Journal of Econometrics 250 (2025) 106010

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Finite- and large-sample inference for ranks using multinomial data with an application to ranking political parties☆

Sergei Bazylik [a], Magne Mogstad [a,b,c], Joseph P. Romano [d], Azeem M. Shaikh [a], Daniel Wilhelm [e,*]

[a] Department of Economics, University of Chicago, United States of America
[b] Statistics Norway, Norway
[c] NBER, United States of America
[d] Departments of Statistics and Economics, Stanford University, United States of America
[e] Departments of Statistics and Economics, LMU Munich, Germany

ARTICLE INFO

JEL classification:
C12
C14
D31
I20
J62

Keywords:
Confidence sets
Multinomial data
Multiple testing
Polls
Ranks
Surveys

ABSTRACT

It is common to rank different categories by means of preferences that are revealed through data on choices. A prominent example is the ranking of political candidates or parties using the estimated share of support each one receives in surveys or polls about political attitudes. Since these rankings are computed using estimates of the share of support rather than the true share of support, there may be considerable uncertainty concerning the true ranking of the political candidates or parties. In this paper, we consider the problem of accounting for such uncertainty by constructing confidence sets for the rank of each category. We consider both the problem of constructing marginal confidence sets for the rank of a particular category as well as simultaneous confidence sets for the ranks of all categories. A distinguishing feature of our analysis is that we exploit the multinomial structure of the data to develop confidence sets that are valid in finite samples. We additionally develop confidence sets using the bootstrap that are valid only approximately in large samples. We use our methodology to rank political parties in Australia using data from the 2019 Australian Election Survey. We find that our finite-sample confidence sets are informative across the entire ranking of political parties, even in Australian territories with few survey respondents and/or with parties that are chosen by only a small share of the survey respondents. In contrast, the bootstrap-based confidence sets may sometimes be considerably less informative. These findings motivate us to compare these methods in an empirically-driven simulation study, in which we conclude that our finite-sample confidence sets often perform better than their large-sample, bootstrap-based counterparts, especially in settings that resemble our empirical application.

1. Introduction

Preferences over different categories are often assessed by means of data on choices. It is natural to summarize this type of data by ranking the different categories according to the share of support each one receives in the data. A prominent example is provided by

☆ The third author acknowledges support from the National Science Foundation, United States (MMS-1949845). The fourth author acknowledges support from the National Science Foundation, United States (SES-1530661). The fifth author acknowledges support from the ESRC Centre for Microdata Methods and Practice at IFS, United States (RES-589-28-0001) and the European Research Council (Starting Grant No. 852332).

* Corresponding author.
E-mail address: d.wilhelm@lmu.de (D. Wilhelm).

https://doi.org/10.1016/j.jeconom.2025.106010
Received 20 December 2022; Received in revised form 31 January 2024; Accepted 24 May 2024
Available online 13 May 2025

surveys or polls of support for political candidates. In this case, individuals choose one political candidate or party from among those available. The resulting rankings may shape public discussion, inform campaigns, and be used as inputs into consequential decisions before the actual election. For example, in U.S. Presidential Elections, the decision about which candidates to feature in nationally televised political debates may hinge on their performance in different polls leading up to the election. Another prominent example is choice-based conjoint analysis, in which respondents select which of several options they would purchase or otherwise choose if given the option. Such analyses are regularly used in marketing analysis both to assess which product features are most valued and thereby inform decisions about which products to introduce. In both these examples, however, it is important to acknowledge the uncertainty surrounding these rankings.

Such data on choices, including polls of political attitudes, commonly feature limited sample sizes and/or categories whose true share of support is small. As explained further below, these features pose challenges to inference methods justified using large-sample arguments. In contrast, this paper considers the problem of constructing confidence sets for the rank of each category that are valid in finite samples, even when some categories are chosen with probability close to zero. We consider two types of confidence sets: marginal confidence sets for the rank of a particular category, by which we mean random sets that contain the rank of a particular category with probability no less than some pre-specified level, as well as simultaneous confidence sets for the ranks of all categories, by which we mean random sets that contain the ranks of all categories with probability no less than some pre-specified level. The former confidence sets provide a way of accounting for uncertainty when answering questions pertaining to the rank of a particular category, whereas the latter confidence sets provide a way of accounting for uncertainty when answering questions pertaining to the ranks of all categories. Our constructions are based off of testing a family of one-sided null hypotheses concerning differences in pairs of success probabilities in a way that controls the familywise error rate in finite samples. In order to do so, we exploit the multinomial structure of the data, which enables the use of a simple conditioning argument.

As a second contribution, we develop bootstrap methods for the construction of these confidence sets. Their validity is justified using large-sample arguments as in Mogstad et al. (2024). However, unlike in Mogstad et al. (2024)'s applications, the estimators of the success probabilities of different categories are necessarily dependent and the bootstrap procedure proposed in this paper explicitly accounts for this dependence. As described in more detail below, the results in Brown et al. (2001) suggest that such bootstrap methods may perform poorly when sample sizes are small and/or some categories are chosen with small probabilities. In particular, approaches that explicitly or implicitly (such as the bootstrap) rely on asymptotic normality of the estimators of the success probabilities perform poorly when the true success probability is small. In such a case, it is well known that the Poisson distribution is in fact a better approximation than the normal. In our simulations, we find not only that the bootstrap-based confidence sets (with or without studentization) may have coverage probability considerably below the desired level, but also that they may be excessively wide. In contrast, the finite-sample confidence sets have coverage probability no less than the desired nominal level and may even be considerably shorter.

We apply our inference procedures to re-examine the ranking of political parties in Australia using data from the 2019 Australian Election Survey. We find that the finite-sample (marginal and simultaneous) confidence sets are remarkably informative across the entire ranking of political parties, even in Australian territories with few survey respondents and/or with parties that are chosen by only a small share of the survey respondents. To illustrate this point further, consider one particular Australian territory, Greater Melbourne. We find that the finite-sample confidence sets are either of similar length to or substantially shorter than their bootstrap-based counterparts (with or without studentization). For instance, at conventional significance levels, the finite-sample marginal confidence set for the rank of the Green Party contains only rank 4. In contrast, the bootstrap-based marginal confidence sets (with or without studentization) contain the ranks 3 to 7, thus exhibiting substantially more uncertainty about the true rank of the Green Party. The studentized procedure leads to especially wide confidence sets for the ranks of parties that are chosen only by a small share of respondents. We find similar patterns in the eight most populous territories, while confidence sets in the remaining seven least populous territories are uninformative due to very small sample sizes. Unlike in Greater Melbourne, however, in some other territories bootstrap-based confidence sets (with or without studentization) may be slightly smaller than their finite-sample counterparts.

These findings motivate us to compare the different confidence sets in a simulation study modeled after our empirical application. The findings of this exercise can be summarized as follows. First, finite-sample marginal confidence sets have coverage probabilities no less than the desired nominal level in all simulation designs, including those with very small sample sizes and/or with parties that are chosen by only a small share of the respondents. Second, bootstrap-based confidence sets without studentization also have coverage probabilities no less than the nominal level, except when sample sizes are very small. In contrast, bootstrap-based confidence sets with studentization may have coverage probabilities less than the nominal level when sample sizes are small and/or the number of parties to be ranked is not too small. Third, the finite-sample confidence sets may produce considerably shorter confidence sets for the ranks compared to the bootstrap-based ones, especially when there are parties that are chosen by only a small share of respondents. However, there are also situations in which the latter methods produce shorter confidence sets than the former, so neither approach always dominates the other in terms of length of their confidence sets.

Our paper is most closely related to the aforementioned paper by Mogstad et al. (2024). We emphasize that the primary contribution of our analysis relative to theirs is to show how one may exploit additional structure given by the multinomial data to construct confidence sets that enjoy finite-sample validity. Importantly, the finite sample guarantee allows for data-generating processes with success probabilities that are arbitrarily close to zero, whereas the asymptotic arguments justifying the bootstrap require these probabilities to be bounded away from zero. Second, we propose a bootstrap method that accounts for the dependence in the estimators of the multinomial success probabilities. Our paper is also related to a recent paper by Klein et al. (2020), who consider the problem of constructing confidence sets analogous to those in Mogstad et al. (2024). We show how a modification of

their procedure can also be used to construct confidence sets that are valid in finite samples in the presence of multinomial data. In our simulations, we find that the resulting confidence sets are often of comparable length to our finite-sample confidence sets, but sometimes meaningfully larger. We refer the reader to Mogstad et al. (2024) for additional comparisons. Other related work includes Goldstein and Spiegelhalter (1996), who propose a different bootstrap-based confidence set to account for uncertainty in reported ranks. As explained by Hall and Miller (2009), Xie et al. (2009) and most recently by Mogstad et al. (2024), however, this method performs poorly in the presence of categories that are chosen with similar frequencies (i.e., in the context of our simulations, when some parties are nearly tied). We confirm this finding in our simulations.

Our paper also draws motivation from earlier work by Brown et al. (2001), who demonstrate that conventional confidence intervals for the probability of success using binomial data may behave poorly in the sense of exhibiting undercoverage, especially when the success probability is close to zero or one, and may also behave erratically in the sense that coverage probabilities may be volatile and non-monotonic in the sample size. In our simulations, we find similar patterns concerning the coverage probabilities for the differences in pairs of success probabilities using multinomial data. For this reason, we view insistence upon finite-sample validity for our confidence sets to be especially compelling in this setting. On the other hand, we find that this poor behavior of confidence sets for the differences in the success probabilities need not translate into similar behavior for the implied confidence sets for the ranks.

We note some key differences between the problems considered in this paper and those of two recent papers in econometrics, Andrews et al. (2018) and Gu and Koenker (2020). In the context of the multinomial setting studied here, Andrews et al. (2018) develop methods for inference on the true success probability for the randomly selected category whose estimated rank is highest. In contrast, as the discussion above makes clear, we develop methods for inference on the true ranks themselves. Gu and Koenker (2020) develop decision rules for selecting the most popular categories (i.e., those with the highest success probabilities), which is more closely related to a literature on subset selection (see Gupta and Panchapakesan, 1979 for a review). We show, however, how our simultaneous confidence sets may be used to create a complimentary object that we refer to as the confidence set for the τ-best. For given value of τ, such a confidence set is a random set that contains the identities of (all of) the categories whose rank is less than or equal to τ with probability approximately no less than some pre-specified level.

Finally, there is also a large literature about the analysis of datasets that contain observations of elicited rankings (e.g., Marden, 1995), but this differs from our setting in which we assume only an individual's top choice is observed rather than their ranking of all available options.

The remainder of the paper is organized as follows. In Section 2.1, we introduce our general setup, including a formal description of the different types of confidence sets we consider and the general testing problem involved in their constructions. Suitable tests that lead to confidence sets that are valid in finite samples are then described in 2.2. The construction of confidence sets for the τ-best that are valid in finite samples is briefly summarized in 2.3. Section 3 described bootstrap-based versions of these same confidence sets. In Section 4.1, we apply our inference procedures to re-examine the ranking of political parties in Australia. Finally, in Section 5, we examine the finite-sample behavior of our inference procedures via a simulation study modeled after our empirical application.

2. Main results

2.1. Setup and notation

Let $j \in J \equiv \{1, \ldots, p\}$ index categories of interest, e.g., parties in an election. There are n independent observations, and each observation falls in category j with probability θ_j. Let X_j denote the observed count for category j from the n observations, e.g. the number of votes party j receives from n voters. Hence, $X \equiv (X_1, \ldots, X_p)'$ is distributed according to the multinomial distribution with parameters n and $\theta \equiv (\theta_1, \ldots, \theta_p)'$.

The rank of category j is defined as

$$r_j \equiv 1 + \sum_{k \in J} \mathbb{1}\{\theta_k > \theta_j\},$$

where $\mathbb{1}\{A\}$ is equal to one if the event A holds and equal to zero otherwise. Let $r \equiv (r_1, \ldots, r_p)'$. Before proceeding, it is useful to provide a simple example to illustrate the way in which ties are handled with this definition of ranks: if $\theta = (0.4, 0.1, 0.1, 0.2, 0.2)'$, then $r = (1, 4, 4, 2, 2)'$.

The primary goal is to construct confidence sets for the rank of a particular category or for the ranks of multiple categories simultaneously. Let $J_0 \subseteq J$ denote the categories of interest. For a given value of $\alpha \in (0, 1)$, we use data X to construct (random) sets

$$R_n \equiv \prod_{j \in J_0} R_{n,j},$$

where the (random) sets $R_{n,j}$, $j \in J_0$, are such that

$$P\left\{r_j \in R_{n,j} \ \forall j \in J_0\right\} \geq 1 - \alpha. \tag{1}$$

If J_0 is a singleton, then sets R_n satisfying (1) are referred to as *marginal confidence sets for the rank of a single category*. If $J_0 = J$, then sets R_n satisfying (1) are referred to as *simultaneous confidence sets for the ranks of all categories*. The remainder of the paper, however, allows J_0 to be any subset of J. In our constructions, $R_{n,j}$ are subsets of J for each $j \in J_0$, allowing for the possibility that the lower endpoint is 1 or the upper endpoint is p to permit both one-sided and two-sided inference.

In addition, we consider the goal of constructing confidence sets for the identities of all categories whose rank is less than or equal to a pre-specified value $\tau \in J$, i.e., for a given value of $\alpha \in (0,1)$, we construct (random) sets $R_n^{\tau-\text{best}}$ that are subsets of J and satisfy

$$P\left\{R_0^{\tau-\text{best}} \subseteq R_n^{\tau-\text{best}}\right\} \geq 1-\alpha\ , \tag{2}$$

where

$$R_0^{\tau-\text{best}} \equiv \{j \in J\ :\ r_j \leq \tau\}\ .$$

Sets satisfying (2) are referred to as *confidence sets for the τ-best categories*.

As in Mogstad et al. (2024), the construction of confidence sets for ranks can be based on tests of the hypotheses

$$H_{j,k} : \theta_j \leq \theta_k$$

for pairs of indices $(j,k) \in J^2$. Which pairs are relevant depends on whether the desired confidence sets for the ranks indicated by J_0 are lower, upper or two-sided confidence bounds:

$$\begin{aligned} \boldsymbol{J}^{\text{lower}} &\equiv \{(j,k) \in J \times J_0 : j \neq k\} \\ \boldsymbol{J}^{\text{upper}} &\equiv \{(j,k) \in J_0 \times J : j \neq k\} \\ \boldsymbol{J}^{\text{two-sided}} &\equiv \boldsymbol{J}^{\text{lower}} \cup \boldsymbol{J}^{\text{upper}} \end{aligned}$$

Suppose a family of tests of the hypotheses $H_{j,k}$ is given. Then, for each $j \in J_0$, let

$$\text{Rej}_j^- \equiv \{k \in J \setminus \{j\} :\ \text{reject } H_{k,j} \text{ and claim } \theta_j < \theta_k\} \tag{3}$$

indicate the set of hypotheses that are rejected in favor of $\theta_j < \theta_k$ and

$$\text{Rej}_j^+ \equiv \{k \in J \setminus \{j\} :\ \text{reject } H_{j,k} \text{ and claim } \theta_j > \theta_k\} \tag{4}$$

the set of hypotheses that are rejected in favor of $\theta_j > \theta_k$. Consider the goal of constructing a two-sided marginal confidence set for the rank of a category j_0, i.e., $J_0 = \{j_0\}$. Then, $\text{Rej}_{j_0}^-$ contains all categories $k \neq j_0$ whose parameter θ_k is claimed to be strictly larger than θ_{j_0}. If these claims were correct, then the lower bound on the rank of category j_0 must be equal to the number of such categories k, denoted by $|\text{Rej}_{j_0}^-|$, plus one. Similarly, $\text{Rej}_{j_0}^+$ contains all categories $k \neq j_0$ whose parameter θ_k is claimed to be strictly smaller than θ_{j_0}. Again, if these claims were correct, then the upper bound on the rank of category j_0 must be the total number of categories, p, minus the number of categories with smaller probability, denoted by $|\text{Rej}_{j_0}^+|$. Therefore, if all claims made in $\text{Rej}_{j_0}^-$ and $\text{Rej}_{j_0}^+$ are correct, the set

$$R_{n,j_0} \equiv \left\{|\text{Rej}_{j_0}^-| + 1, \ldots, p - |\text{Rej}_{j_0}^+|\right\}.$$

contains the rank of category j_0, r_{j_0}. More generally, for an arbitrary set of indices $J_0 \subseteq J$, if all claims made in Rej_j^- and Rej_j^+, $j \in J_0$, are correct, then the set

$$R_n \equiv \prod_{j \in J_0} R_{n,j} \qquad \text{with} \qquad R_{n,j} \equiv \left\{|\text{Rej}_j^-| + 1, \ldots, p - |\text{Rej}_j^+|\right\} \tag{5}$$

contains the ranks of all categories in J_0, $(r_j : j \in J_0)$. Of course, it cannot be guaranteed that tests of $H_{j,k}$ never falsely reject, but the probability of such mistakes can be controlled. More specifically, for the set R_n to satisfy the coverage statement in (1), the number of false claims must be controlled in the sense that the familywise error rate for testing $H_{j,k}$ for the relevant pairs of indices $I \subset J^2$ is no larger than α, i.e.,

$$FWER_I \equiv P\left\{\text{reject at least one true hypothesis } H_{j,k},\ (j,k) \in I\right\} \leq \alpha. \tag{6}$$

For two-sided confidence sets, the relevant set of indices I is $J^{\text{two-sided}}$ and, for one-sided confidence sets, it is either J^{lower} or J^{upper}. The following theorem is a slight generalization (allowing for a general set J_0 of indices) of Theorem 3.4 in Mogstad et al. (2024) and summarizes the above discussion.

