta d_{1j}) g(M_{1j}+d_{1j}, X_2, \ldots, M_{1j}-d_{1j}, \ldots, X_n)1_{A_1}. \end{aligned}$$

The p-values thus are

$$p^{\delta}_{1j} = \frac{\int_{D_{1j}}^{\infty} \exp(\delta z) g(M_{1j}+z, X_2, \ldots, M_{1j}-z, \ldots, X_n)\, dz}{\int_{\max\{X_2-M_{1j},0\}}^{\infty} \exp(\delta z) g(M_{1j}+z, X_2, \ldots, M_{1j}-z, \ldots, X_n)\, dz}.$$

As before in Part 1 of Theorem 1, the conditioning reduces to the case where $n=3$. Once again, it is sufficient to show that $p_{12} \geq p_{13}$. We have the same two cases. If $X_2 < M_{13}$, then

$$\begin{aligned} p^{\delta}_{12} &= \frac{\int_0^{\infty} \exp(\delta(z+D_{12})) g(X_1+z, X_2-z, X_3)\, dz}{\int_{-D_{12}}^{\infty} \exp(\delta(z+D_{12})) g(X_1+z, X_2-z, X_3)\, dz} \\ &= \frac{\int_0^{\infty} \exp(\delta z) g(X_1+z, X_2-z, X_3)\, dz}{\int_{-D_{12}}^{\infty} \exp(\delta z) g(X_1+z, X_2-z, X_3)\, dz}, \\ p^{\delta}_{13} &= \frac{\int_0^{\infty} \exp(\delta(z+D_{13})) g(X_1+z, X_2, X_3-z)\, dz}{\int_{-D_{13}}^{\infty} \exp(\delta(z+D_{13})) g(X_1+z, X_2, X_3-z)\, dz} \\ &= \frac{\int_0^{\infty} \exp(\delta z) g(X_1+z, X_2, X_3-z)\, dz}{\int_{-D_{13}}^{\infty} \exp(\delta z) g(X_1+z, X_2, X_3-z)\, dz}. \end{aligned}$$

The same argument in Figure 2 shows that $p^{\delta}_{12} \geq p^{\delta}_{13}$. This is again true for the case where $X_2 \geq M_{13}$ as well. □

In other words, Procedure 2′ can be summarized as: Find the minimum δ such that $p^{\delta}_{[1][2]} \leq \alpha$. And by construction, Procedure 2′ gives exact $1-\alpha$ confidence bound for $\theta_{[1]} - \max_{j\neq[1]} \theta_j$.

PART 2 OF THEOREM 1. *Assume the model* (1) *holds and* $g(x)$ *is a Schur-concave function. Procedure* 2 (*the lower bound of unadjusted pairwise confidence interval*) *gives a conservative* $1-\alpha$ *lower confidence bound for* $\theta_{[1]} - \max_{j\neq[1]} \theta_j$.

PROOF. When Procedure 2 reports $-\infty$ as a confidence lower bound, it is definitely valid and conservative. It remains to show that when Procedure 2 reports

a finite confidence lower bound, it is smaller than the confidence lower bound reported by Procedure 2′.

If Procedure 2 reports a finite confidence lower bound δ^*, then $\delta^* \geq 0$. Also

$$\frac{\alpha}{2} = \frac{\int_{D_{12}}^{\infty} \exp(\delta^* z) g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz}{\int_{-\infty}^{\infty} \exp(\delta^* z) g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz} \tag{7}$$

as Procedure 2 is constructed from an unadjusted two-tail pairwise confidence interval. However, as $\delta^* \geq 0$, we have

$$\frac{\int_{-\infty}^{0} \exp(\delta^* z) g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz}{\int_{0}^{\infty} \exp(\delta^* z) g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz} \leq 1,$$

$$\frac{\int_{-\infty}^{\infty} \exp(\delta^* z) g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz}{\int_{0}^{\infty} \exp(\delta^* z) g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz} \leq 2.$$

Multiplying this to (7), we have

$$\alpha \geq \frac{\int_{D_{12}}^{\infty} \exp(\delta^* z) g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz}{\int_{0}^{\infty} \exp(\delta^* z) g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz},$$

indicating that δ^* is smaller than the confidence bound that Procedure 2′ would report. Hence δ^* is a valid and conservative. □

Note that Procedure 2 reporting $-\infty$ in case of $\delta^* \leq 0$ is rather extreme. In reality, we can always just adopt Procedure 2′ in the case when Procedure 1 rejects. In fact, by Procedure 2′, the multinomial example for polling in Section 1.1 can give a stronger lower confidence bound, that $\pi_{\text{Trump}}/\max_{j \neq \text{Trump}} \pi_j \geq 1.108$ (Trump leads the field by at least 10.8%).

