ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2020)
69, *Part* 3, *pp.* 589–606

A joint confidence region for an overall ranking of populations

Martin Klein,

US Food and Drug Administration, Silver Spring, USA

Tommy Wright

US Census Bureau, and Georgetown University, Washington DC, USA

and Jerzy Wieczorek

Colby College, Waterville, USA

[Received May 2019. Final revision January 2020]

Summary. National statistical agencies lack statistical methodology to express uncertainty in their released estimated overall rankings. For example, the US Census Bureau produced an 'explicit' ranking of the states based on observed sample estimates during 2011 of mean travel time to work. Current literature provides measures of uncertainty in estimated individual ranks, but not a direct measure of uncertainty for the estimated overall ranking. We construct and visualize a joint confidence region for the true unknown overall ranking that provides a measure of uncertainty in the estimated overall ranking.

Keywords: Bonferroni correction; Independence; Official statistics; Ranking; Visualization

1. Introduction

We present a simple novel measure of uncertainty for an estimated overall ranking by constructing a joint confidence region for the true unknown overall ranking as follows: by observing how K known confidence intervals for K means overlap or not, by obtaining a confidence set containing each population rank, and ultimately obtaining the joint confidence region for the overall ranking.

Rankings (explicit or implicit) of $K \geqslant 2$ populations or governmental units based on sample survey data are usually released without direct statistical statements of uncertainty on estimated overall rankings. Our main objective is to provide simple statistical methodology for expressing uncertainty in released overall rankings based on data from sample surveys by statistical agencies. A visualization facilitates communication with wide audiences.

Formally, assume K disjoint populations with associated independent continuous random variables $Y_1, \ldots, Y_K$ and respective cumulative distribution functions $F_1(y), \ldots, F_K(y)$. Let θ_k be a real-valued characteristic (parameter) related to $F_k(y)$, for $k = 1, \ldots, K$. Although the values of $\theta_1, \ldots, \theta_K$ are unknown, it is desired to rank the K populations from smallest to largest on the basis of these unknown values, i.e. based on

Address for correspondence: Tommy Wright, Center for Statistical Research and Methodology, US Bureau of the Census, 4600 Silver Hill Road, Washington DC 20233, USA.
E-mail: tommy.wright@census.gov

0035–9254/20/69589

$$\theta_{(1)} < \theta_{(2)} < \ldots < \theta_{(k)} < \ldots < \theta_{(K)}. \quad (1)$$

If $Y_{k1}, \ldots, Y_{kn_k}$ is a probability sample of size n_k from the kth population where the statistic $\hat{\theta}_k = \hat{\theta}_k(Y_{k1}, \ldots, Y_{kn_k})$ is an estimator of θ_k for $k = 1, \ldots, K$, we rank the K populations on the basis of the observed ranking of the values, $\hat{\theta}_1, \ldots, \hat{\theta}_K$, i.e.

$$\hat{\theta}_{(1)} < \hat{\theta}_{(2)} < \ldots < \hat{\theta}_{(k)} < \ldots < \hat{\theta}_{(K)}. \quad (2)$$

For example, data from the US Census Bureau's American Community Survey (ACS) produced an *explicit ranking* of the $K = 51$ states (including Washington DC) based on observed sample estimates during 2011 of θ_k the mean travel time to work (in minutes) for workers 16 years old and over who did not work at home (henceforth 'mean travel time to work') for state k, where $k = 1, \ldots, 51$. Also, given estimates in a table without an explicit ranking, users tend to compare states by looking for smallest or largest estimates and for relative standings between the states. We refer to such tables as motivating *'implicit' rankings*.

Because rankings based on the observed values of $\hat{\theta}_1, \ldots, \hat{\theta}_K$ can vary because of sampling variability, *widely understood* and *robust* (valid and applicable in many situations) statements of uncertainty should accompany each released ranking.