Theorem 2.1. *For $J_0 \subseteq J$, let I be equal to J^{lower}, J^{upper}, or $J^{\text{two-sided}}$. Let R_n be defined by (3), (4), and (5), where the family of hypotheses $H_{j,k}$, $(j,k) \in I$, is tested using a procedure that satisfies (6) for some $\alpha \in (0,1)$. Then, R_n satisfies (1).*

Instead of controlling the coverage probability in finite samples as in (1), the confidence sets proposed in Mogstad et al. (2024) only asymptotically control the coverage probability. Their constructions assume the availability of bootstrap confidence sets that simultaneously cover the differences $\theta_j - \theta_k$ for all relevant pairs of indices (j,k). While their paper does not explicitly show how these can be constructed when $X_1, \ldots, X_p$ are not independent (which by construction is the case with multinomial data), it is not difficult to propose an appropriate bootstrap procedure (see Appendix A).

The general approach in Mogstad et al. (2024) does not require X to follow a multinomial distribution. The purpose of the remainder of this paper is to show that with this additional distributional assumption it is possible to construct confidence sets for ranks that control the coverage probability not only asymptotically, but in finite samples.

Remark 2.1 (*Definition of Rank*)**.** To simplify the exposition in this remark, suppose we are interested in a single category, $J_0 = \{j_0\}$. In the presence of ties, the rank of a category can be defined in different ways. For any $j \in J$, let $\underline{r}_j \equiv 1 + \sum_{k \in J} \mathbb{1}\{\theta_k > \theta_j\}$ and $\bar{r}_j \equiv p - \sum_{k \in J} \mathbb{1}\{\theta_k < \theta_j\}$ be the smallest (i.e., best) and largest (i.e., worst) possible rank of category j. If category j_0 is not tied with any other category, then $\underline{r}_{j_0} = \bar{r}_{j_0}$ and the rank is unique. On the other hand, when category j_0 is tied with at least one other category, then $\underline{r}_{j_0} < \bar{r}_{j_0}$ and different definitions of the rank may select different values from the interval $R_{j_0} \equiv [\underline{r}_{j_0}, \bar{r}_{j_0}]$. An inspection of the proof of Theorem 2.1 reveals that the confidence set R_n not only covers our definition of the rank, r_{j_0}, in the sense of (1), but also any other "reasonable" definition of the rank in the sense that

$$P\left\{ R_{j_0} \subseteq R_n^{\text{cont}} \right\} \geq 1 - \alpha,$$

where $R_n^{\text{cont}} \equiv [\min(R_n), \max(R_n)]$ is the interval from the smallest to the largest value in the confidence set R_n. □

2.2. Marginal and simultaneous confidence sets for ranks

In light of the previous discussion, for the construction of a confidence set satisfying (1), it remains to propose a procedure for testing the family of hypotheses $H_{j,k}$, $(j,k) \in I$, that controls $FWER_I$. In this section, we propose a test of the individual hypothesis $H_{j,k}$ with nominal level $\beta_{j,k}$ and then choose the constants $(\beta_{j,k} : (j,k) \in I)$ in a way that controls $FWER_I$ in the sense of (6).

Let $S_{j,k} \equiv X_j + X_k$. One can show that the conditional distribution of X_j given $S_{j,k} = s$ is binomial based on s trials and success probability $\theta_j/(\theta_j + \theta_k)$; a simple proof appears in the proof of Theorem 2.2. Notice that $H_{j,k}$ is equivalent to $\theta_{j,k} \leq 1/2$, where $\theta_{j,k} = \theta_j/(\theta_j + \theta_k)$. Conditioning on $S_{j,k}$ eliminates nuisance parameters and reduces the testing problem to a one-parameter problem of testing a binomial probability. An exact level $\beta_{j,k}$ test may be therefore be easily constructed. In particular, the (possibly randomized) test of $H_{j,k}$ defined by the critical function

$$\phi(x,s) = \begin{cases} 1, & \text{if } x > C(s) \\ \gamma(s), & \text{if } x = C(s) \\ 0, & \text{if } x < C(s) \end{cases} \tag{7}$$

with constants $\gamma(s)$, $C(s)$ determined by

$$\sum_{i=C(s)+1}^{s} \binom{s}{i} (1/2)^s + \gamma(s) \binom{s}{C(s)} (1/2)^s = \beta_{j,k} \qquad \forall s, \tag{8}$$

is an exact level $\beta_{j,k}$ test of $H_{j,k}$. The test has rejection probability equal to $\beta_{j,k}$ when $\theta_{j,k} = 1/2$. Moreover, since the binomial family of distributions has monotone likelihood ratio, the test has rejection probability strictly less than $\beta_{j,k}$ whenever $\theta_{j,k} < 1/2$. In fact, Theorem 2.2 below shows that, for testing $H_{j,k}$, the test $\phi(X_j, S_{j,k})$ is uniformly most powerful level $\beta_{j,k}$ among all level $\beta_{j,k}$ unbiased tests based on $(X_1, \ldots, X_p)$; i.e., it is UMPU. If $\gamma(s) > 0$ and one wishes to avoid randomization of the test, then one may simply reject $H_{j,k}$ iff $X_j > C(S_{j,k}) + 1$. The p-value for this slightly conservative approach of testing $H_{j,k}$ when $S_{j,k} = s$ can be written as

$$\hat{p}_{j,k} \equiv \frac{1}{2^s} \sum_{i=X_j}^{s} \binom{s}{i}. \tag{9}$$

The following theorem summarizes the above discussion.

Theorem 2.2. *For any $(j,k) \in J^2$, $j \neq k$, and $\beta_{j,k} \in (0,1)$, the test $\phi(X_j, S_{j,k})$ defined by (7) and (8) is a UMPU level $\beta_{j,k}$ test of $H_{j,k}$.*

This theorem shows that ϕ defines a level $\beta_{j,k}$ test of $H_{j,k}$. To satisfy (6) one could combine the individual tests, i.e., choose the $(\beta_{j,k} : (j,k) \in I)$, by a Bonferroni correction or by the Holm (1979) procedure, for example. Theorem 2.1 then implies that the confidence set R_n, based on such a procedure, has coverage probability no less than $1 - \alpha$.

The steps necessary for construction of the proposed confidence sets for the ranks, using the non-randomized test with p-value in (9), are summarized as follows:

Algorithm 2.1.

1. Choose the set $J_0 \subseteq J$ of categories of interest.
2. Set I equal to one of J^{lower}, J^{upper}, or $J^{\text{two-sided}}$, depending on whether lower, upper, or two-sided confidence bounds on the ranks of categories in J_0 are desired.
3. Test the family of hypotheses $H_{j,k}$, $(j,k) \in I$, so that $FWER_I$ is controlled. For instance:

   - **Bonferroni:** $H_{j,k}$ is rejected iff

   $$\hat{p}_{j,k} \leq \frac{\alpha}{|I|}.$$

   - **Holm:** order the p-values $\hat{p}_{j,k}$, $(j,k) \in I$, from the smallest to the largest, $\hat{p}_{(1)} \leq \cdots \leq \hat{p}_{(|I|)}$, and denote the corresponding hypotheses by $H_{(1)}, \ldots, H_{(|I|)}$. Then, $H_{(l)}$ is rejected iff

$$\hat{p}_{(l')} \leq \frac{\alpha}{|I| + 1 - l'} \quad \forall l' \leq l.$$

4. For each $j \in J_0$, collect the rejected hypotheses as in (3) and (4).
5. Construct R_n, the confidence set for the ranks $(r_j : j \in J_0)$, as in (5).

Since the Holm procedure rejects at least as many hypotheses as Bonferroni (with probability one) and thus leads to confidence sets that are at least as short as those based on Bonferroni, it is to be preferred. However, in simulations (Section 5), we find that the two methods lead to almost identical confidence sets, and both methods are optimal in many high-dimensional settings (Lehmann and Romano, 2022, Chapter 13.5).

Theorems 2.1 and 2.2 imply that the resulting confidence set for the ranks is valid in finite samples:

Corollary 2.1. *The confidence set R_n constructed by Algorithm 2.1 satisfies (1).*

One important aspect to note is that we have not imposed any assumptions on the vector of probabilities θ, besides it being the vector of probabilities associated with a multinomial distribution for p categories. In particular, the confidence set R_n satisfies the coverage result (1) regardless of whether any of the elements of θ are equal to each other ("ties") or close to each other ("near-ties"). This is an important feature of our confidence sets for the ranks because it ensures that the coverage property does not break down when some categories are observed with equal or similar counts. In contrast, a "naive" bootstrap confidence set is valid only in the absence of ties and may substantially under-cover when there are near-ties (see Remark 3.6 and the simulations in Section 5).

Remark 2.2 (*Clopper–Pearson*)**.** An alternative approach to the construction of confidence sets for the ranks that are valid in finite samples could be based on Clopper–Pearson intervals for binomial probabilities (Clopper and Pearson, 1934). To see this, note that one could form a Clopper–Pearson interval separately for each element of θ. With a Bonferroni correction one could then combine the marginal confidence intervals into a simultaneous confidence set for the vector θ, which would be valid in finite samples. Given this simultaneous confidence set for θ, one could apply the approach by Klein et al. (2020) to form a simultaneous confidence set for the ranks of all categories, which would also be valid in finite samples. We compare this approach with ours in the simulations of Section 5 and find that the two methods often perform similarly well, but sometimes the Clopper–Pearson intervals are meaningfully wider. One reason why this construction may lead to wide confidence sets is that Klein et al. (2020)'s approach implicitly tests whether two success probabilities are equal by checking whether the Clopper–Pearson intervals for the two success probabilities overlap or not. This construction is excessively crude compared to the use of confidence sets for the differences. □

2.3. Confidence sets for the τ-best

Let $R_n \equiv \prod_{j \in J_0} R_{n,j}$ be a simultaneous lower confidence bound on the ranks of all categories, i.e., each $R_{n,j}$ has upper bound equal to p and R_n satisfies (1) for $J_0 = J$. Then, the projection

$$R_n^{\tau\text{-best}} \equiv \left\{ j \in J : \tau \in R_{n,j} \right\} \tag{10}$$

is a confidence set for the τ-best categories:

Corollary 2.2. *If $R_n \equiv \prod_{j \in J_0} R_{n,j}$ is defined as in Theorem 2.1 for $J_0 = J$ and $I = J^{\text{lower}}$, then $R_n^{\tau\text{-best}}$ as defined in (10) satisfies (2).*

The construction of such a confidence set using the non-randomized test with p-values in (9) can thus be summarized as follows:

Algorithm 2.2.

1. Set $J_0 = J$ and $I = J^{\text{lower}}$.
2. Perform Steps 3–5 of Algorithm 2.1 to obtain $R_n \equiv \prod_{j \in J_0} R_{n,j}$.
3. Construct $R_n^{\tau\text{-best}}$ as defined in (10).

Remark 2.3 (τ-*Worst*)**.** A confidence set for the τ-worst categories, $R_0^{\tau\text{-worst}} \equiv \{j \in J : r_j \geq p - \tau + 1\}$ can be constructed in a similar fashion as in Algorithm 2.3 for the τ-best. First, set $J_0 = J$ and $I = J^{\text{upper}}$. Then perform Steps 3–5 of Algorithm 2.1 to obtain $R_n \equiv \prod_{j \in J_0} R_{n,j}$. Finally, construct the confidence set $R_n^{\tau\text{-worst}} \equiv \left\{ j \in J : p - \tau + 1 \in R_{n,j} \right\}$. □

3. Bootstrap confidence sets

As was previously seen, confidence sets for ranks can be based on simultaneous tests of the hypotheses $H_{j,k}$ with $(j,k) \in I$ for an appropriate set of indices $I \subset J^2$. In a similar way, inference for ranks can also be based on simultaneous bootstrap confidence sets $C_n(1-\alpha, I)$ for the differences $(\theta_j - \theta_k : (j,k) \in I)$, which we now describe.

Let $X^* \equiv (X_1^*, \ldots, X_p^*)$ denote a bootstrap draw from the multinomial distribution with parameters n and $\hat{\theta} \equiv X/n$. Define the bootstrap estimator $\hat{\theta}^* \equiv X^*/n$ and

$$\hat{\sigma}_{j,k}^* \equiv \sqrt{\hat{\theta}_j^*(1-\hat{\theta}_j^*) + \hat{\theta}_k^*(1-\hat{\theta}_k^*) + 2\hat{\theta}_j^*\hat{\theta}_k^*}\,.$$

Consider the bootstrap statistic

$$T^*_{\text{lower},n}(I) \equiv \max_{(j,k)\in I} \frac{\hat\theta^*_j - \hat\theta^*_k - (\hat\theta_j - \hat\theta_k)}{\hat\sigma^*_{j,k}/\sqrt{n}},$$

where we adopt the convention that $0/0 = 0$ and $c/0 = \text{sign}(c)\infty$ for $c \neq 0$, and denote by $c_{\text{lower},n}(1-\alpha, I)$ the $(1-\alpha)$-quantile of $T^*_{\text{lower},n}(I)$ conditional on the data.[1] We can then construct lower confidence bounds for the vector of differences $\Delta_I \equiv (\theta_j - \theta_k : (j,k) \in I)$ by

$$C_{\text{lower},n}(1-\alpha, I) \equiv \prod_{(j,k)\in I} C_{\text{lower},n,j,k}(1-\alpha, I)$$

with

$$C_{\text{lower},n,j,k}(1-\alpha, I) \equiv \left[\hat\theta_j - \hat\theta_k - c_{\text{lower},n}(1-\alpha, I)\frac{\hat\sigma_{j,k}}{\sqrt{n}}, \infty \right)$$

and $\hat\sigma_{j,k} \equiv \sqrt{\hat\theta_j(1-\hat\theta_j) + \hat\theta_k(1-\hat\theta_k) + 2\hat\theta_j\hat\theta_k}$. As long as all θ_j, $j \in J$, are nonzero, this confidence set covers the vector of true differences with probability $1-\alpha$, asymptotically as the sample size n tends to infinity:

$$\lim_{n\to\infty} P\{\Delta_I \in C_{\text{lower},n}(1-\alpha, I)\} = 1-\alpha. \tag{11}$$

Appendix A provides a formal justification of this claim and further shows that the coverage probability is no less than $1-\alpha$ when some $\theta_j = 0$. Let I be equal to one of the sets J^{lower}, J^{upper}, or $J^{\text{two-sided}}$ depending on which type of confidence set for the ranks is desired. Consider the test that rejects $H_{j,k}$ iff $C_{\text{lower},n,j,k}(1-\alpha, I)$ lies entirely above zero. Then, based on this test, form the sets Rej^-_j and Rej^+_j as in (3) and (4). The bootstrap confidence set for the ranks of categories in J_0 can then be constructed as in (5); denote the resulting confidence set by $R^{\text{boot}}_n \equiv \prod_{j\in J_0} R^{\text{boot}}_{n,j}$. By an argument similar to that in Theorem 3.3 in Mogstad et al. (2024), the probability that this confidence set covers the true ranks is bounded from below by the probability that the vector of differences, Δ_I, is covered by $C_{\text{lower,n}}(1-\alpha, I)$. Therefore, the validity of the bootstrap in the sense of (11) implies that the bootstrap confidence set for the ranks also covers the true ranks with probability at least $1-\alpha$ in the limit as $n \to \infty$. The following result formalizes this discussion and shows that the validity of the bootstrap does not require any further assumptions:

Theorem 3.1. *For R^{boot}_n defined in the previous paragraph, we have*

$$\liminf_{n\to\infty} P\left\{ r_j \in R^{\text{boot}}_{n,j} \ \forall j \in J_0 \right\} \geq 1-\alpha. \tag{12}$$

Remark 3.1 (*Exactness*)**.** It is possible to show that there exists θ such that

$$\lim_{n\to\infty} P\left\{ R_j \in R^{\text{boot}}_{n,j} \ \forall j \in J_0 \right\} = 1-\alpha,$$

where R_j is defined as in Remark 2.1. In this sense, the bootstrap-based confidence sets described above are non-conservative. □

Remark 3.2 (*Uniformity*)**.** The validity of the bootstrap confidence set for the vector of differences as in (11) also holds uniformly over data-generating processes in certain classes of distributions. Such a statement could be established using the results in Romano and Shaikh (2012). One important assumption for the applicability of their results to our setting is that the elements of θ need to be bounded away from 0 and 1. As in Mogstad et al. (2024), uniform validity of the confidence sets for the differences then implies uniform validity of the confidence sets for the ranks. □

Remark 3.3 (*Stepwise Improvements*)**.** One could use stepdown procedures from Romano and Wolf (2005) to improve the confidence sets for the ranks. See Mogstad et al. (2024) for more details. □

Remark 3.4 (*Two-sided Confidence Sets*)**.** Suppose the goal is to construct a rectangular two-sided confidence set for the ranks of categories in J_0. Instead of testing the one-sided hypotheses $H_{j,k}$ for all pairs in the large set $J^{\text{two-sided}}$ with the one-sided confidence sets for the differences, one could also test whether the differences are zero using a smaller number of two-sided confidence sets for the differences.