5. Verifying other ranks: Is the runner-up really the second best, etc.? Often we will be interested in verifying ranks beyond the winner. More generally, we could imagine declaring that the first j populations are all in the correct order, that is,

$$\theta_{[1]} > \cdots > \theta_{[j]} > \max_{k>j} \theta_{[k]}. \tag{8}$$

Let j_0 denote the largest j for which (8) is true. Note that j_0 is both random and unknown, because it depends on both the data and population ranks. Procedure 3 declares that $j_0 \geq j$ if the unadjusted pairwise tests between $X_{[k]}$ and $X_{[k+1]}$, reject at level α for *all* of $k = 1, \ldots, j$.

In terms of the Iowa polling example of Section 1, we would like to produce a statement of the form "Trump has the most support, Cruz has the second-most and Rubio has the third-most." Procedure 3 performs unadjusted pairwise tests to ask if Cruz is really the runner-up upon verifying that Trump is the best, and if Rubio

is really the second runner-up upon verifying that Cruz is the runner-up, etc., until we can no longer infer that a certain population really holds its rank.

While we aim to declare more populations to be in the correct order, declaring too many populations, that is, out-of-place populations, to be in the right order is undesirable. It is possible to consider false discovery rate (the expected portion of out-of-place populations declared) here, but we restrict our derivation to FWER (the probability of having any out-of-place populations declared).

Formally, let $\hat{j}_0$ denote the number of ranks validated by a procedure (the number of rejections). Then the FWER of $\hat{j}_0$ is the probability that too many rejections are made; that is, $\mathbb{P}[\hat{j}_0 > j_0]$. For example, suppose that the top three data ranks and population ranks coincide, but not the fourth ($j_0 = 3$). Then we will have made a Type I error if we declare that the top five ranks are correct ($\hat{j}_0 = 5$), but not if we declare that the top two are correct ($\hat{j}_0 = 2$). In other words, $\hat{j}_0$ is a lower confidence bound for j_0.

To show that Procedure 3 is valid, we will prove the validity of a more liberal Procedure 3′, described in Algorithm 1. Procedure 3 is equivalent to Procedure 3′ for the most part, except that Procedure 3 conditions on a larger event $\{X_{[j]} \geq \max_{k>j} X_{[k]}\}$ in Line 7.

THEOREM 7. *Procedure 3′ is a stepwise procedure that an estimate $\hat{j}_0$ of j_0 at the FWER controlled at α, where j_0 is given by*

$$j_0 = \max_j \left\{ \theta_{[1]} > \cdots > \theta_{[j]} > \max_{k>j} \theta_{[k]} \right\}.$$

PROOF. We will first show that Procedure 3′ falls into the sequential goodness-of-fit testing framework proposed by Fithian, Taylor and Tibshirani (2015). We thus analyze Procedure 3′ as a special case of the BasicStop procedure on random hypothesis, described in the same paper. This enables us to construct valid selective p-values and derive Procedure 3′.

Application of the sequential goodness-of-fit testing framework. Upon observing $X_{[1]} \geq \cdots \geq X_{[n]}$, we can set up a sequence of nested models

$$\mathcal{M}_1(X) \subseteq \cdots \subseteq \mathcal{M}_n(X) \qquad \text{where } \mathcal{M}_j(X)^c = \left\{ \theta : \theta_{[1]} > \cdots > \theta_{[j]} > \max_{k>j} \theta_{[k]} \right\}.$$

If we define the jth null hypothesis as

$$\widetilde{H}_{0j} : \theta_{[j]} \leq \max_{k>j} \theta_{[k]},$$

then $\widetilde{H}_{01}, \ldots, \widetilde{H}_{0j}$ are all false if and only if $\theta \notin \mathcal{M}_j(X)$.