In this paper, a collection of joint confidence intervals for $\theta_1, \ldots, \theta_K$ forms the basis for the measure that is presented. Knowledge of the specific complex sampling design and estimation methodology for each population is not required. In Section 2, we present a simple mathematical result. Section 3 uses this mathematical result to provide general theory for constructing a joint confidence region for the overall ranking. Examples using the ACS's travel time to work data are given in Section 4. Section 5 gives discussion.

1.1. Overview of the American Community Survey

Conducted by the US Census Bureau, the ACS's sampling design is basically a national stratified random sample with sampling and estimation following a finite population design-based framework. Data are collected continually throughout the year. The ACS provides data every year for the preceding calendar year and for the preceding 5-year period—giving communities current information that is needed to plan investments and services. The sample survey generates data that help to determine how hundreds of billions of dollars in federal and state funds are distributed each year. Currently, over 3500000 housing unit addresses are contacted each year by Internet, mail, telephone or face to face to provide data for statistical estimates at various geographic levels—large and small. In addition to travel time to work, the ACS questionnaire asks about age, sex, race, family and relationships, income and benefits, health insurance, education, veteran status, disabilities, where people work and how they get there, and where they live and how much they pay for some essentials. (See `https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.html`.)

1.2. Travel time to work

There is considerable interest each year from the public, the media, city planners and engineers who study traffic patterns when travel time to work rankings are released by the Census Bureau. Increases in the travel times to work have been observed over time. A *Washington Post* article (Siddiqui, 2018) noted that

> 'There's a massive body of social science and public health research on the negative effects of commuting on personal and social well-being. Longer commutes are linked with increased rates of obesity, high

> cholesterol, high blood pressure, back and neck pain, divorce, depression, and death. At the societal level, people who commute more are less likely to vote. They're more likely to be absent from work. They're less likely to escape poverty. They have kids who are more likely to have emotional problems.'

Rankings help to identify factors and living conditions of areas with lower ranks for travel time to work for possible adoption by areas with higher ranks.

Among 'ranking tables' based on many topics using data collected by the ACS for 2011 is Table 1. From Table 1, the 51 states (including Washington DC) are ranked from largest to smallest by estimated mean travel time to work, $\hat{\theta}_k$ (see `https://factfinder.census.gov/bkmk/table/1.0/en/ACS/11_1_YR/R0801.US01PRF`). From the 'statistical significance' column, we see the results of 50 separate tests of significance ($\alpha = 0.1$) for Alabama as the selected state with each of the other states. Alabama is not statistically significantly different from Tennessee, Michigan, Nevada, Mississippi, South Carolina or Rhode Island.

In Table 1, the margin of error MOE_k gives uncertainty in the estimate $\hat{\theta}_k$ for each state separately; and the tests of significance compare one state's estimate with those of each of the other states. However, a direct assessment of the uncertainty in the estimated overall ranking would involve all the states simultaneously and their relative standing to each other. The (estimated) rank of a state (e.g. of Alabama) is informed by data from all states. No direct measure of uncertainty is given for estimated individual ranks. We seek to provide a measure of uncertainty that is directly focused on the estimated overall ranking.

Although we illustrate our method by using all 51 states to be consistent with the published ACS ranking tables, the same method could be applied to data subsets that are specified *a priori* by analysts who are interested in subsets of states or other domains (e.g. rankings of states in the north-east, rankings of states with low urbanization or rankings of large metropolitan areas).

1.3. Selected ranking literature

In a seminal paper from the ranking and selection literature, Bechhofer (1954) presented a procedure for computing sample sizes n_k for ranking K populations where the ranking is based on the observed sample means. Assuming the usual Bayesian set-up of priors on the parameters θ_k, the focus is on how to go from posteriors on the parameters θ_k to a ranking of the parameters. The literature (Shen and Louis, 1998) suggested that 'Ranking posterior means can perform poorly ...'. Others who have provided similar reports include Brand *et al.* (1996), Goldstein and Spiegelhalter (1996), Laird and Louis (1989) and Morris and Christiansen (1986). Govindarajulu and Harvey (1974) pointed out that simply choosing the ranking with the highest posterior probability may not be an ideal approach, even if it were possible. Louis (1984) argued that any ranking of populations based on θ_k should consider the collection or ensemble $\{\theta_1, \theta_2, \ldots, \theta_K\}$ and not the θ_k individually; we agree. Also see Klein and Wright (2011). Goldstein and Spiegelhalter (1996) suggested the bootstrap as a means of obtaining interval estimates for ranks, as did Hall and Miller (2009) and Wright *et al.* (2013, 2014, 2019).