To see this let $C_{\text{symm},n}(1-\alpha, I) \equiv \prod_{(j,k)\in I} C_{\text{symm},n,j,k}(1-\alpha, I)$ with

$$C_{\text{symm},n,j,k}(1-\alpha, I) \equiv \left[\hat\theta_j - \hat\theta_k \pm c_{\text{symm},n}(1-\alpha, I)\frac{\hat\sigma_{j,k}}{\sqrt{n}} \right],$$

[1] In a given bootstrap sample, the ratio inside the max of $T^*_{\text{lower},n}$ can have a zero denominator and/or zero numerator. For instance, when two categories j and k both have small success frequencies in the data ($\hat\theta_j$ and $\hat\theta_k$ are small), then it is possible that a given bootstrap sample does not contain any success for either of the two categories, i.e., $\hat\theta^*_j = \hat\theta^*_j = \hat\sigma^*_{j,k} = 0$, and the denominator is zero. When $\hat\theta_j < \hat\theta_k$, then the resulting critical value $c_{\text{lower},n}(1-\alpha, I)$ equals ∞. On the other hand, when the frequencies in the data are equal ($\hat\theta_j = \hat\theta_k$), then both the numerator and denominator of the ratio are zero and the maximum is determined by other bootstrap samples that produce a positive ratio.

where $c_{\text{symm},n}(1-\alpha, I)$ denotes the $(1-\alpha)$-quantile of

$$T^*_{\text{symm},n}(I) \equiv \max_{(j,k)\in I} \frac{\left|\hat{\theta}^*_j - \hat{\theta}^*_k - (\hat{\theta}_j - \hat{\theta}_k)\right|}{\hat{\sigma}^*_{j,k}/\sqrt{n}}$$

conditional on the data. Then, set $I = J^{\text{upper}}$ and compute

$$N_j^- \equiv \{k \in J \setminus \{j\} : C_{\text{symm},n,j,k}(1-\alpha, I) \text{ lies entirely below zero }\}$$
$$N_j^+ \equiv \{k \in J \setminus \{j\} : C_{\text{symm},n,j,k}(1-\alpha, I) \text{ lies entirely above zero }\}$$

which indicate the categories k with probabilities strictly larger or smaller than θ_j. A confidence set for the ranks of categories in J_0 can then be formed as in (5), replacing Rej_j^- and Rej_j^+ by N_j^- and N_j^+, respectively. By arguments analogous to those for the one-sided confidence sets, the resulting confidence set for the ranks of categories in J_0 then covers the true ranks with probability approaching at least $1-\alpha$ as $n \to \infty$. □

Remark 3.5 (*Studentization*)**.** The bootstrap procedure in Remark 3.4 may perform poorly in the sense of under-covering the true ranks when there are many categories and all estimated probabilities $\hat{\theta}_1, \ldots, \hat{\theta}_p$ are small. In such situations, the ratio in the definition of the bootstrap statistic may evaluate to $0/0$ on many bootstrap samples, leading to a critical value that is too small. In addition, the bootstrap procedure may perform poorly in the sense of yielding confidence sets that are very wide when there are two or more categories with small estimated probabilities. In such situations, there may be bootstrap samples without any successes for categories j and k, so the ratio in the definition of the bootstrap statistic evaluates to ∞ and the resulting critical value is equal to ∞; see Footnote 1. For these reasons, it may be beneficial not to studentize $T^*_{\text{symm},n}(I)$, in which case one would also remove $\hat{\sigma}_{j,k}$ from the expression of $C_{\text{symm},n,j,k}(1-\alpha, I)$. These two approaches are compared further in the simulations in Section 5.3. □

Remark 3.6 (*"Naive" Bootstrap*)**.** Suppose the goal is to construct a confidence set for the rank of a single category. The confidence set R_n based on Algorithm 2.1 was shown to be valid in finite samples, regardless of the value of the vector θ. In particular, there may be an arbitrary number of ties or near-ties in θ. Similarly, the confidence set R_n based on the bootstrap as proposed in this section is asymptotically valid regardless of the number of ties or near-ties in θ. On the other hand, the bootstrap as proposed by, e.g., Goldstein and Spiegelhalter (1996) performs poorly when, for some $k \neq j$, θ_k is (close to) equal to θ_j. For concreteness, consider the following "naive" bootstrap procedure. For a category j, denote by $\hat{\theta}^*_j$ the estimator of θ_j computed on a bootstrap sample and let $\hat{r}^*_j$ be the rank computed using the bootstrap estimators $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_p$. Confidence sets for r_j could then be constructed using upper and/or lower empirical quantiles of $\hat{r}^*_j$ conditional on the data. Mogstad et al. (2024) show that this intuitive approach fails to deliver the desired coverage property when there are ties (unless $p = 2$). In fact, the coverage probability tends to zero as p grows. For further discussion, see Xie et al. (2009) and Hall and Miller (2009). In contrast, our bootstrap approach does not rely on a consistent estimator of the distribution of estimated ranks but rather on the availability of simultaneous bootstrap confidence sets for the differences Δ_I with asymptotic coverage no less than the desired level. Such simultaneous confidence sets are available under weak conditions and, in particular, do not restrict the configuration of the vector of probabilities θ. In comparison to Xie et al. (2009), our bootstrap procedure also circumvents smoothing of the indicator in the definition of the ranks and thus the need for choosing such a smoothing parameter. □

4. Ranking political parties by voters' support in the Australian Election Study 2019

In this section, we apply the inference procedures from Sections 2 and 3 to examine the ranking of political parties by their share of voters' support in the Australian Election Study (AES). The AES has fielded representative surveys after every federal election since 1987 and provides the most comprehensive source of evidence on political attitudes in Australia (Cameron and McAllister, 2019). We use AES data from 2019 with address-based stratified random sampling from the Geocoded National Address File (G-NAF) (Bean et al., 2019).[2] Table 1 shows a total of 3944 sampled eligible voters from all fifteen Australian territories resulting in 1211 respondents. In the subsequent analysis we work with respondents and refer to them as "sample".

To examine which political parties are on the top and the bottom of the ranking in each Australian territory, we use respondents' answers to the survey question "Generally speaking, do you usually think of yourself as Liberal, Labor, National, Greens or other (specify)?". The answer categories include political parties; "Skipped" and "No answer" categories that we group in a single "No answer" category; "Independent", "Swing Voter" and "No party" categories that we group in a single "No party" category. For more populous territories, we observe between seven and ten categories with positive support shares.

By applying the inference procedures from Sections 2 and 3 we compute (i) the marginal confidence set for the rank of a particular political party and (ii) the simultaneous confidence set for the ranks of all parties. Thus, (i) is relevant if one is interested whether a particular party is on the top (bottom) of ranking by the voters' support, and (ii) is relevant if one is interested in the entire ranking of parties.

[2] The original sampling methodology description reads: "Within the parameters outlined above, the new AES sample was selected from the G-NAF database using a stratified sample design in accordance with the geographical distribution of the Australian residential population aged 18 years and over. GNAF sample selections were supplied by the MasterSoft Group. A total of 3944 sample records were randomly generated within 15 geographic strata (see Table 2) to ensure sufficient sample was utilized to achieve the desired number of responses for the AES" (Bean et al., 2019). We interpret this sampling as being i.i.d. within territories.

Table 1
Australian election study 2019 G-NAF stratified random sample of 3944 eligible voters resulting in 1211 respondents. In the subsequent analysis we work with respondents and refer to them as "sample". The last column shows the number of categories with positive support share in each territory measured by answers to AES2019 survey question "Generally speaking, do you usually think of yourself as Liberal, Labor, National, Greens or other(specify)?". The answers include political parties, "Skipped" and "No answer" categories that we group in a single "No answer" category; "Independent", "Swing Voter" and "No party" categories that we group in a single "No party" category.

Territory	Sampled voters	Respondents	Categories
Greater Sydney	816	238	8
Greater Melbourne	780	234	7
Rest of New South Wales	445	144	8
Rest of Queensland	402	121	10
Greater Brisbane	378	115	8
Greater Perth	319	93	7
Rest of Victoria	248	82	8
Greater Adelaide	217	81	9
Rest of Western Australia	87	26	7
Australian Capital Territory	67	24	4
Rest of South Australia	63	17	4
Rest of Tasmania	47	16	3
Greater Hobart	35	12	4
Greater Darwin	24	6	3
Rest of Northern Territories	16	2	2
Total:	3944	1211	

4.1. Marginal confidence sets for the ranks of political parties in Greater Melbourne

Consider first a particular territory, Greater Melbourne. Fig. 1 shows the point estimates and standard errors for voters' support share for each category with positive number of supporters. The leftmost panel in row A shows considerable variation in point estimates across categories from 0.371 for the most supported (Labor Party) to 0.004 for the least supported (National Party). Support shares on the top and the bottom of the ranking are close to each other, while the shares in the middle are better separated.

The middle panel in row A of Fig. 1 presents the 95% marginal confidence sets for the rank of each category implemented using five procedures: the exact Holm (**"exactHolm"**) described in the Algorithm 2.1, Clopper–Pearson (**"CP"**) as in Remark 2.2, non-studentized (**"boot"**) and studentized (**"bootStud"**) versions of the bootstrap as in Section 3 and the "naive" bootstrap (**"naive"**) as in Remark 3.6. The first two methods have been shown to be valid in finite samples, and the studentized and non-studentized bootstrap are motivated by asymptotic validity. The "naive" bootstrap is asymptotically valid in the absence of (near-)ties (see Remark 3.6), but invalid otherwise.

The resulting confidence sets exhibit four pronounced patterns. First, the "naive" bootstrap confidence sets are the tightest. As indicated in Remark 3.6, the "naive" bootstrap produces confidence sets that fail to cover the true ranks with the desired probability when there are (near-) ties. Our simulations in Section 5 confirm that, in datasets like the one from Greater Melbourne, the "naive" bootstrap does indeed produce short confidence sets at the expense of its coverage frequency lying substantially below the desired nominal level. Second, the exact Holm procedure produces weakly shorter confidence sets than Clopper–Pearson and the studentized and non-studentized bootstraps. For example, the exact Holm confidence sets for the ranks of the categories "No party" and "Greens" contain only ranks three and four, respectively, while Clopper–Pearson and the studentized and non-studentized bootstraps produce confidence sets containing at least two ranks. Third, both the studentized and non-studentized bootstrap confidence sets are wide in the middle of the ranking. Their length for the 4th category, "Greens", is equal to four compared to the maximum possible length of six. Fourth, the studentized bootstrap produces extremely wide confidence sets at the bottom of the ranking for the last two categories. These confidence sets cover the entire ranking and correspond to infinite critical values. In contrast, the length of the finite-sample valid confidence sets for the last two categories is two. These patterns suggest that the finite-sample valid confidence sets for parties in Greater Melbourne are informative, and the exact Holm procedure is the most informative among the valid procedures (i.e., excluding the "naive" bootstrap).

As discussed in Footnote 1, the studentized bootstrap confidence sets are wide when the ratio in the bootstrap test statistic has a denominator that is equal to zero and a positive numerator in a substantial fraction of the bootstrap samples. This circumstance arises when at least two categories have small but positive shares in the data so that, in the bootstrap samples, $\hat{\theta}_j^* = \hat{\theta}_k^* = \hat{\sigma}_{j,k}^* = 0$ while, in the data, $|\hat{\theta}_j - \hat{\theta}_k| > 0$. One solution is to group the categories with small shares together. The leftmost panel in row B of Fig. 1 shows support shares when we group "National Party", "One Nation" and "No answer" into a single category "Other". The middle panel in row B shows that in the middle or at the bottom of the ranking, both the studentized and non-studentized

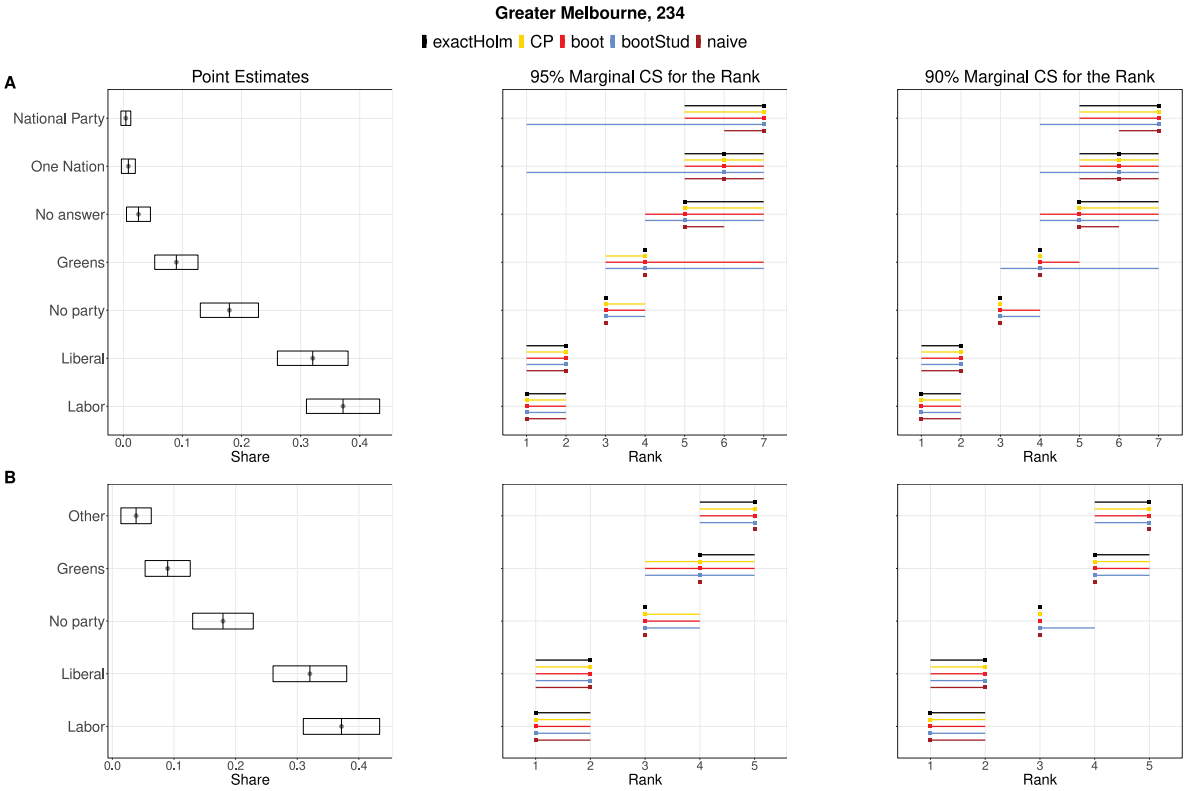


Fig. 1. **Left** column shows point estimates of categories' support shares in Greater Melbourne and ±1.96se. **Middle** column shows the 95% marginal confidence sets for the rank of each category computed by five procedures for each category. **Right** column shows the 90% marginal confidence sets for the rank of each category computed by five procedures for each category. In row **A** we use originally defined choice categories. In row **B**, we group "National Party", "One Nation" and "No answer" into a single category "Other".

bootstraps no longer produce confidence sets as wide as in row A. Notably, the exact Holm confidence sets are still tighter than Clopper–Pearson confidence sets.