In other words, $\mathcal{M}_j(X)$ is a family of distributions that does not have all first j ranks correct. As we will see later, each step in Procedure 3′ is similar to testing

ALGORITHM 1. Procedure 3′, a more liberal version of Procedure 3

input : $X_1, \ldots, X_n$
output: $\hat{j}_0$, an estimate for j_0
`# Initialization`
$\tau_j \leftarrow [j]$;
`# Consider` τ_j `as part of the observation and the fixed realization of the random index` $[j]$
$X_{\tau_0} \leftarrow \infty$;
$j \leftarrow 0$;
rejected $\leftarrow$ **true**;
while rejected **do**
  $j \leftarrow j + 1$;
  $D_{\tau_j \tau_{j+1}} \leftarrow X_{\tau_j} - X_{\tau_{j+1}}$;
  Set up the distribution of $D_{\tau_j \tau_{j+1}}$, conditioned on
    - the variables $X_{\tau_1}, \ldots, X_{\tau_{j-1}}, X_{\tau_{j+2}}, \ldots, X_{\tau_n}$, and
    - the event $\{X_{\tau_{j-1}} \geq X_{\tau_j} \geq \max_{k>j} X_{\tau_k}\}$;

  `# The distribution of` $D_{\tau_j \tau_{j+1}}$ `depends only on` $\theta_{\tau_j} - \theta_{\tau_{j+1}}$ `now`
  test $H_0 : \theta_{\tau_j} - \theta_{\tau_{j+1}} \leq 0$ against $H_1 : \theta_{\tau_j} - \theta_{\tau_{j+1}} > 0$ according to the distribution of $D_{\tau_j \tau_{j+1}}$;
  Set rejected as the output of the test;
end
$\hat{j}_0 \leftarrow j - 1$;

$\widetilde{H}_{0j}$, stating that without the first j ranks correct, it is hard to explain the observations. Thus, returning $\hat{j}_0 = j$ amounts to rejecting $\widetilde{H}_{01}, \ldots, \widetilde{H}_{0j}$, or equivalently determining that the models $\mathcal{M}_1(X), \ldots, \mathcal{M}_j(X)$ do *not* fit the data.

While the null hypotheses $\widetilde{H}_{0j}$ provided intuition in the setting up the nested models, they are rather cumbersome to work with. Inspired by Fithian, Taylor and Tibshirani (2015), we will instead consider another sequence of random hypothesis that are more closely related to the nest models,

$$H_{0j} : \theta \in \mathcal{M}_j(X),$$

or equivalently, that $\theta_{[1]}, \ldots, \theta_{[j]}$ are not the best j parameters in order.

Adapting this notation, the FWER can be viewed as $\mathbb{P}[\text{reject } H_{0(j_0+1)}]$.

Special case of the BasicStop procedure. While impractical, Procedure 3′ can be thought of as performing all n tests first, producing a sequence of p-values p_j,

and returning

$$\hat{j}_0 = \min\{j : p_j > \alpha\} - 1. \tag{9}$$

This is a special case of the BasicStop procedure. Instead of simply checking that Procedure 3′ fits all the requirement for FWER control in BasicStop, we will give the construction of Procedure 3′, assuming that we are to estimate j_0 with BasicStop.

In general, the FWER for BasicStop can be rewritten as $\mathbb{P}[p_{j_0+1} \le \alpha]$. This is however difficult to analyze, as j_0 itself is random and dependent on X, thus we break the FWER down as follows:

$$\begin{aligned}\mathbb{P}[p_{j_0+1} \le \alpha] &= \sum_j \mathbb{P}[p_{j_0+1} \le \alpha \mid j_0 = j]\mathbb{P}[j_0 = j] \\ &= \sum_j \mathbb{P}[p_{j+1} \le \alpha \mid j_0 = j]\mathbb{P}[j_0 = j] \\ &= \sum_j \mathbb{P}\big[p_{j+1} \le \alpha \mid \theta \in \mathcal{M}_{j+1}(X) \setminus \mathcal{M}_j(X)\big]\mathbb{P}[j_0 = j].\end{aligned}$$

We emphasize here that θ is *not* random, but $\mathcal{M}_{j+1}$ is. Thus it suffices to construct the p-values such that

$$\mathbb{P}\big[p_j \le \alpha \mid \theta \in \mathcal{M}_j(X) \setminus \mathcal{M}_{j-1}(X)\big] \le \alpha \qquad \text{for all } j. \tag{10}$$