Bayesian methods (e.g. Laird and Louis (1989) and Shen and Louis (1998)) tend to provide measures of uncertainty in estimated individual ranks; they do not provide measures of uncertainty in the estimated overall ranking (see Section 5). Stated another way, existing Bayesian methods seem to answer the question 'How good is the estimated individual rank for a specific state?'. We seek to answer the question 'How good is the estimated overall ranking for all the states?', i.e. the estimated overall ranking depends on the observed sample for each state, and these samples all have sampling error. Other samples could result in alternative estimated overall rankings. Thus we seek a quantification of this uncertainty via a joint confidence region for the overall ranking (Sections 3 and 4).

Table 1. Mean travel time to work of workers 16 years old and over who did not work at home†

Rank	*Geographical area*	*Statistical significance?*	*Estimated mean (min)*	*Margin of error*
	USA		25.5	±0.1
51	Maryland		32.2	±0.2
50	New York		31.5	±0.2
49	New Jersey		30.5	±0.2
48	District of Columbia		30.1	±0.5
47	Illinois		28.2	±0.2
46	Massachusetts		28.0	±0.2
45	Virginia		27.7	±0.2
44	California		27.1	±0.1
44	Georgia		27.1	±0.3
42	New Hampshire		26.9	±0.5
41	Pennsylvania		25.9	±0.1
40	Florida		25.8	±0.2
39	Hawaii		25.7	±0.4
38	West Virginia		25.6	±0.5
37	Washington		25.5	±0.2
36	Delaware		25.3	±0.6
35	Connecticut		25.0	±0.3
34	Arizona		24.8	±0.2
34	Texas		24.8	±0.1
32	Colorado		24.5	±0.3
32	Louisiana		24.5	±0.2
30	Tennessee	‡	24.2	±0.2
29	Michigan	‡	24.1	±0.2
29	Nevada	‡	24.1	±0.4
27	*Alabama*	*§*	*23.9*	*±0.2*
27	Mississippi	‡	23.9	±0.4
25	South Carolina	‡	23.6	±0.3
24	Indiana		23.5	±0.2
23	Maine		23.4	±0.4
23	North Carolina		23.4	±0.2
23	Rhode Island	‡	23.4	±0.5
20	Missouri		23.1	±0.2
20	Ohio		23.1	±0.1
18	Minnesota		23.0	±0.2
17	Kentucky		22.9	±0.2
16	Oregon		22.5	±0.3
15	Vermont		21.9	±0.5
15	Wisconsin		21.9	±0.2
13	Utah		21.6	±0.3
12	New Mexico		21.4	±0.4
11	Arkansas		21.3	±0.4
10	Oklahoma		21.1	±0.2
9	Idaho		19.7	±0.4
8	Kansas		18.9	±0.3
7	Iowa		18.8	±0.2
6	Alaska		18.4	±0.5
5	Montana		18.2	±0.5
4	Nebraska		18.1	±0.3
4	Wyoming		18.1	±0.8
2	North Dakota		16.9	±0.6
2	South Dakota		16.9	±0.5

†Source: 2011 1-year ACS, ranking table R0801, US Census Bureau. For more information on the ACS, see `https://www.census.gov/programs-surveys/acs/`.
‡Indicates when an estimate is not statistically significantly different from the estimate for the selected state (Alabama).
§Indicates that the selected state is being compared with each of the other 50 states.

The primary objective in this paper is to present a frequentist joint confidence region for the overall ranking whose coverage probability has a guaranteed lower bound. The approach proposed does not require intensive computations that are often needed by Bayesian or bootstrap methods. However, remark 8 of Section 5 shows how our method can also be easily applied for Bayesian inference.