An alternative solution to the division by zero in bootstrap samples is to reduce the confidence level. The rightmost column of Fig. 1 shows the 90% marginal confidence sets for the ranks using both the original set of categories in row A and with the three smallest categories grouped in row B. Indeed, with lower confidence level we no longer observe studentized bootstrap confidence sets covering the entire ranking. However, both finite-sample methods still produce weakly smaller confidence sets than both types of bootstrap. Furthermore, the panel with grouped small categories in row B shows that the exact Holm confidence sets for "Greens" are less informative than for the original categories, regardless of whether the confidence level is 90% or 95%.

4.2. Marginal confidence sets for the eight most populous territories

Next, we move beyond the example of Greater Melbourne and consider all fifteen Australian territories. Appendix Figs. 7(a) and 7(b) show the point estimates for support shares in all territories, and Appendix Figs. 8(a) and 8(b) show the 95% marginal confidence sets for the rank of each category in each territory computed using the same five procedures as for Greater Melbourne. Similar to our illustration for Greater Melbourne, both bootstrap procedures produce wide confidence sets in the middle and at the bottom of the ranking in the majority of populous territories. Due to small sample sizes the seven least populous territories have mostly uninformative confidence sets for all categories. Therefore, we focus our analysis on the eight most populous territories.

Fig. 8(a) shows that none of the valid methods produce tighter confidence sets than all other valid methods uniformly across all categories in all territories. For example, the finite-sample valid confidence sets are weakly tighter than both types of bootstrap confidence sets for each category in Greater Melbourne. In comparison, the bootstrap confidence sets are weakly tighter in the top part of the ranking in Greater Sydney. We summarize this finding in the top panel of Table 2 that shows the percentage

Table 2
Each cell shows the percentage of pairwise comparisons across all categories in the eight most populous territories where the inference procedure in a row produces wider 95% marginal confidence sets for the ranks than the procedure in a column. The **top panel** shows results for the original set of categories in each territory. The **bottom panel** shows results when we group all categories except "Liberal", "Labor", "Greens" and "No party" into a single category "Other".

Original set of categories				
	exactHolm	CP	boot	bootStud
exactHolm		6.2	12.3	1.5
CP	3.1		9.2	0.0
Boot	29.2	27.7		0.0
bootStud	47.7	49.2	41.5	

Small categories grouped				
	exactHolm	CP	Boot	bootStud
exactHolm		2.5	22.5	2.5
CP	10.0		22.5	0.0
Boot	7.5	0.0		0.0
bootStud	15.0	5.0	27.5	

of category×territory cases across the eight most populous territories where each method produces strictly wider 95% marginal confidence sets for the rank than other methods. For example, the first row shows that the exact Holm confidence sets are strictly wider than Clopper–Pearson confidence sets in 6.2% of category×territory cases, strictly wider than the non-studentized bootstrap confidence sets in 12.3% of cases, and strictly wider than the studentized bootstrap confidence sets in 1.5% of cases. The first column shows that Clopper–Pearson confidence sets are strictly wider than the exact Holm confidence sets in 3.1% of category×territory cases, the non-studentized bootstrap confidence sets are strictly wider than the exact Holm in 29.2% of cases, and the studentized bootstrap confidence sets are strictly wider than the exact Holm confidence sets in 47.7% of cases. Notably, the share of cases where the studentized and non-studentized bootstrap confidence sets are strictly wider than Clopper–Pearson or the exact Holm confidence sets is substantially larger than the share of cases where Clopper–Pearson and the exact Holm confidence sets are strictly wider than both types of bootstrap confidence sets.

The bottom panel of Table 2 shows the same comparisons when we group all categories except "Liberal", "Labor", "Greens" and "No party" into a single category "Other". As discussed above, this grouping prevents zeros in the bootstrap test statistic denominator and excessively wide bootstrap confidence sets. As a result, the percentage of cases in rows 3 and 4 where both types of bootstrap confidence sets are strictly wider than the finite-sample valid confidence sets decreases. Interestingly, the percentage of cases where the exact Holm confidence sets are strictly wider than Clopper–Pearson confidence sets is lower with grouping, and the percentage of cases where Clopper–Pearson confidence sets are wider than the exact Holm confidence sets is larger.

4.3. Marginal versus simultaneous confidence sets

The analysis of marginal confidence sets answers questions about the rank of a particular party, but one may be interested in the ranking of all parties described by simultaneous confidence sets. Fig. 2 compares 95% marginal confidence sets for the ranks to 95% simultaneous confidence sets for the ranks produced by the exact Holm, Clopper–Pearson, the studentized and non-studentized bootstrap procedures for categories in Greater Melbourne. Naturally, simultaneous confidence sets are weakly wider than marginal confidence sets for each procedure. This feature is more pronounced for the studentized bootstrap confidence sets due to infinite critical values in the bootstrap test statistic for categories at the bottom of the ranking. In contrast, the finite-sample valid simultaneous confidence sets are still informative, and the exact Holm produces weakly tighter confidence sets than all other procedures. Appendix Fig. 9 shows that with confidence level reduced to 90% we no longer observe studentized bootstrap confidence sets as wide as on Fig. 2 since the critical value becomes finite.

The top panel of Table 3 shows that our findings hold in the eight most populous Australian territories with the original set of choice categories. The studentized bootstrap almost always produces strictly wider 95% simultaneous confidences sets than other methods and never produces tighter confidence sets. The exact Holm simultaneous confidence sets are never strictly wider than Clopper–Pearson or the studentized bootstrap confidence sets, and are strictly wider than the non-studentized bootstrap confidence sets in only 9.2% of category×territory cases. In contrast, both Clopper–Pearson and the non-studentized bootstrap simultaneous confidence sets are strictly wider than the exact Holm confidence sets in 26.2% of category×territory cases.

In the bottom panel of Table 3 we group all categories except "Liberal", "Labor", "Greens" and "No party" into a single category "Other". As a result, both types of bootstrap confidence sets become tighter. In particular, the non-studentized bootstrap simultaneous confidence sets are never strictly wider than the confidence sets produced by any other inference procedure. The

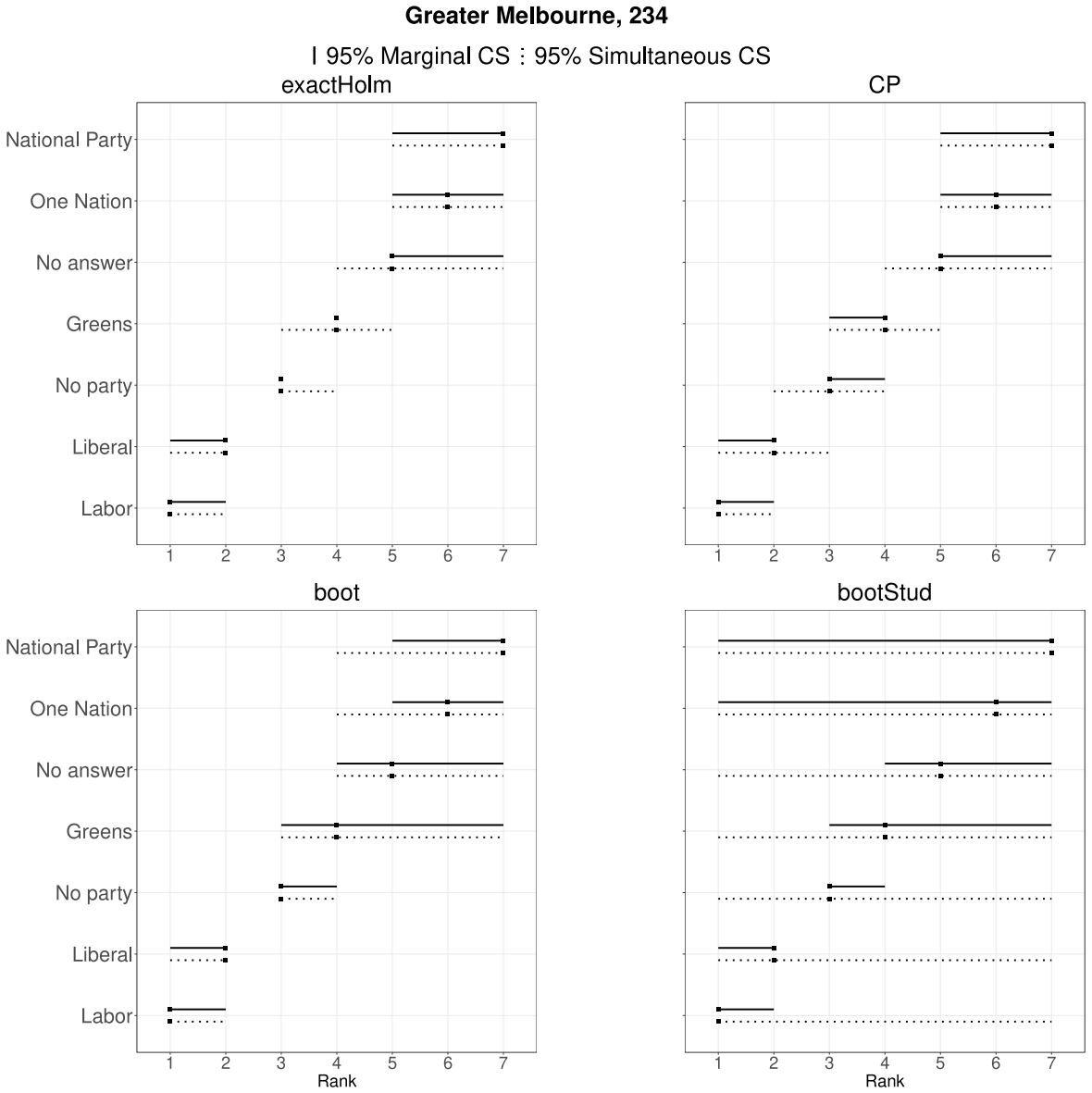


Fig. 2. 95% marginal and 95% simultaneous confidence sets for the ranks of categories in Greater Melbourne.

exact Holm confidence sets, however, are still never wider than Clopper–Pearson or the studentized bootstrap confidence sets and are strictly wider than the non-studentized bootstrap confidence sets in only 5% of category×territory cases.

5. Simulations

In this section, we examine the finite-sample performance of the following approaches to constructing confidence sets for ranks:

"exactBonf": the confidence set computed through Algorithm 2.1 using the Bonferroni correction.

"exactHolm": the confidence set computed through Algorithm 2.1 using the Holm correction.

"CP": the confidence set based on Clopper–Pearson confidence sets for binomial probabilities as described in Remark 2.2.

Table 3
Each cell shows the percentage of pairwise comparisons across all categories in the eight most populous territories where the inference procedure in a row produces wider 95% simultaneous confidence sets for the ranks than the procedure in a column. The **top panel** shows results for the original set of categories in each territory. The **bottom panel** shows results when we group all categories except "Liberal", "Labor", "Greens" and "No party" into a single category "Other".

Original set of categories				
	exactHolm	CP	Boot	bootStud
exactHolm		0.0	9.2	0.0
CP	26.2		23.1	0.0
Boot	26.2	21.5		0.0
bootStud	87.7	83.1	76.9	

Small categories grouped				
	exactHolm	CP	Boot	bootStud
exactHolm		0.0	5.0	0.0
CP	20.0		25.0	0.0
Boot	0.0	0.0		0.0
bootStud	32.5	20.0	37.5	

"boot": the bootstrap (not studentized) confidence set based on two-sided confidence sets for the differences as described in Remark 3.5.

"bootStud": the bootstrap (studentized) confidence set based on two-sided confidence sets for the differences as described in Remark 3.4.

"naive": the "naive" bootstrap confidence set as described in Remark 3.6.

All simulations are based on 1000 Monte Carlo samples and nominal coverage of 95%. Bootstrap confidence sets are based on 10,000 bootstrap samples except in Section 5.2, where for computational reasons we use 1000 bootstrap samples. Coverage is defined as in Remark 2.1, i.e., coverage of the set of possible ranks R_j.

We consider three different designs, starting with one that is calibrated to the dataset on which the empirical application is based. Second, we investigate whether the confidence sets exhibit erratic behavior in coverage frequencies similar to that reported for confidence sets for binomial proportions (Brown et al., 2001). The first two designs consider only data generating processes with three or seven categories to be ranked. In the final simulation design, we analyze the behavior of the confidence sets as the number of categories increases.

5.1. AES design

The simulation design in this subsection is calibrated to the AES data for Greater Melbourne as in Section 4.1. The estimated vector of success probabilities is

$$\hat{\theta}_{AES} = (0.372, 0.321, 0.179, 0.090, 0.026, 0.009, 0.004)$$

and the number of respondents is $n_{AES} = 234$. The vector of success probabilities employed in the simulations, θ, is parametrized as

$$\theta = (1 - \kappa)\frac{1}{p}\iota + \kappa\hat{\theta}_{AES},$$

where ι denotes a vector of ones and $\kappa \in [0, 1]$. So, when $\kappa = 1$, then the vector of probabilities is the same as in the data set. When $\kappa = 0$, then all probabilities are equal, and values of κ between 0 and 1 generate probabilities between the two extremes. A parameter $\tau \in \{0.5, 1, 2\}$ is introduced to vary the sample size as $n = \tau\, n_{AES}$. So, when $\tau = 1$, then the sample size in the simulation is equal to the one in the data set, but we also consider half and double that sample size.

We begin by recalling the four main findings about 95% marginal confidence sets for the ranks of categories in Greater Melbourne from Section 4.1:

1. naive bootstrap confidence sets are the tightest
2. exactHolm confidence sets are weakly tighter than confidence sets produced by all other valid procedures (i.e. all except the naive bootstrap)
3. boot and bootStud confidence sets are very wide in the middle of the ranking

4. bootStud confidence sets are very wide at the bottom of the ranking

Below we explore each of these findings in depth by focusing on lengths and empirical coverage frequencies of confidence sets for the rank of the 1st (top of the ranking), 4th (middle of the ranking) and 7th (bottom of the ranking) categories in Table 4.

First, naive bootstrap confidence sets are the tightest for all three categories in all simulations. Table 4 shows that this tightness comes at the cost of severe under-coverage for the 1st and the 7th categories. For the 7th category naive bootstrap confidence sets under-cover in all but two simulations, and empirical coverage frequency may be as low as 59.5%. Note that even for $\kappa = 1$ the success probabilities for the bottom categories are not well separated, which explains more pronounced under-coverage for the 7th category. In contrast, finite-sample methods (exactHolm and CP) cover the rank with the frequency no smaller than the desired level for all parametrizations and sample sizes. Both boot and bootStud confidence sets cover the 1st category with the frequency close to the nominal level, especially when success probabilities are equal ($\kappa = 0$) and the sample size is small ($\tau = 0.5$). When the categories are better separated ($\kappa = 1$) as in the dataset, then both finite-sample and bootstrap procedures cover the true rank with probability (close to) one.