Considerations for conditioning. By smoothing, we are free to condition on additional variables in (10). A logical choice that simplified (10) is conditioning on the variables $\mathcal{M}_{j-1}(X)$ and $\mathcal{M}_j(X)$. Note that the choice of the model $\mathcal{M}_j(X)$, once again, based solely on the random indices $[1], \ldots, [j]$, so conditioning on both $\mathcal{M}_{j-1}(X)$ and $\mathcal{M}_j(X)$ is equivalent to conditioning on the random indices $[1], \ldots, [j]$, which in turns is equivalent to conditioning on the σ-field generated by the partition of the observation space X

$$\left\{\left\{X_{\tau_1} \ge \cdots \ge X_{\tau_j} \ge \max_{k>j} X_{\tau_k}\right\} : \tau \text{ is any permutation of } (1, \ldots, n)\right\},$$

or colloquially, the set of all possible choices of $[1], \ldots, [j]$. Within each set in this partition, the event $\{\theta \in \mathcal{M}_j(X) \setminus \mathcal{M}_{j-1}(X)\}$ is simply $\{\theta_{\tau_1} > \cdots > \theta_{\tau_j}$ and $\theta_{\tau_j} \le \max_{k>j} \theta_{\tau_k}\}$, a trivial event.

As a brief summary, we want to construct p-values p_j such that

$$\mathbb{P}_{\substack{\theta_{\tau_1} > \cdots > \theta_{\tau_j} \\ \theta_{\tau_j} \le \max_{k>j} \theta_{\tau_k}}} \left[p_j \le \alpha \mid X_{\tau_1} \ge \cdots \ge X_{\tau_j} \ge \max_{k>j} X_{\tau_k}\right].$$

Construction of the p-values. To avoid the clutter in the subscripts, we will drop the τ in the subscript. Hence our goal is now

$$\mathbb{P}_{\substack{\theta_1>\cdots>\theta_j \\ \theta_j\le\max_{k>j}\theta_k}}\Big[p_j \le \alpha \mid X_1 \ge \cdots \ge X_j \ge \max_{k>j} X_k\Big].$$

Construction of p_j for other permutations τ can be obtained similarly.

There are many valid options for p_j (such as constant α). We will follow the idea in the proof of Part 1 of Theorem 1 here. p_j is intended to test $H_{0j} : \theta \in \mathcal{M}_j(X)$, which is equivalent to the union of the null hypotheses:

1. $\theta_k \le \theta_{k+1}$ for $k = 1, \ldots, j-1$, and
2. $\theta_j \le \theta_k$ for $k = j+1, \ldots, n$. (The union of these null hypotheses is $\widetilde{H}_{0j}$.)

Since the joint distribution of X, restricted to $\{X_1 \ge \cdots \ge X_j \ge \max_{k>j} X_k\}$, remains in the exponential family, we can construct the p-values for each of the hypotheses above by conditioning on the variables corresponding to the nuisance parameters here, similar to the proof of Part 1 of Theorem 1. Then we can take p_j as the maximum of such p-values.

For the hypothesis $H_{0jk} : \theta_j \le \theta_k$, we can construct p_{jk}, by considering the survival function of the conditional law

$$\begin{aligned}&\mathcal{L}_{\theta_j=\theta_k}\Big(D_{jk} \mid \Big\{X_1 \ge \cdots \ge X_j \ge \max_{\ell>j} X_\ell\Big\}, X_{\setminus\{j,k\}}, M_{jk}\Big)\\ &\quad= \mathcal{L}_{\theta_j=\theta_k}\Big(D_{jk} \mid \Big\{X_{j-1} \ge X_j \ge \max_{\substack{\ell>j\\ \ell\ne k}} X_\ell \text{ and } X_j \ge M_{jk}\Big\}, X_{\setminus\{j,k\}}, M_{jk}\Big).\end{aligned}$$

Once again, $X_{j+1} = \max_{\ell>j} X_\ell$ is simply shorthand for simplifying our notation. Now the p-values are similar to the ones in equation (4), for $k > j$:

$$p_{jk} = \frac{\int_{D_{jk}}^{X_{j-1}} g(X_1, \ldots, M_{jk}+z, \ldots, M_{jk}-z, \ldots, X_n)\,dz}{\int_{\max\{X_{j+1}-M_{jk},0\}}^{X_{j-1}} g(X_1, \ldots, M_{jk}+z, \ldots, M_{jk}-z, \ldots, X_n)\,dz}.$$

We can graphically represent p_{jk} in Figure 4, a diagram analogous to Figure 2.

We have $p_{j(j+1)} \ge \max_{k>j} p_{jk}$ by Section 3: the upper truncation for X_j can be represented by cropping Figure 2 along a vertical line, shown in Figure 4. Considering $p_{j(j+1)}$ is sufficient in rejecting all the H_{0jk}. We will take $p_{j*} = p_{j(j+1)}$, noting that this is the p-value that Procedure 3′ would produce. In fact, p_{j*} is also the p-value we would have constructed if we were to reject only $\widetilde{H}_{0j}$.