2. Main result

One could imply uncertainty in an estimated ranking (2) through confidence intervals and hypothesis tests for individual parameters θ_ks, and for the pairwise differences $\theta_k - \theta_{k'}$ (e.g. Wright *et al.* (2019)). This is the approach that is currently taken by the Census Bureau's ACS and illustrated earlier with Table 1. However, this approach does not provide a direct measure of uncertainty for the estimated individual ranks nor the estimated overall ranking. Alternatively, one may consider the individual ranks as the parameters of interest, and inferences can be drawn on them and the overall ranking directly. The unknown true ranks are denoted by $r_1, \ldots, r_K$, and they are defined such that the population with the smallest θ_k has rank 1, the population with the second smallest θ_k has rank 2, and so on. Formally, we define the rank for the kth population as

$$r_k = \sum_{j=1}^{K} I(\theta_j \leqslant \theta_k) = 1 + \sum_{j:j\neq k} I(\theta_j \leqslant \theta_k), \qquad \text{for } k = 1, \ldots, K. \tag{3}$$

The estimated overall ranking, computed on the basis of the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_K$, is denoted by $(\hat{r}_1, \ldots, \hat{r}_K)$, where

$$\hat{r}_k = 1 + \sum_{j:j\neq k} I(\hat{\theta}_j \leqslant \hat{\theta}_k), \qquad \text{for } k = 1, 2, \ldots, K. \tag{4}$$

Naturally, uncertainty in the estimators $\hat{\theta}_1, \ldots, \hat{\theta}_K$ is propagated to the estimated ranking. An understandable measure of uncertainty should accompany a released overall ranking.

Although the values of $\theta_1, \ldots, \theta_K$ are unknown, suppose that for each $k \in \{1, 2, \ldots, K\}$ we know real numbers $L_k < U_k$ such that

$$\theta_k \in (L_k, U_k), \tag{5}$$

i.e. although each θ_k is unknown, we do know that θ_k is contained in the open interval (L_k, U_k).

For deriving a confidence region for the ranking, there will be no loss of generality in assumption (5) because, when we construct the confidence region in Section 3, we shall replace the intervals in expression (5) with joint confidence intervals and the main result will then be used to obtain a probability statement.

For each $k \in \{1, 2, \ldots, K\}$, define

$$\left.\begin{aligned} I_k &= \{1, 2, \ldots, K\} \backslash \{k\}, \\ \Lambda_{\mathrm{L}k} &= \{j \in I_k : U_j \leqslant L_k\}, \\ \Lambda_{\mathrm{R}k} &= \{j \in I_k : U_k \leqslant L_j\}, \\ \Lambda_{\mathrm{O}k} &= \{j \in I_k : U_j > L_k \text{ and } U_k > L_j\} = I_k \backslash (\Lambda_{\mathrm{L}k} \cup \Lambda_{\mathrm{R}k}). \end{aligned}\right\} \tag{6}$$

For each $k \in \{1, 2, \ldots, K\}$, and $j \in I_k$:

(a) $j \in \Lambda_{\mathrm{L}k}$ if and only if $(L_j, U_j) \cap (L_k, U_k) = \emptyset$ and (L_j, U_j) lies to the left of (L_k, U_k);
(b) $j \in \Lambda_{\mathrm{R}k}$ if and only if $(L_j, U_j) \cap (L_k, U_k) = \emptyset$ and (L_j, U_j) lies to the right of (L_k, U_k);

(c) $j \in \Lambda_{Ok}$ if and only if $(L_j, U_j) \cap (L_k, U_k) \neq \emptyset$.

It follows that Λ_{Lk}, Λ_{Rk} and Λ_{Ok} are mutually exclusive, and $\Lambda_{Lk} \cup \Lambda_{Rk} \cup \Lambda_{Ok} = I_k$. For a finite set A, let $|A|$ denote the number of elements in A.