Second, in contrast to our finding for Greater Melbourne, the finite-sample valid confidence sets are not uniformly tighter than the bootstrap confidences sets in simulations. In most parametrizations where boot and bootStud confidence sets are tighter, however, the difference in the size of the average confidence sets is below 0.1 and never exceeds 0.3, where the reference size of the confidence set covering the entire ranking is 6.0.

Third, both CP and exactHolm methods produce much tighter confidence sets than both types of bootstrap for the 4th category when $\kappa = 1$, i.e., when success probabilities are equal to point estimates from Greater Melbourne respondents' sample. Specifically, the average length of exactHolm confidence sets is less than one-third of the average length of boot and bootStud confidence sets when $\tau = 1$ and less than one-quarter of their average size when $\tau = 2$. Furthermore, the average length of exactHolm confidence sets is more than 20% lower than the average length of CP confidence sets in both instances. Despite the shorter average length, exactHolm confidence sets' empirical coverage frequency is above the desired level. Notice that both finite-sample valid confidence sets are of length zero in the majority of simulations with $\kappa = 1$ and $\tau = 2$ and contain only a single value 4. In contrast, most boot and bootStud confidence sets are of length one or above, meaning that they contain at least two values and are not as informative as finite-sample valid confidence sets.

Fourth, when the success probabilities are equal to point estimates from the respondents' sample of Greater Melbourne ($\kappa = 1$), CP and exactHolm produce much tighter confidence sets for the 7th category than bootStud. For $\tau = 0.5$ and $\tau = 1$ the average length of CP and exactHolm confidence sets is more than two times smaller than the average length of bootStud confidence set. The difference becomes less pronounced with the increase in sample size, but the average length of the exactHolm confidence set is still more than 50% smaller than the average length of bootStud confidence set for $\tau = 2$. Furthermore, the exactHolm confidence set is substantially shorter than the boot and CP confidence sets for all values of τ. Tighter exactHolm confidence sets for the 7th category still provide empirical coverage frequency above the desired level.

Table 4 also highlights common features of the inference procedures. For equal success probabilities ($\kappa = 0$) the set of ranks R_j is $[1, 7]$ for all categories j, and the average length of all confidence sets barely decreases for larger sample sizes τ. When success probabilities differ ($\kappa > 0$) larger sample size τ means differences in success probabilities are easier to detect, and all confidence sets for the rank decrease in length. Similarly, the average length of all confidence sets except for bootStud for 7th category decreases as we increase κ, which means differences in success probabilities are larger and thus, again, easier to detect. For bootStud at the bottom of the ranking, the effect of better separation for $\kappa > 0$ is mitigated by decreasing success probabilities leading to more frequent division by zero events in bootstrap test statistics and thus larger critical values.

Finally, in addition to the five confidence sets we used in Section 4 we include exactBonf in all simulations. Table 4 shows that, as expected, exactBonf confidence sets are uniformly wider than exactHolm confidence sets, but the difference in the average length is not substantial.

5.2. Erratic coverage

Brown et al. (2001) found that coverage frequencies of some confidence intervals for binomial proportions may vary in highly non-monotonic ways with the sample size and the success probability. Furthermore, they found that coverage frequencies may be far below the desired level, especially for small sample sizes and/or small success probabilities. Motivated by this "erratic" behavior of coverage in the binomial case, in this subsection, we compare our confidence set (exactBonf), which is valid in finite samples, with the bootstrap confidence sets (boot and bootStud), which are justified by asymptotic validity. In particular, we are interested in their coverage properties in small samples and/or scenarios with small success probabilities. To this end we consider three categories and set the vector of success probabilities as $\theta = (\pi, \pi, 1 - 2\pi)$, where π is varied between $1/100$ and $1/3$. The sample size is varied between 10 and 100.

For the different values of π, Fig. 3 shows the frequencies of the confidence set $C_{\text{symm},n}(1 - \alpha, I)$ simultaneously covering all differences involving the first category, Δ_I, where $I = J^{\text{two-sided}}$ with $J_0 = \{1\}$. The coverage frequencies are plotted as functions of the sample size n for the bootstrap with (panel (a)) and without (panel (b)) studentization. Both bootstrap approaches lead to simultaneous confidence sets for the differences that considerably under-cover for small sample sizes and or small probabilities π, analogously to the findings in Brown et al. (2001). The coverage probability for $\pi = 1/100$ can be even lower than 0.4 when sample sizes are small ($n \leq 20$).

Fig. 4 shows the coverage frequencies of the resulting confidence sets for the rank based on the two bootstrap approaches (panels (a) and (b)) and, for comparison, also the coverage frequencies of the finite sample method (panel (c)). Perhaps somewhat

Table 4
Average lengths and empirical coverage frequencies from 1000 Monte Carlo samples for the 95% marginal confidence sets for the rank of the 1st (**top panel**), 4th (**middle panel**) and 7th (**bottom panel**) categories.

κ	τ	Length:						Coverage Frequency:					
		exactBonf	exactHolm	CP	Boot	BootStud	Naive	exactBonf	exactHolm	CP	Boot	bootStud	Naive
1st category:													
0	0.5	5.971	5.970	5.962	5.944	5.927	5.089	0.985	0.985	0.980	0.960	0.950	0.741
	1	5.965	5.965	5.957	5.949	5.934	5.107	0.985	0.985	0.980	0.967	0.957	0.714
	2	5.963	5.962	5.955	5.938	5.943	5.103	0.988	0.988	0.984	0.970	0.973	0.727
0.5	0.5	3.223	3.120	3.043	3.001	3.622	1.757	1.000	1.000	1.000	0.994	1.000	0.989
	1	1.895	1.797	1.778	1.689	1.965	1.252	1.000	1.000	1.000	0.999	1.000	0.997
	2	1.258	1.210	1.197	1.130	1.252	0.960	1.000	1.000	1.000	1.000	1.000	1.000
1	0.5	1.490	1.430	1.362	1.302	1.476	1.051	1.000	1.000	1.000	1.000	1.000	0.998
	1	1.046	1.003	0.957	0.897	0.995	0.826	1.000	1.000	1.000	0.999	1.000	0.999
	2	0.926	0.889	0.841	0.774	0.854	0.731	1.000	1.000	0.999	0.998	1.000	0.998
4th category:													
0	0.5	5.977	5.977	5.971	5.944	5.935	5.094	1.000	1.000	1.000	1.000	1.000	0.982
	1	5.957	5.955	5.950	5.925	5.924	5.105	0.999	0.998	0.998	0.997	0.996	0.975
	2	5.951	5.951	5.940	5.924	5.916	5.090	1.000	1.000	1.000	1.000	1.000	0.971
0.5	0.5	5.203	5.192	5.116	4.867	5.130	3.553	1.000	1.000	1.000	0.999	1.000	0.993
	1	4.232	4.192	4.141	4.159	4.388	2.734	1.000	1.000	1.000	1.000	1.000	0.998
	2	3.274	3.170	3.239	3.503	3.761	1.882	1.000	1.000	1.000	1.000	1.000	1.000
1	0.5	2.540	2.472	2.595	3.782	3.953	0.941	1.000	1.000	1.000	1.000	1.000	1.000
	1	1.007	0.861	1.036	2.771	3.353	0.355	1.000	1.000	1.000	1.000	1.000	1.000
	2	0.233	0.171	0.222	0.821	1.314	0.058	1.000	1.000	1.000	1.000	1.000	1.000
7th category:													
0	0.5	5.968	5.966	5.959	5.914	5.910	5.079	0.985	0.985	0.983	0.984	0.982	0.672
	1	5.970	5.968	5.961	5.946	5.937	5.112	0.985	0.985	0.981	0.985	0.984	0.672
	2	5.967	5.967	5.959	5.951	5.948	5.185	0.985	0.985	0.980	0.986	0.985	0.728
0.5	0.5	4.315	4.288	4.267	3.994	4.253	3.036	0.995	0.995	0.995	1.000	1.000	0.908
	1	3.393	3.345	3.366	3.354	3.550	2.537	0.996	0.994	0.996	1.000	1.000	0.943
	2	2.802	2.748	2.799	2.865	3.018	2.214	0.998	0.998	0.999	1.000	1.000	0.964
1	0.5	2.331	2.331	2.393	2.777	5.097	1.220	1.000	1.000	1.000	1.000	1.000	0.595
	1	1.892	1.879	1.930	2.247	3.926	1.237	1.000	1.000	1.000	1.000	1.000	0.856
	2	1.525	1.492	1.630	2.003	2.277	1.034	1.000	1.000	1.000	1.000	1.000	0.986

surprisingly, the coverage frequencies for the bootstrap methods are above $1-\alpha$ in all scenarios, even when success probabilities and/or sample sizes are small. Hence, under-coverage of the differences observed in Fig. 3 does not lead to under-coverage of the rank. The reason for this phenomenon is that correct coverage of the rank only requires that the confidence sets for the differences do not lead to incorrect claims about the signs of the differences. While not necessarily covering the true values of the differences, the bootstrap confidence sets for the differences lead to the correct determination of their signs and thus to coverage of the rank. As expected, the finite-sample method covers the rank with the desired probability even for small samples sizes and/or small success probabilities.

5.3. Larger number of categories

In this subsection, we consider a simulation design in which the success probabilities are all equal, i.e. $\theta_j = 1/p$ for all $j = 1, \ldots, p$, and we increase p from 5 to 50. For the different values of p, Figs. 5 and 6 show the coverage frequencies and lengths of the different confidence sets as functions of the sample size n.

First of all, the finite-sample methods exactBonf, exactHolm, and CP cover the rank with coverage frequencies close to one in all scenarios. As expected the naive bootstrap fails to cover the true rank with the desired probability. Its coverage may be considerably below 0.4 and even approaches zero when there are many categories. Interestingly, however, the bootstrap method based on the studentized statistic also considerably under-covers when there are (moderately) many categories. For instance, with $p = 20$ categories, its coverage frequency can be well below 0.8 for small sample sizes. For more categories ($p = 50$), the under-coverage occurs up to larger sample sizes.

In terms of length, the exactBonf, exactHolm, CP, and boot perform similarly. As expected bootStud and naive lead to shorter confidence sets in those scenarios in which they under-cover.

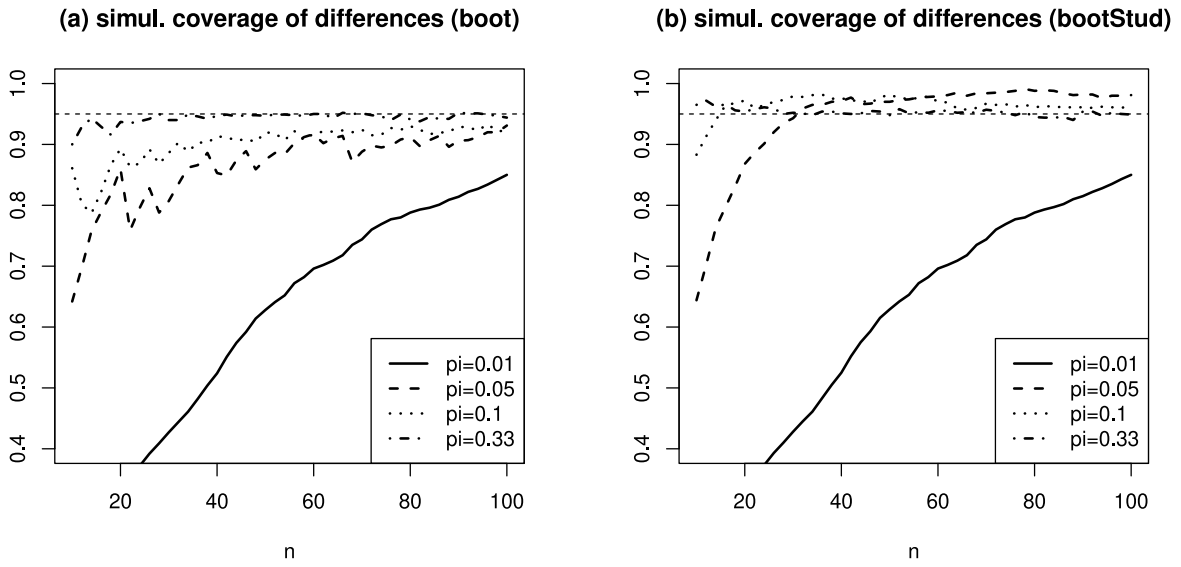


Fig. 3. Frequencies of simultaneously covering all differences involving the first category, i.e., the frequency of $\Delta_I \in C_{\text{symm},n}(1-\alpha, I)$. The horizontal dashed line marks the desired coverage level $1-\alpha$.

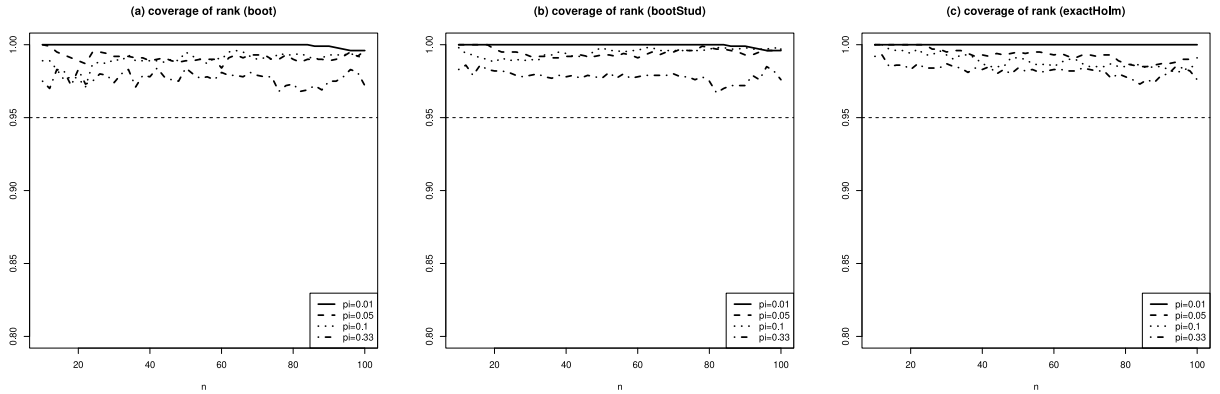


Fig. 4. Coverage frequencies for the rank of the first category. The horizontal dashed line marks the desired coverage level $1-\alpha$.