Upon constructing p_j, one should realize that the p-values for testing $\theta_k \le \theta_{k+1}$ would have been constructed in earlier iterations of BasicStop, as p_{k*}. In other words, $p_j = \max_{k\le j} p_{k*}$ is the sequence of p-values that works with BasicStop. However, from (9),

$$\hat{j}_0 = \min\Big\{j : \max_{k\le j} p_{k*} > \alpha\Big\} - 1 = \min\{j : p_{j*} > \alpha\} - 1,$$

so it is safe to apply BasicStop to p_{j*} directly, yielding Procedure 3′. □

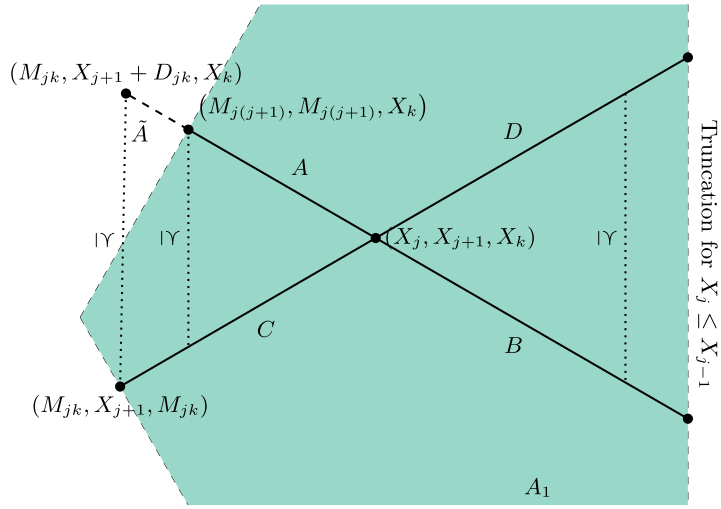

FIG. 4. *The two p-values constructed corresponds to taking integrals of g along these segments that lie on a level set of* $x_j + x_{j+1} + x_k$. *The dashed line corresponds to extension in* (5). *The dotted line on the far right is the truncation that enforces* $X_j < X_{j-1}$.

PART 3 OF THEOREM 1. *Assume the model* (1) *holds and* $g(x)$ *is a Schur-concave function. Procedure* 3 *is a conservative stepwise procedure with FWER no larger than* α.

PROOF. The p-values $p_{j(j+1)}$ obtained in Procedure 3′ are always smaller than their counterpart in Procedure 3, as the upper truncation at X_{j-1} is on the upper tail. Therefore, Procedure 3 is conservative and definitely valid. □

6. Discussion. Combining ideas from conditional inference and multiple testing, we have proven the validity of several very simple and seemingly "naive" procedures for significance testing of sample ranks. In particular, we have shown that an unadjusted pairwise test comparing the winner with the runner-up is a valid significance test for the first rank. Our result complements and extends preexisting analogous results for location and location-scale families with independence between observations. Our approach is considerably more powerful than previously known solutions. We provide similarly straightforward conservative methods for producing a lower confidence bound for the difference between the winner and runner up, and for verifying ranks beyond the first.

Claims reporting the "winner" are commonly made in the scientific literature, usually with no significance level reported or an incorrect method applied. For example, Uhls and Greenfield (2012) asked $n = 20$ elementary and middle school students which of seven personal values they most hoped to embody as adults, with "Fame" (8 responses) being the most commonly selected, with "Benevolence" (5

responses) second. The authors' main finding—which appeared in the abstract, the first paragraph of the article, and later a CNN.com headline [Alikhani (2011)]—was that "Fame" was the most likely response, accompanied by a significance level of 0.006, which the authors computed by testing whether the probability of selecting "Fame" was larger than $1/7$. The obvious error in the authors' reasoning could have been avoided if they had performed an equally straightforward two-tailed binomial test of "Fame" versus "Benevolence," which would have produced a p-value of 0.58.

Reproducibility. A git repository containing with the code generating the image in this paper is available at https://github.com/kenhungkk/verifying-winner.

Acknowledgment. We thank Jason C. Hsu for helpful discussions.

REFERENCES

ALIKHANI, L. (2011). Study: Tween TV today is all about fame. Available at http://thechart.blogs.cnn.com/2011/08/05/study-tweens-aim-for-fame-above-all-else/.