2.1. Main result

Under the scenario that was described above, it follows that, for each $k \in \{1, 2, \ldots, K\}$,

$$r_k \in \{|\Lambda_{Lk}| + 1, |\Lambda_{Lk}| + 2, |\Lambda_{Lk}| + 3, \ldots, |\Lambda_{Lk}| + |\Lambda_{Ok}| + 1\}. \tag{7}$$

Proof. Let $k \in \{1, 2, \ldots, K\}$. Because Λ_{Lk}, Λ_{Rk} and Λ_{Ok} are mutually exclusive, and $\Lambda_{Lk} \cup \Lambda_{Rk} \cup \Lambda_{Ok} = I_k$, we can write the rank of the kth population as

$$\begin{aligned} r_k &= 1 + \sum_{j:j\neq k} I(\theta_j \leqslant \theta_k) = 1 + \sum_{j \in I_k} I(\theta_j \leqslant \theta_k) \\ &= 1 + \sum_{j \in \Lambda_{Lk}} I(\theta_j \leqslant \theta_k) + \sum_{j \in \Lambda_{Rk}} I(\theta_j \leqslant \theta_k) + \sum_{j \in \Lambda_{Ok}} I(\theta_j \leqslant \theta_k). \end{aligned} \tag{8}$$

We note that $j \in \Lambda_{Lk} \Rightarrow U_j \leqslant L_k \Rightarrow L_j < \theta_j < U_j \leqslant L_k < \theta_k < U_k \Rightarrow I(\theta_j \leqslant \theta_k) = 1$, and $j \in \Lambda_{Rk} \Rightarrow U_k \leqslant L_j \Rightarrow L_k < \theta_k < U_k \leqslant L_j < \theta_j < U_j \Rightarrow I(\theta_j \leqslant \theta_k) = 0$; and therefore, continuing from equation (8), we have

$$\begin{aligned} r_k &= 1 + \sum_{j \in \Lambda_{Lk}} I(\theta_j \leqslant \theta_k) + \sum_{j \in \Lambda_{Rk}} I(\theta_j \leqslant \theta_k) + \sum_{j \in \Lambda_{Ok}} I(\theta_j \leqslant \theta_k) \\ &= 1 + \sum_{j \in \Lambda_{Lk}} 1 + \sum_{j \in \Lambda_{Rk}} 0 + \sum_{j \in \Lambda_{Ok}} I(\theta_j \leqslant \theta_k) \\ &= 1 + |\Lambda_{Lk}| + \sum_{j \in \Lambda_{Ok}} I(\theta_j \leqslant \theta_k). \end{aligned}$$

Because $\Sigma_{j \in \Lambda_{Ok}} I(\theta_j \leqslant \theta_k) \in \{0, 1, \ldots, |\Lambda_{Ok}|\}$ it follows that

$$r_k = 1 + |\Lambda_{Lk}| + \sum_{j \in \Lambda_{Ok}} I(\theta_j \leqslant \theta_k) \in \{|\Lambda_{Lk}| + 1, |\Lambda_{Lk}| + 2, |\Lambda_{Lk}| + 3, \ldots, |\Lambda_{Lk}| + |\Lambda_{Ok}| + 1\}.$$

This completes the proof.

2.2. Choice of L_k and U_k

As can be seen explicitly from the main result (7), smaller values of $|\Lambda_{Ok}|$ imply a smaller set size for r_k in expression (7), which we desire. Smaller values of $|\Lambda_{Ok}|$ tend to follow from smaller differences $U_k - L_k$. Hence, we want to choose real numbers L_k and U_k that are near each other so that the intervals in expression (5) are as short as possible subject to the constraint $\theta_k \in (L_k, U_k)$.

In the main result, we assume that L_k and U_k are fixed constants. In Section 3, we take L_k and U_k as the end points of a confidence interval for θ_k. Hence, in Section 3, L_k and U_k are random variables (i.e. statistics).