Appendix A. Asymptotic validity of the bootstrap

For the arguments in this section, we slightly change notation by indexing population quantities by P, the underlying probability mechanism that specifies the multinomial sampling probabilities. For instance, let $\theta \equiv \theta(P) \equiv (\theta_1(P), \ldots, \theta_p(P))'$ denote the probabilities of a particular P and similarly

$$\Delta_I(P) \equiv (\Delta_{j,k}(P) : (j,k) \in I) ,$$

where

$$\Delta_{j,k}(P) \equiv \theta_j(P) - \theta_k(P)$$

and $I \subseteq J^2$. As in the main text let

$$\hat{\sigma}^2_{j,k} \equiv \hat{\theta}_j(1-\hat{\theta}_j) + \hat{\theta}_k(1-\hat{\theta}_k) + 2\hat{\theta}_j\hat{\theta}_k \tag{13}$$

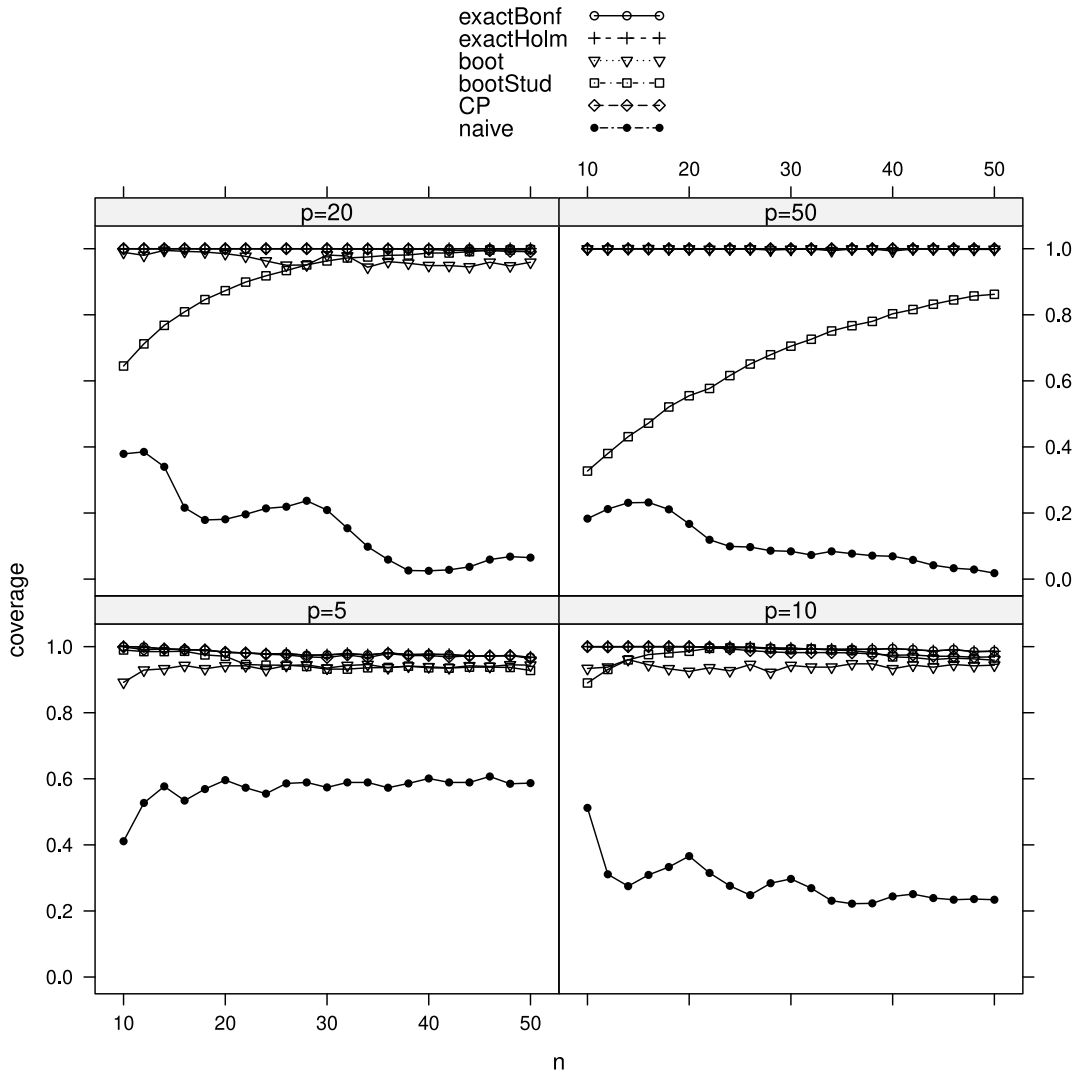


Fig. 5. Coverage frequencies of confidence sets for different n and p.

define the cumulative distribution functions

$$L_{\text{lower},n}(x, I, P) \equiv P\left\{\max_{(j,k)\in I} \frac{\hat{\theta}_j - \hat{\theta}_k - \Delta_{j,k}(P)}{\hat{\sigma}_{j,k}/\sqrt{n}} \leq x\right\}, \tag{14}$$

$$L_{\text{upper},n}(x, I, P) \equiv P\left\{\max_{(j,k)\in I} \frac{\Delta_{j,k}(P) - (\hat{\theta}_j - \hat{\theta}_k)}{\hat{\sigma}_{j,k}/\sqrt{n}} \leq x\right\}, \tag{15}$$

$$L_{\text{symm},n}(x, I, P) \equiv P\left\{\max_{(j,k)\in I} \frac{|\hat{\theta}_j - \hat{\theta}_k - \Delta_{j,k}(P)|}{\hat{\sigma}_{j,k}/\sqrt{n}} \leq x\right\}. \tag{16}$$

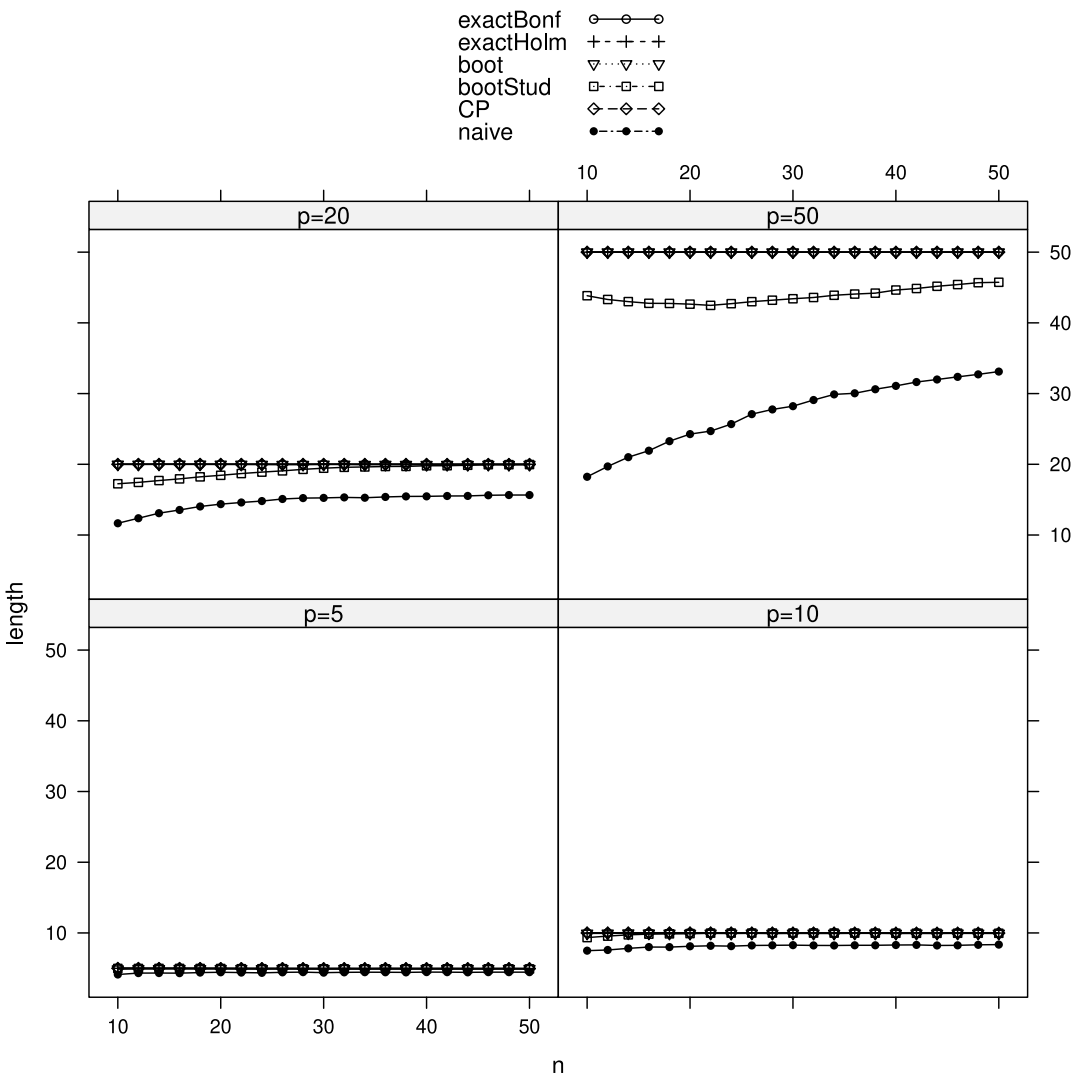


Fig. 6. Length of confidence sets for different n and p.

Let $\hat{P}_n$ be an estimate of P, where $\hat{P}_n$ specifies the empirical frequencies $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$. Then the bootstrap quantiles can be written as

$$c_{l,n}(1-\alpha, I) = L_{l,n}^{-1}(1-\alpha, I, \hat{P}_n)$$

for $l \in \{\text{lower}, \text{upper}, \text{symm}\}$. Here, it is understood that, for a cumulative distribution function $F(x)$ on the real line, the quantity $F^{-1}(1-\alpha)$ is defined to be $\inf\{x \in \mathbf{R} : F(x) \geq 1-\alpha\}$. The bootstrap simply replaces the unknown frequencies θ with its empirical counterpart $\hat{\theta}$, i.e. $\hat{\theta} = \theta(\hat{P}_n)$.

Consider the rectangular confidence set for the vector of differences $\Delta_I(P)$ defined by

$$C_{l,n}(1-\alpha, I) \equiv \prod_{(j,k)\in I} C_{l,n,j,k}(1-\alpha, I)$$

where $C_{l,n,j,k}(1-\alpha, I)$ could be defined in various ways:

$$C_{\text{lower},n,j,k}(1-\alpha, I) \equiv \left[\hat{\theta}_j - \hat{\theta}_k - c_{\text{lower},n}(1-\alpha, I)\frac{\hat{\sigma}_{j,k}}{\sqrt{n}}, \infty \right). \tag{17}$$

$$C_{\text{upper},n,j,k}(1-\alpha, I) \equiv \left(-\infty, \hat{\theta}_j - \hat{\theta}_k + c_{\text{upper},n}(1-\alpha, I)\frac{\hat{\sigma}_{j,k}}{\sqrt{n}} \right], \tag{18}$$

$$C_{\text{symm},n,j,k}(1-\alpha, I) \equiv \left[\hat{\theta}_j - \hat{\theta}_k \pm c_{\text{symm},n}(1-\alpha, I)\frac{\hat{\sigma}_{j,k}}{\sqrt{n}} \right], \tag{19}$$

$$C_{\text{equi},n,j,k}(1-\alpha, I) \equiv C_{\text{lower},n}\left(1-\frac{\alpha}{2}, I\right) \bigcap C_{\text{upper},n}\left(1-\frac{\alpha}{2}, I\right). \tag{20}$$

The following lemma shows that these bootstrap confidence sets are asymptotically valid in the sense that they cover the true vector of differences with probability approaching $1-\alpha$:

Lemma A.1. *For any $l \in \{\text{lower}, \text{upper}, \text{symm}\}$ and any $I \subset J^2$,*

$$\lim_{n\to\infty} P\left\{\Delta_I(P) \in C_{l,n}(1-\alpha, I)\right\} \geq 1-\alpha$$

with equality if $\theta_j(P) > 0$ for all $j \in J$.

Proof. We begin by considering the case where $\theta_j(P) > 0$ for all $j \in J$. First, consider the joint behavior of $(\sqrt{n}(\hat{\theta}_j - \hat{\theta}_k)/\hat{\sigma}_{j,k})$ for all $\binom{p}{2}$ distinct pairs (j,k) with $j \neq k$. Toward this end, let $J_n(P)$ denote the joint distribution of $\sqrt{n}(\hat{\theta}_1 - \theta_1(P), \ldots, \hat{\theta}_p - \theta_p(P))$. By the multivariate Central Limit Theorem, $J_n(P)$ converges in distribution to $J(P)$, the multivariate normal distribution with mean 0 and covariance matrix $\Sigma = \Sigma(P)$, where Σ has (j,k) entry $\theta_j(P)(1-\theta_j(P))$ if $j = k$ and $-\theta_j(P)\theta_k(P)$ if $j \neq k$. Moreover, in a triangular array setup, if P_n is a sequence of multinomial probabilities with $\theta(P_n) \to \theta(P)$, then $J_n(P_n)$ converges in distribution to $J(P)$. To see why, apply the Cramér–Wold device and the Lindeberg CLT. Since, $\theta(\hat{P}_n) \to \theta(P)$ almost surely (componentwise, by the Strong Law of Large Numbers), it follows that

$$\rho\left(J_n(\hat{P}_n), J_n(P)\right) \to 0 \quad \text{almost surely},$$

where ρ is any metric metrizing weak convergence in $\mathbf{R}^p$. Next, let $J'_n(P)$ denote the joint distribution of

$$\sqrt{n}[(\hat{\theta}_j - \hat{\theta}_k) - (\theta_j(P) - \theta_k(P))]$$

for all $j < k$. So $J'_n(P)$ is a distribution on $\mathbf{R}^{p'}$, where $p' = \binom{p}{2}$. (The pairs can be ordered in any fashion, but for the sake of argument, they are ordered as $(1,2), \ldots (1,p)$ followed by $(2,3), \ldots, (2,p)$, etc.) By the Continuous Mapping Theorem, $J'_n(P)$ converges in distribution to $J'(P)$, the multivariate normal distribution with mean 0 and covariance matrix $\Sigma' = \Sigma'(P)$. Note Σ' can easily be obtained from Σ, but its exact form is not actually required. Again, this convergence is locally uniform in the sense that $J'_n(P_n)$ converges in distribution to $J'(P)$ whenever $\theta(P_n) \to \theta(P)$. Since $\theta(\hat{P}_n) \to \theta(P)$ almost surely, we have

$$\rho'\left(J'_n(\hat{P}_n), J'_n(P)\right) \to 0 \quad \text{almost surely},$$

where ρ' metrizes weak convergence on $\mathbf{R}^{p'}$. Finally, we can consider the joint distribution of studentized differences; to this end, let $J^*_n(P)$ denote the joint distribution of the p' variables

$$\frac{\sqrt{n}[(\hat{\theta}_j - \hat{\theta}_k) - (\theta_j(P) - \theta_k(P))]}{\hat{\sigma}_{j,k}}, \tag{21}$$

where $\hat{\sigma}^2_{j,k}$ is given in (13). Under P, $\theta(\hat{P}_n) \to \theta(P)$ almost surely, and so $\hat{\sigma}^2_{j,k}$ converges almost surely to $\sigma^2_{j,k}(P)$ given by

$$\sigma^2_{j,k}(P) = \theta_j(P)(1-\theta_j(P)) + \theta_k(P)(1-\theta_k(P)) + 2\theta_j(P)\theta_k(P).$$

by a multivariate Slutsky Theorem (or the Continuous Mapping Theorem), $J^*_n(P)$ converges in distribution to $J^*(P)$, the multivariate normal distribution with mean 0 and covariance matrix $\Sigma^* = \Sigma^*(P)$, where Σ^* is easily obtained from Σ' (as Σ^* is the correlation matrix corresponding to the covariance matrix Σ'). Under P_n with $\theta(P_n) \to \theta(P)$, it also follows that $\hat{\sigma}_{j,k}$ converges almost surely to $\sigma_{j,k}(P)$. To see why, first show $\hat{\theta}_j$ converges to θ_j with probability one under P_n; since $\hat{\theta}_j$ can be viewed an average of bounded i.i.d. variables, this convergence follows easily by the well-known 4th moment argument and the Borel–Cantelli Lemma. Hence, under P_n, we also have $J^*(P_n)$ converges in distribution to $J^*(P)$, and then for the same reasons as for J_n and J'_n, we also have

$$\rho'\left(J^*_n(\hat{P}_n), J^*_n(P)\right) \to 0 \quad \text{almost surely}$$

and

$$\rho'\left(J^*_n(\hat{P}_n), J^*(P)\right) \to 0 \quad \text{almost surely}.$$

All of these results carry over if we consider a subset $I \subset J^2$, by the Continuous Mapping Theorem. For example, if $J^*_n(I,P)$ refers to the joint distribution of the variables (21), but only for $(j,k) \in I$, then it follows that

$$\rho'\left(J^*_n(I, \hat{P}_n), J^*_n(I,P)\right) \to 0 \quad \text{almost surely}.$$

Bootstrap consistency of the distributions (14)–(16) now follows by the Continuous Mapping Theorem. Indeed, for any P_n with $\theta(P_n) \to \theta(P)$,

$$L_{\text{lower},n}(x, I, P_n) \to L_{\text{lower},n}(x, I, P) \text{ for all } x,$$

where $L_{\text{lower},n}(\cdot, I, P))$ is the distribution of $\max_{(j,k)\in I} Z_{j,k}$ and $Z = (Z_{j,k} : (j,k) \in I)$ is multivariate normal with distribution $J^*(I, P)$. Note that this distribution is continuous everywhere and strictly increasing. Hence, since $\theta(\hat{P}_n) \to \theta(P)$ almost surely, we also have, for all x,

$$L_{\text{lower},n}(x, I, \hat{P}_n) \to L_{\text{lower},n}(x, I, P) \text{ almost surely .}$$

It also follows that bootstrap quantiles are consistent in the sense that

$$L^{-1}_{\text{lower},n}(1-\alpha, I, \hat{P}_n) \to L^{-1}_{\text{lower},n}(1-\alpha, I, P) \text{ almost surely .}$$

By Slutsky,

$$P\left\{ \max_{(j,k)\in I} \frac{\hat{\theta}_j - \hat{\theta}_k - \Delta_{j,k}(P)}{\hat{\sigma}_{j,k}/\sqrt{n}} \le L^{-1}_{\text{lower},n}(1-\alpha, I, \hat{P}_n) \right\} \to 1-\alpha .$$

By "inverting" this probability statement, we can conclude the intervals

$$\left[\hat{\theta}_j - \hat{\theta}_k - \frac{\hat{\sigma}_{j,k}}{\sqrt{n}} L^{-1}_{\text{lower},n}(1-\alpha, I, \hat{P}_n),\ \infty \right)$$

jointly cover the true $\theta_j(P) - \theta_k(P)$ with asymptotic probability $1-\alpha$. The arguments for upper and two-sided confidence bounds are analogous.