BERGER, R. L. (1980). Minimax subset selection for the multinomial distribution. *J. Statist. Plann. Inference* **4** 391–402. MR0596773

BERGER, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24** 295–300. MR0687187

BESAG, J. and CLIFFORD, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* **76** 633–642.

BOFINGER, E. (1991). Selecting "demonstrably best" or "demonstrably worst" exponential populations. *Aust. N. Z. J. Stat.* **33** 183–190. MR1144383

EDWARDS, D. G. and HSU, J. C. (1983). Multiple comparisons with the best treatment. *J. Amer. Statist. Assoc.* **78** 965–971.

FINNER, H. and STRASSBURGER, K. (2002). The partitioning principle: A powerful tool in multiple decision theory. *Ann. Statist.* **30** 1194–1213. MR1926174

FITHIAN, W., SUN, D. L. and TAYLOR, J. E. (2014). Optimal inference after model selection. Preprint. Available at arXiv:1410.2597.

FITHIAN, W., TAYLOR, J. E. and TIBSHIRANI, R. J. (2015). Selective sequential model selection. Preprint. Available at arXiv:1512.02565.

GUPTA, S. S., HUANG, D.-Y. and PANCHAPAKESAN, S. (1984). On some inequalities and monotonicity results in selection and ranking theory. In *Inequalities in Statistics and Probability* (*Lincoln, Neb.*, 1982) 211–227. IMS, Hayward, CA.

GUPTA, S. S. and LIANG, T. (1989). Selecting the best binomial population: Parametric empirical Bayes approach. *J. Statist. Plann. Inference* **23** 21–31. MR1029237

GUPTA, S. S. and NAGEL, K. (1967). On selection and ranking procedures and order statistics from the multinomial distribution. *Sankhyā* **29** 1–34.

GUPTA, S. S. and PANCHAPAKESAN, S. (1971). On multiple decision (subset selection) procedures. Technical report, Purdue Univ., West Lafayette, IN.

GUPTA, S. S. and PANCHAPAKESAN, S. (1985). Subset selection procedures: Review and assessment. *Amer. J. Math. Management Sci.* **5** 235–311. MR0859941

GUPTA, S. S. and WONG, W.-Y. (1976). On subset selection procedures for Poisson processes and some applications to the binomial and multinomial problems. Technical report.

GUTMANN, S. and MAYMIN, Z. (1987). Is the selected population the best? *Ann. Statist.* **15** 456–461. MR0885752

HSU, J. C. (1984). Constrained simultaneous confidence intervals for multiple comparisons with the best. *Ann. Statist.* **12** 1136–1144. MR0751303

HSU, J. (1996). *Multiple Comparisons*: *Theory and Methods*. CRC Press, Boca Raton, FL.

KANNAN, N. and PANCHAPAKESAN, S. (2009). Does the selected normal population have the smallest variance? *Amer. J. Math. Management Sci.* **29** 109–123. MR2751256

MARSHALL, A. W., OLKIN, I. and ARNOLD, B. C. (2011). *Inequalities*: *Theory of Majorization and Its Applications*, 2nd ed. Springer, New York. MR2759813

MAYMIN, Z. and GUTMANN, S. (1992). Testing retrospective hypotheses. *Canad. J. Statist.* **20** 335–345. MR1190577

NETTLETON, D. (2009). Testing for the supremacy of a multinomial cell probability. *J. Amer. Statist. Assoc.* **104** 1052–1059. MR2750236

NG, H. K. T. and PANCHAPAKESAN, S. (2007). Is the selected multinomial cell the best?. *Sequential Anal.* **26** 415–423. MR2359863

QUINNIPIAC UNIVERSITY POLL INSTITUTE (2016). First-timers put Trump ahead in Iowa GOP caucus, Quinnipiac University poll finds; Sanders needs first-timers to tie Clinton in Dem caucus. Available at https://poll.qu.edu/images/polling/ia/ia02012016_Ifsmb28.pdf/.

STEFANSSON, G., KIM, W.-C. and HSU, J. C. (1988). On confidence sets in multiple comparisons. In *Statistical Decision Theory and Related Topics IV* **2** 89–104. Springer, New York. MR0927125

UHLS, Y. T. and GREENFIELD, P. M. (2012). The value of fame: Preadolescent perceptions of popular media and their relationship to future aspirations. *Dev. Psychol.* **48** 315–326.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, BERKELEY
951 EVANS HALL, SUITE 3840
BERKELEY, CALIFORNIA 94720-3840
USA
E-MAIL: kenhung@berkeley.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
301 EVANS HALL
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: wfithian@berkeley.edu