3. Joint confidence region for an overall ranking

Assume that $\{(L_1, U_1), (L_2, U_2), \ldots, (L_K, U_K)\}$ is a collection of confidence intervals for the unknown parameters $\theta_1, \theta_2, \ldots, \theta_K$ respectively, and the joint coverage probability of these intervals is greater than or equal to $1 - \alpha$, i.e. we assume that

$$P\left[\bigcap_{k=1}^{K}\{\theta_k \in (L_k, U_k)\}\right] \geqslant 1-\alpha.$$

In this setting, $L_1, L_2, \ldots, L_K, U_1, U_2, \ldots, U_K$ are random variables. By the main result,

$$\bigcap_{k=1}^{K}\{\theta_k \in (L_k, U_k)\} \Rightarrow \bigcap_{k=1}^{K}[r_k \in \{|\Lambda_{Lk}|+1, |\Lambda_{Lk}|+2, |\Lambda_{Lk}|+3, \ldots, |\Lambda_{Lk}|+|\Lambda_{Ok}|+1\}],$$

where, for each $k \in \{1, 2, \ldots, K\}$, r_k is the rank that is defined in equation (3), and Λ_{Lk} and Λ_{Ok} are as defined in expression (6). Therefore, it follows that

$$P\left(\bigcap_{k=1}^{K}[r_k \in \{|\Lambda_{Lk}|+1, |\Lambda_{Lk}|+2, |\Lambda_{Lk}|+3, \ldots, |\Lambda_{Lk}|+|\Lambda_{Ok}|+1\}]\right) \geqslant P\left[\bigcap_{k=1}^{K}\{\theta_k \in (L_k, U_k)\}\right] \geqslant 1-\alpha.$$

Thus we have shown that

$$\{(r_1, \ldots, r_K) : r_k \in \{|\Lambda_{Lk}|+1, |\Lambda_{Lk}|+2, |\Lambda_{Lk}|+3, \ldots, |\Lambda_{Lk}|+|\Lambda_{Ok}|+1\} \text{ for } k = 1, \ldots, K\} \quad (9)$$

is a *joint confidence region* (or set) for the overall ranking $(r_1, \ldots, r_K)$ having joint coverage probability of at least $1-\alpha$.

A natural question to ask regarding the joint confidence region for the overall ranking is as follows: 'Is the estimated ranking $(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_K)$ contained in the joint confidence region for the overall ranking (9)?'. The following result gives a condition on the joint confidence intervals $(L_1, U_1), \ldots, (L_K, U_K)$, that will ensure that the estimated ranking is in fact contained in the joint confidence region for the overall ranking. The result states that, if $(L_1, U_1), \ldots, (L_K, U_K)$ are constructed such that the estimator $\hat{\theta}_k \in (L_k, U_k)$ for all $k \in \{1, 2, \ldots, K\}$ with probability 1 (which is so with our approach), then the estimated ranking $(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_K)$ is contained in the joint confidence region (9) with probability 1.

Result 1. If $P[\bigcap_{k=1}^{K}\{\hat{\theta}_k \in (L_k, U_k)\}] = 1$, then

$$P\left(\bigcap_{k=1}^{K}[\hat{r}_k \in \{|\Lambda_{Lk}|+1, |\Lambda_{Lk}|+2, |\Lambda_{Lk}|+3, \ldots, |\Lambda_{Lk}|+|\Lambda_{Ok}|+1\}]\right) = 1.$$

Proof. If the observed values of $L_1, \ldots, L_K, U_1, \ldots, U_K, \hat{\theta}_1, \ldots, \hat{\theta}_K$ are such that $\hat{\theta}_k \in (L_k, U_k)$ for all $k \in \{1, \ldots, K\}$, then an argument similar to that used in the proof of the main result gives

$$\begin{aligned}\hat{r}_k &= 1 + \sum_{j:j\neq k} I(\hat{\theta}_j \leqslant \hat{\theta}_k) \\ &= 1 + |\Lambda_{Lk}| + \sum_{j\in\Lambda_{Ok}} I(\hat{\theta}_j \leqslant \hat{\theta}_k) \\ &\in \{|\Lambda_{Lk}|+1, |\Lambda_{Lk}|+2, |\Lambda_{Lk}|+3, \ldots, |\Lambda_{Lk}|+|\Lambda_{Ok}|+1\},\end{aligned}$$