The above argument maintained the assumption that $\theta_j(P) > 0$ for all $j \in J$. We now consider the case where that need not be true. To this end, first suppose that $\theta_j(P) = \theta_k(P) = 0$ for all $(j,k) \in I$. In this case, $\hat{\theta}_j = \hat{\theta}_k = \hat{\sigma}_{j,k} = 0$ for all $(j,k) \in I$. But then, bootstrap samples from $\hat{P}_n$ also satisfy $\hat{\theta}^*_j = \hat{\theta}^*_k = \hat{\sigma}^*_{j,k} = 0$ for all $(j,k) \in I$. Hence, our convention that $0/0 = 0$ and $c/0 = \text{sign}(c)\infty$ for $c \neq 0$, implies that $L^{-1}_{\text{lower},n}(1-\alpha, I, \hat{P}_n) = 0$ w.p.1. It follows that

$$P\{\Delta_I(P) \in C_{\text{lower},n}(1-\alpha, I)\} = 1 .$$

Now suppose that $\theta_j(P) > 0$ or $\theta_k(P) > 0$ for some $(j,k) \in I$. The same argument implies that

$$P\{\theta_j(P) - \theta_k(P) \in C_{\text{lower},n,j,k}(1-\alpha, I)\} = 1$$

for any $(j,k) \in I$ with $\theta_j(P) = \theta_k(P) = 0$. To complete the proof, it suffices to show that

$$\liminf_{n\to\infty} P\{\theta_j(P) - \theta_k(P) \in C_{\text{lower},n,j,k}(1-\alpha, I) \text{ for all } (j,k) \in I^*\} \ge 1-\alpha, \tag{22}$$

where

$$I^* = \{(j,k) \in I : \theta_j(P) > 0 \text{ or } \theta_k(P) > 0\}.$$

Since

$$L^{-1}_{\text{lower},n}(1-\alpha, I, \hat{P}_n) \ge L^{-1}_{\text{lower},n}(1-\alpha, I^*, \hat{P}_n),$$

the desired convergence in (22) can be established simply by arguing as in the first part of the theorem. ■

Appendix B. Proofs of the main results

Proof of Theorem 2.1. We prove the theorem for the two-sided case ($I = J^{\text{two-sided}}$), but the derivations are very similar for the one-sided cases. Define

$$S^- \equiv \bigcup_{j\in J_0} S^-_j \quad \text{with} \quad S^-_j \equiv \{(j,k) \in I : j \neq k \text{ and } \theta_j \le \theta_k\}$$
$$S^+ \equiv \bigcup_{j\in J_0} S^+_j \quad \text{with} \quad S^+_j \equiv \{(j,k) \in I : j \neq k \text{ and } \theta_j \ge \theta_k\}$$

and

$$R^- \equiv \bigcup_{j\in J_0} R^-_j \quad \text{with} \quad R^-_j \equiv \{(j,k) \in I : j \neq k, \text{ reject } H_{k,j}, \text{ and claim } \theta_j < \theta_k\}$$
$$R^+ \equiv \bigcup_{j\in J_0} R^+_j \quad \text{with} \quad R^+_j \equiv \{(j,k) \in I : j \neq k, \text{ reject } H_{j,k}, \text{ and claim } \theta_j > \theta_k\}$$

Suppose $S^- \cap R^+ = \emptyset$ and $S^+ \cap R^- = \emptyset$. Then:

$$\forall j \in J_0 : S^-_j \cap R^+_j = \emptyset \text{ and } S^+_j \cap R^-_j = \emptyset$$

$$\Rightarrow \quad \forall j \in J_0 : \theta_j > \theta_k \ \forall (j,k) \in R_j^+ \text{ and } \theta_j < \theta_k \ \forall (j,k) \in R_j^-$$
$$\Rightarrow \quad \forall j \in J_0 : \theta_j > \theta_k \ \forall k \in \text{Rej}_j^+ \text{ and } \theta_j < \theta_k \ \forall k \in \text{Rej}_j^-$$
$$\Rightarrow \quad \forall j \in J_0 : r_j \leq p - |\text{Rej}_j^+| \text{ and } r_j \geq 1 + |\text{Rej}_j^-|$$

The third implication uses the fact that the number of pairs (j,k) in R_j^+ (or R_j^-) is equal to the number of k in Rej_j^+ (or Rej_j^-). Therefore,

$$P\left\{r_j \in R_{n,j} \ \forall j \in J_0\right\} \geq P\left\{S^- \cap R^+ = \emptyset \text{ and } S^+ \cap R^- = \emptyset\right\} = 1 - FWER_I$$

and the desired result follows from (6). ■

Proof of Theorem 2.2. We first show that the distribution of X_j given $S_{j,k} = s$ is binomial based on s trials and success probability $\theta_j/(\theta_j + \theta_k)$. To see this note that

$$P\{X_j = x | S_{j,k} = s\} \propto P\{X_j = x \text{ and } X_k = s - x\} = P\{X_j = x\} P\{X_k = s - x | X_j = x\},$$

where the symbol $\propto$ means there is a constant out in front that can depend on s and θ_j, θ_k (which we will see depends on θ_j and θ_k through $\theta_j/(\theta_j + \theta_k)$). Continuing,

$$P\{X_j = x | S_{j,k} = s\} \propto \binom{n}{x} \theta_j^x (1-\theta_j)^{n-x} \binom{n-x}{s-x} \left(\frac{\theta_k}{1-\theta_k}\right)^{s-x} \left(\frac{1-\theta_j-\theta_k}{1-\theta_j}\right)^{n-s}$$

$$\propto \binom{s}{x} (\theta_j/\theta_k)^x \propto \binom{s}{x} \left(\frac{\theta_j}{\theta_j+\theta_k}\right)^x \left(\frac{\theta_k}{\theta_j+\theta_k}\right)^{s-x},$$

which is the binomial family of distributions with univariate parameter $\theta_{j,k} = \theta_j/(\theta_j + \theta_k)$. Clearly, a test of $H_{j,k}$ is equivalent to the hypothesis specifying $\theta_{j,k} \leq 1/2$. As is well known, this family of distributions has monotone likelihood ratio. Therefore, by Corollary 3.4.1 of Lehmann and Romano (2022) and conditional on $S_{j,k}$, there exists a UMP level $\beta_{j,k}$ test for testing $H_{j,k}$ given by (7) with constants $\gamma(s)$ and $C(s)$ determined by

$$E_{\theta_k}[\phi(X_j, S_{j,k}) | S_{j,k} = s] = \beta_{j,k} \quad \forall s.$$

Here, $E_{\theta_k}[\cdot]$ refers to the expectation under which $\theta_j = \theta_k$. It is easy to see that the equation determining the constants can be written as in (8), which shows that the test has exact rejection probability (conditional on $S_{j,k}$) equal to $\beta_{j,k}$. By monotone likelihood ratio (and still conditional on $S_{j,k}$), the (conditional) power function of the test is nondecreasing, and so the conditional rejection probability is bounded above by $\beta_{j,k}$ for all θ_j and θ_k satisfying $\theta_{j,k} \leq 1/2$

Thus far, we have shown that $\phi(X_j, S_{j,k})$ is the UMP level $\beta_{j,k}$ test, conditional on $S_{j,k}$. In order to argue that it is UMPU level $\beta_{j,k}$ among all level $\beta_{j,k}$ unbiased tests (unconditionally), consider the boundary of the parameter space $\omega_{j,k} = \{(\theta_1, \ldots, \theta_p);\ \theta_j = \theta_k\}$. The family of distributions of $(X_1, \ldots, X_p)$ still is multinomial, but now $T = (S_{j,k},\ X_i, i \neq j, i \neq k)$ is complete and sufficient for $\omega_{j,k}$. Hence, any test, say ψ, that is similar on $\omega_{j,k}$, i.e., it satisfies that the rejection probability is equal to $\beta_{j,k}$ for all $\theta \in \omega_{j,k}$ or $E_\theta(\psi) = \beta_{j,k}$ for all $\theta \in \omega_{j,k}$, must satisfy that it is conditionally level $\beta_{j,k}$ given T or $E_\theta(\psi|T) = \beta_{j,k}$ for all $\theta \in \omega_{j,k}$. In other words, all similar tests have Neyman structure, by Theorem 4.3.2 in Lehmann and Romano (2022). Therefore, the optimal unconditional level $\beta_{j,k}$ test must be obtained by finding the optimal conditional level $\beta_{j,k}$, conditional on T. Finally, note that specifying the conditional distribution of the data $(X_1, \ldots, X_p)$ given T is equivalent to specifying the conditional distribution of X_j given $S_{j,k}$, since conditionally all other X_i with $i \neq j$ and $i \neq k$ are now fixed. But, we have found the optimal conditional test above based on the conditional distribution of X_j given $S_{j,k}$. (Note that we could have argued by writing the family of distributions in the canonical multiparameter exponential form discussed in Section 4.4 of Lehmann and Romano (2022), but the notation becomes messy, stemming from the fact that the rank of the multinomial family is $p - 1$ and not p, and consequently the argument gets obscured.) ■

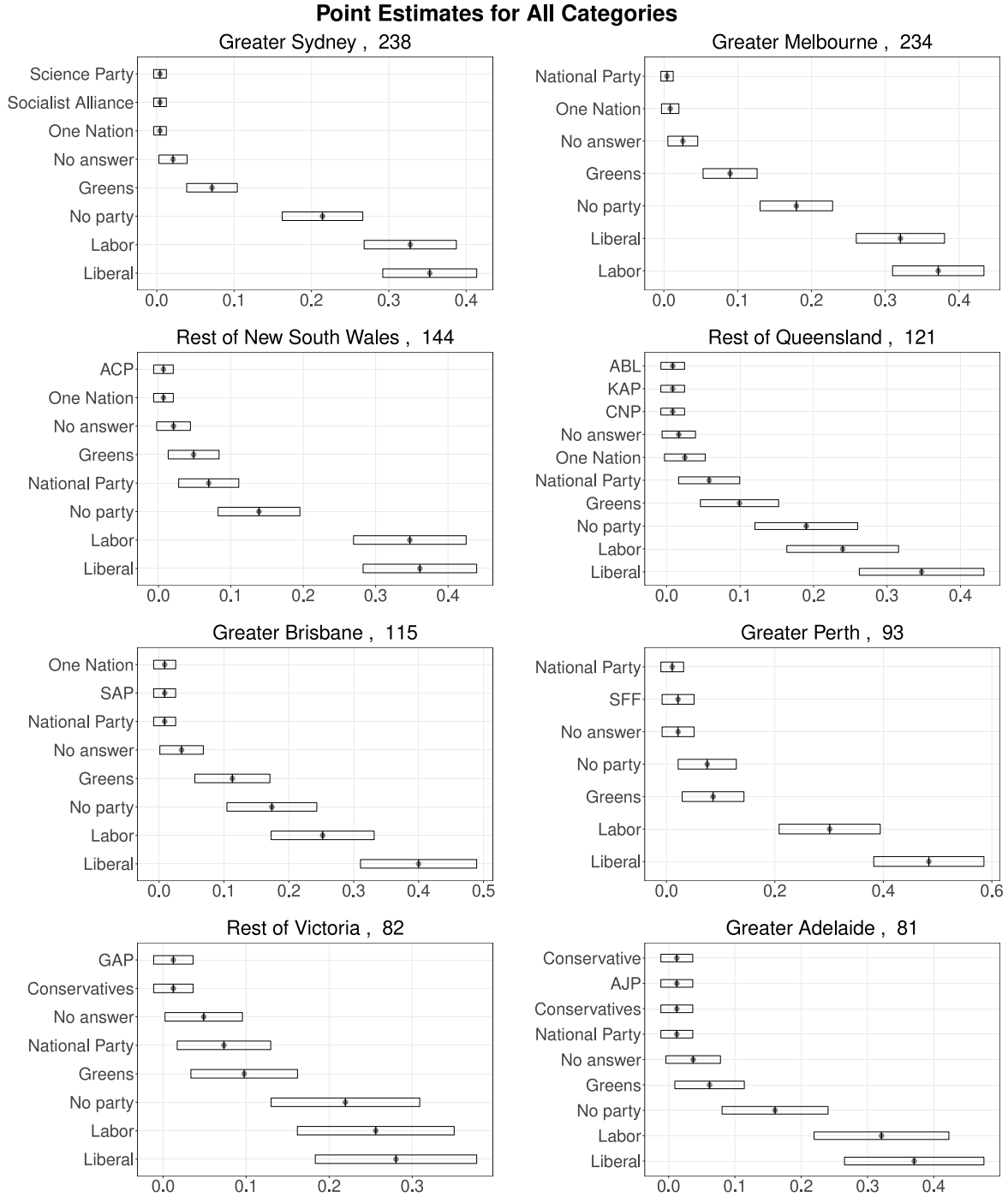
Proof of Theorem 3.1. First, by an argument analogous to the one in Theorem 3.3 in Mogstad et al. (2024),

$$P\left\{r_j \in R_{n,j}^{\text{boot}} \ \forall j \in J_0\right\} \geq P\left\{\Delta_I \in C_{\text{lower},n}(1-\alpha, I)\right\},$$

where I is equal to one of the sets J^{lower}, J^{upper}, or $J^{\text{two-sided}}$, which was used to construct R_n^{boot}. The desired result therefore follows from Lemma A.1. ■

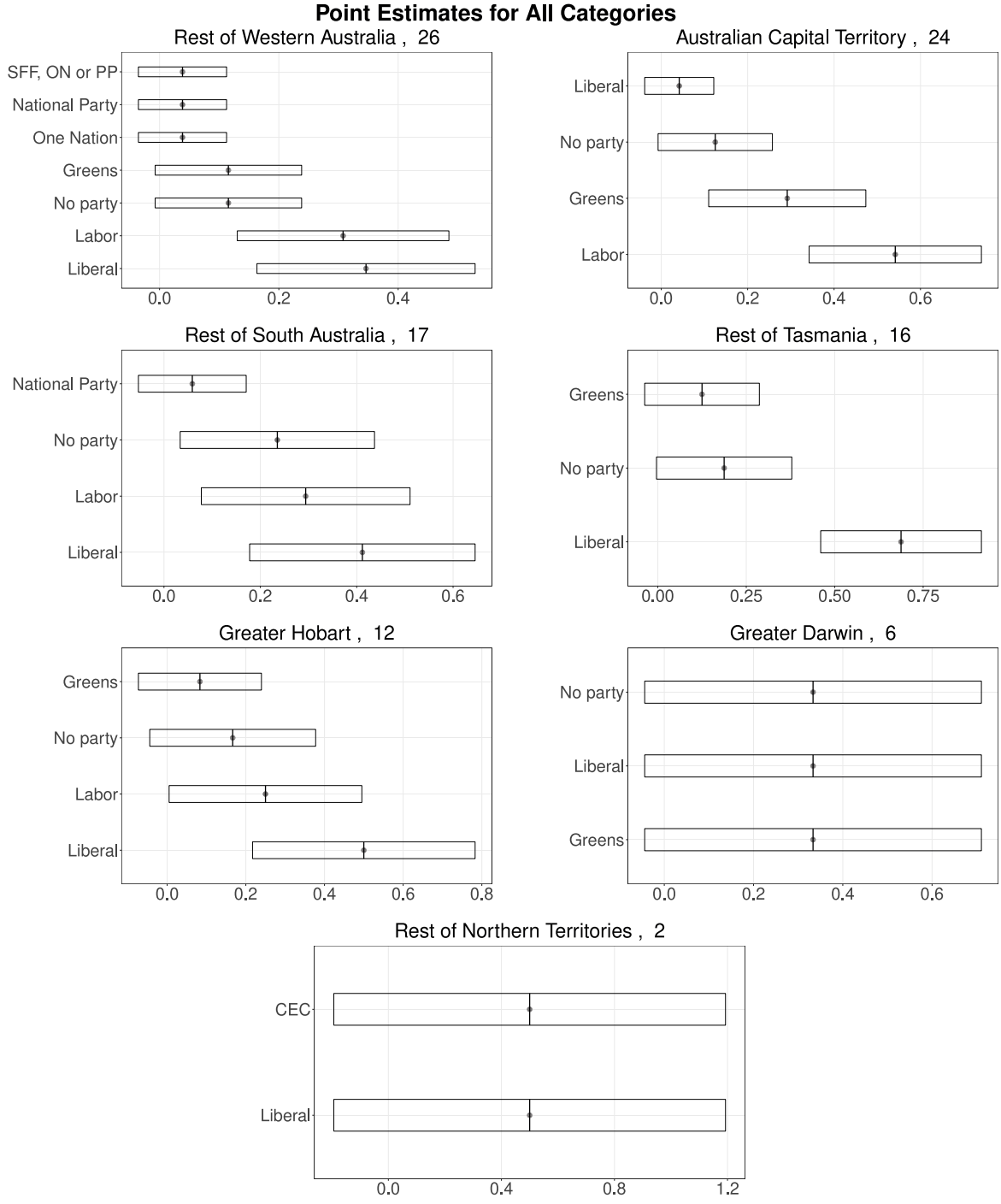
Appendix C. Supporting results for the empirical application

See Figs. 7–9.



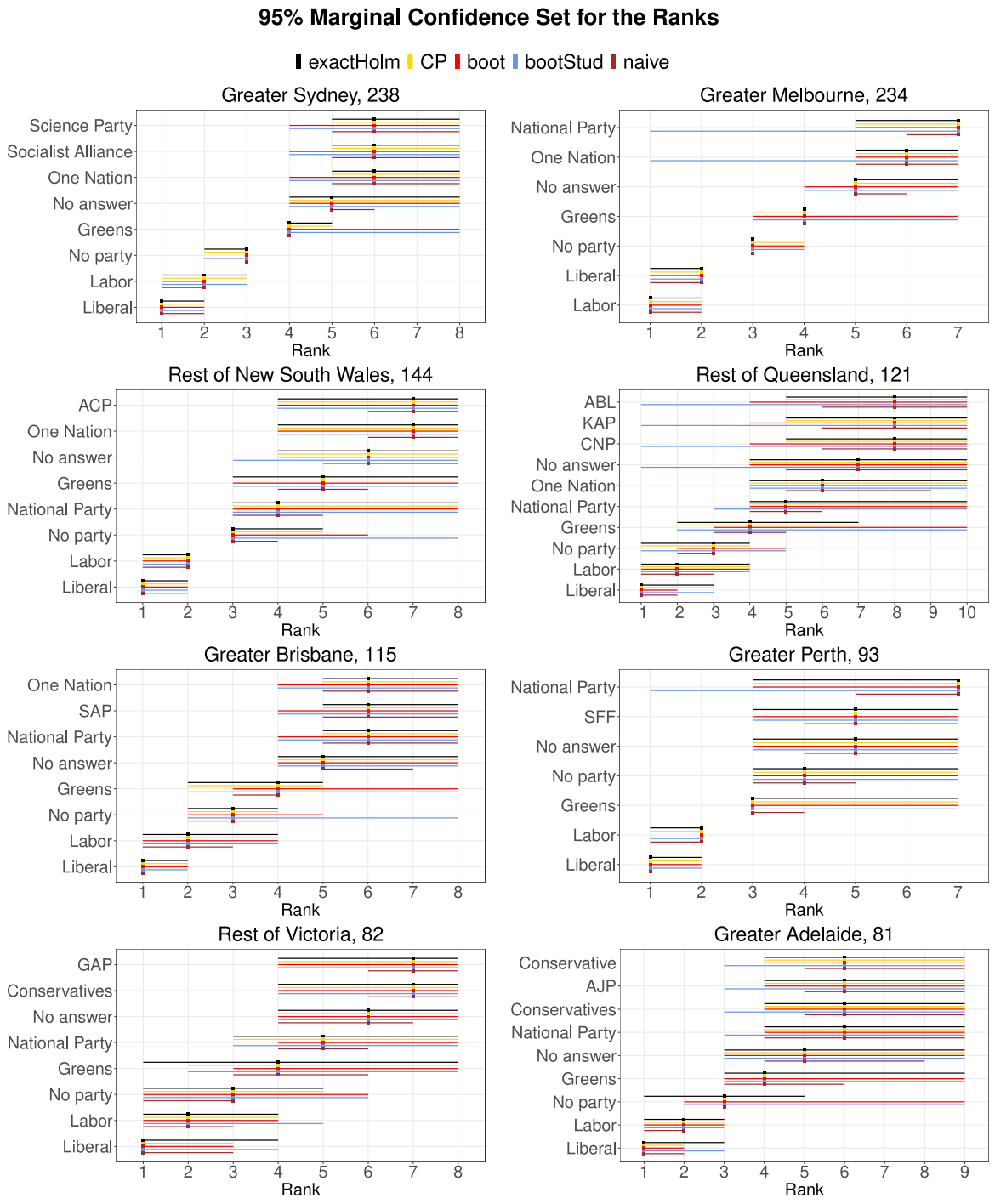
(a)

Fig. 7. Point estimates of categories support shares in fifteen Australian territories from AES2019 and ±1.96se.



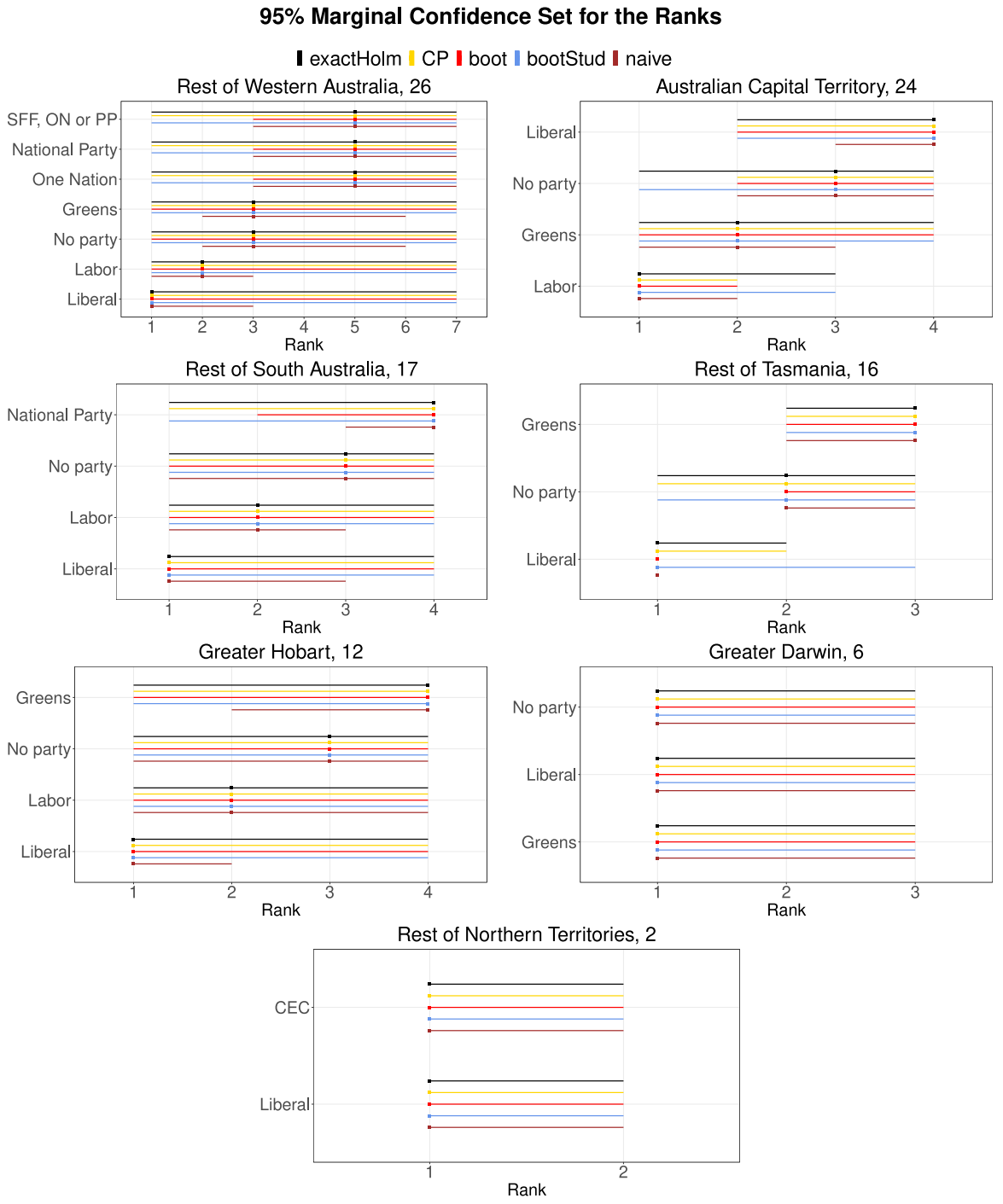
(b)

Fig. 7. (*continued*).



(a)

Fig. 8. 95% marginal confidence sets for the ranks of categories in fifteen Australian territories ranked by their support share in AES2019. Each panel shows the confidence sets for the ranks computed by five methods for each party.



(b)

Fig. 8. (*continued*).

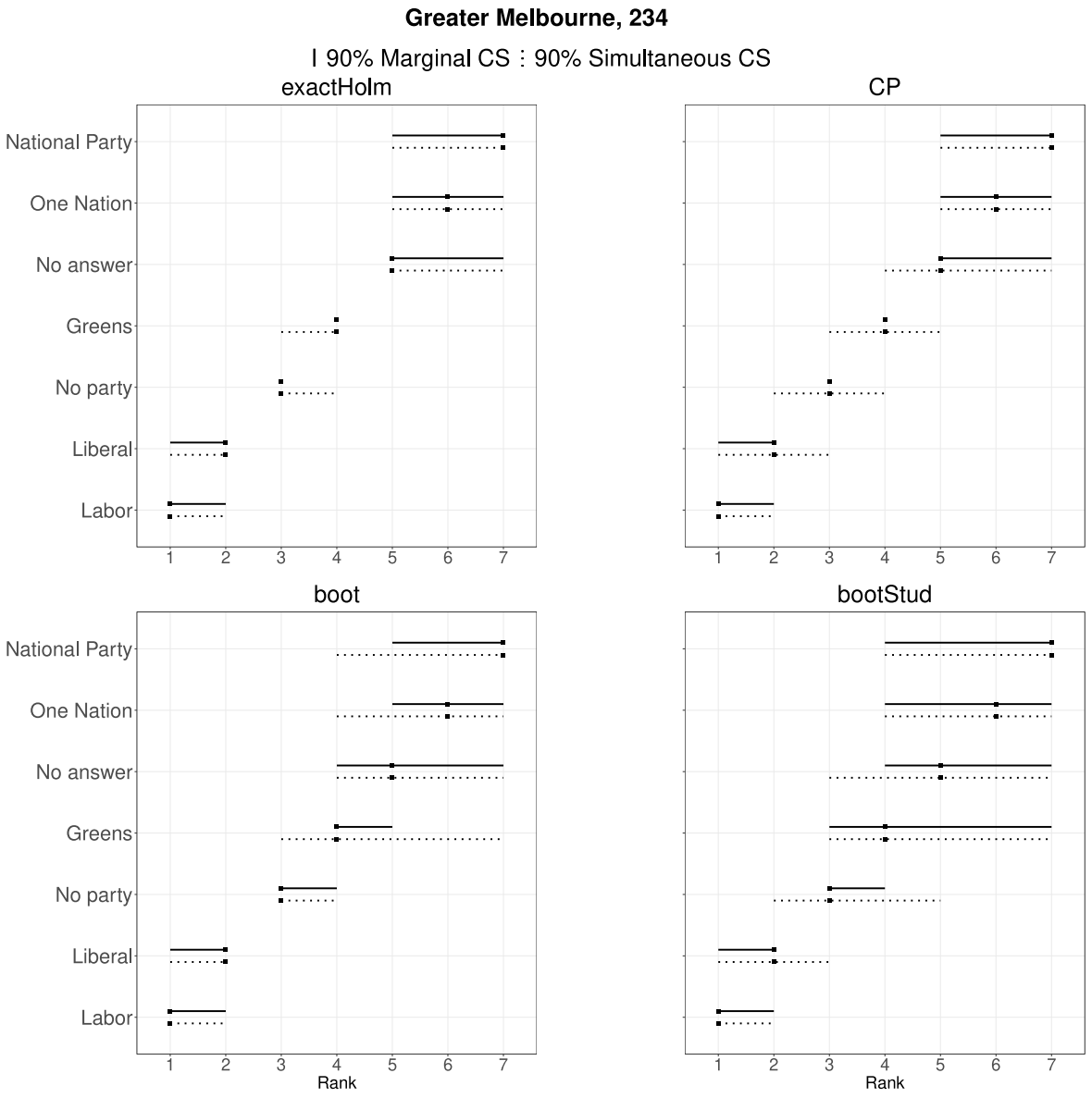


Fig. 9. 90% marginal and 90% simultaneous confidence sets for the ranks of categories in Greater Melbourne.

References

Andrews, I., Kitagawa, T., McCloskey, A., 2018. Inference on winners. Working Paper CWP 31/18, CeMMAP.

Bean, C., Cameron, S., Gibson, R., Makkai, T., McAllister, I., Sheppard, J., 2019. Australian election study 2019: Voters technical report. The Australian National University.

Brown, L.D., Cai, T.T., DasGupta, A., 2001. Interval estimation for a binomial proportion. Statist. Sci. 16 (2), 101–117.

Cameron, S., McAllister, I., 2019. The 2019 Australian federal election: Results from the Australian election study. The Australian National University.

Clopper, C.J., Pearson, E.S., 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26 (4), 404–413.

Goldstein, H., Spiegelhalter, D.J., 1996. League tables and their limitations: Statistical issues in comparisons of institutional performance. J. R. Stat. Soc. Ser. A 159 (3), 385–443.

Gu, J., Koenker, R., 2020. Invidious comparisons: Ranking and selection as compound decisions. arXiv preprint arXiv:2012.12550.

Gupta, S.S., Panchapakesan, S., 1979. Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations. John Wiley & Sons, New York.

Hall, P., Miller, H., 2009. Using the bootstrap to quantify the authority of an empirical ranking. Ann. Statist. 37 (6B), 3929–3959.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

Klein, M., Wright, T., Wieczorek, J., 2020. A joint confidence region for an overall ranking of populations. J. R. Stat. Soc. Ser. C. Appl. Stat. 69 (3), 589–606.

Lehmann, E.L., Romano, J.P., 2022. Testing Statistical Hypotheses, fourth ed. Springer Cham.

Marden, J.I., 1995. Analyzing and Modeling Rank Data, first ed. Chapman & Hall/CRC.
Mogstad, M., Romano, J.P., Shaikh, A.M., Wilhelm, D., 2024. Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries. Rev. Econ. Stud. 91 (1), 476–518.
Romano, J.P., Shaikh, A.M., 2012. On the uniform asymptotic validity of subsampling and the bootstrap. Ann. Stat. 40 (6), 2798–2822.
Romano, J.P., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73 (4), 1237–1282.
Xie, M., Singh, K., Zhang, C.-H., 2009. Confidence intervals for population ranks in the presence of ties and near ties. J. Amer. Statist. Assoc. 104 (486), 775–788.