for all $k \in \{1, \ldots, K\}$. Thus we have established that

$$\bigcap_{k=1}^{K}\{\hat{\theta}_k \in (L_k, U_k)\} \Rightarrow \bigcap_{k=1}^{K}[\hat{r}_k \in \{|\Lambda_{\mathrm{L}k}|+1, |\Lambda_{\mathrm{L}k}|+2, |\Lambda_{\mathrm{L}k}|+3, \ldots, |\Lambda_{\mathrm{L}k}|+|\Lambda_{\mathrm{O}k}|+1\}]$$

and, therefore,

$$P\left(\bigcap_{k=1}^{K}[\hat{r}_k \in \{|\Lambda_{\mathrm{L}k}|+1, |\Lambda_{\mathrm{L}k}|+2, |\Lambda_{\mathrm{L}k}|+3, \ldots, |\Lambda_{\mathrm{L}k}|+|\Lambda_{\mathrm{O}k}|+1\}]\right) \geqslant P\left[\bigcap_{k=1}^{K}\{\hat{\theta}_k \in (L_k, U_k)\}\right] = 1.$$

Hence the result follows. □

In general, the joint confidence region in expression (9) contains more than one possible overall ranking. However, if the values of θ_k are sufficiently different from each other such that $(L_k, U_k) \cap (L_{k'}, U_{k'}) = \emptyset$ for all $k \neq k'$ and $k = 1, 2, \ldots, K$, then it follows immediately that the joint confidence region contains only one overall ranking, and it is the estimated ranking $(\hat{r}_1, \ldots, \hat{r}_K)$; when this happens, we would have the 'tightest' possible joint confidence region.

4. Joint confidence region construction: two examples

For simplicity, this section illustrates two ways to construct joint confidence regions from the set of familiar $\hat{\theta}_k \pm z_{\alpha/2}\mathrm{SE}_k$ individual confidence intervals, assuming that each SE_k is known, although in most applications this is only approximately true. However, for full generality, it is straightforward to adapt either construction to work with any appropriate method for constructing individual $100(1-\alpha)\%$ confidence intervals for each θ_k. Nonetheless, for ease of exposition, let us assume that $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K$ are independently distributed such that $\hat{\theta}_k \sim N(\theta_k, \mathrm{SE}_k)$ for $k = 1, 2, \ldots, K$ with $\theta_1, \theta_2, \ldots, \theta_K$ unknown and $\mathrm{SE}_1, \mathrm{SE}_2, \ldots, \mathrm{SE}_K$ known. For a given θ_k, an individual $100(1-\alpha)\%$ confidence interval is

$$(\hat{\theta}_k - z_{\alpha/2}\mathrm{SE}_k, \hat{\theta}_k + z_{\alpha/2}\mathrm{SE}_k).$$

To construct a joint confidence region as discussed in Section 3, we consider two cases of the joint confidence intervals for $\theta_1, \ldots, \theta_K$:

(a) using Bonferroni correction and
(b) using independence.

(Note that $\mathrm{MOE}_k = z_{\alpha/2}\mathrm{SE}_k$ for $k = 1, \ldots, K$ are assumed known.)

4.1. Joint confidence intervals for $\theta_1, \ldots, \theta_K$ by using Bonferroni correction

We apply the Bonferroni correction to obtain a collection of confidence intervals whose joint coverage for $\theta_1, \theta_2, \ldots, \theta_K$ is greater than or equal to $1-\alpha$; these intervals are given by

$$(\hat{\theta}_k - z_{(\alpha/K)/2}\mathrm{SE}_k, \hat{\theta}_k + z_{(\alpha/K)/2}\mathrm{SE}_k), \qquad \text{for } k = 1, 2, \ldots, K. \tag{10}$$

The Bonferroni inequality (Mukhopadhyay (2000), page 157) states that, for events $A_1, \ldots, A_K$,

$$P\left(\bigcap_{i=1}^{K} A_i\right) \geqslant \sum_{i=1}^{K} P(A_i) - (K-1).$$

Applying the Bonferroni inequality, we see that

$$P\left[\bigcap_{k=1}^{K}\{\theta_k \in (\hat{\theta}_k - z_{(\alpha/K)/2}\mathrm{SE}_k, \hat{\theta}_k + z_{(\alpha/K)/2}\mathrm{SE}_k)\}\right]$$
$$\geqslant \sum_{k=1}^{K} P\{\theta_k \in (\hat{\theta}_k - z_{(\alpha/K)/2}\mathrm{SE}_k, \hat{\theta}_k + z_{(\alpha/K)/2}\mathrm{SE}_k)\} - (K-1)$$
$$= \sum_{k=1}^{K}(1-\alpha/K) - (K-1) = K - \alpha - (K-1) = 1-\alpha.$$

Thus the Bonferroni-corrected confidence intervals given by expression (10) have joint coverage probability greater than or equal to $1-\alpha$. We apply the proposed methodology to the ACS travel time to work data for the year 2011. In this example, θ_k is the mean travel time (in minutes) to work for state k (including Washington DC) where $k = 1, 2, \ldots, 51$. The fifth column of Table 2 shows the Bonferroni-corrected joint confidence intervals for $\theta_1, \theta_2, \ldots, \theta_{51}$ as given by expression (10) with $\alpha = 0.10$. Table 2 also shows the joint confidence region (sixth column) for the ranking $(r_1, \ldots, r_{51})$ obtained by using expression (9) as applied to the Bonferroni-corrected confidence intervals for $\theta_1, \theta_2, \ldots, \theta_K$. (In some cases, it is convenient to show that k ranges over the names of the states rather than over the integers $1, \ldots, K$.)

To illustrate some of the details on part of one row in Table 2, we focus on Alabama. For $\alpha = 0.10$, $z_{(\alpha/51)/2} = 3.1$. The Bonferroni-corrected joint confidence interval for θ_{Alabama} is given by expression (10):

$$\left(23.9 - \frac{3.1 \times 0.2}{1.645}, 23.9 + \frac{3.1 \times 0.2}{1.645}\right) = (23.5, 24.3). \tag{11}$$

To obtain the portion of the joint confidence region for r_{Alabama}, we refer to the observed ranking and note that

$$\Lambda_{\text{L,Alabama}} = \{\text{Missouri, Ohio, Minnesota, } \ldots, \text{South Dakota}\} \text{ implies } |\Lambda_{\text{L, Alabama}}| = 20,$$
$$\Lambda_{\text{R,Alabama}} = \{\text{Maryland, New York, } \ldots, \text{ Washington DC, Connecticut, Arizona, Texas}\}$$
$$\text{implies } |\Lambda_{\text{R, Alabama}}| = 18$$

and

$$\Lambda_{\text{O,Alabama}} = \{\text{Delaware, Colorado, } \ldots, \text{Nevada, Mississippi, Rhode Island}\} \text{ implies}$$
$$|\Lambda_{\text{O,Alabama}}| = 12.$$

Hence the portion of the joint confidence region for r_{Alabama} by using expression (9) is

$$\{20+1, 20+2, 20+3, \ldots, 20+12, 20+12+1\} = \{21, \ldots, 33\}. \tag{12}$$

The Bonferroni portion for other rows of Table 2 are obtained similarly.

4.2. Joint confidence intervals for $\theta_1, \ldots, \theta_K$ by using independence

In the independence situation, because $\hat{\theta}_1, \ldots, \hat{\theta}_K$ are independently distributed such that $\hat{\theta}_k \sim N(\theta_k, \mathrm{SE}_k)$ for $k = 1, 2, \ldots, K$ with $\theta_1, \theta_2, \ldots, \theta_K$ unknown and $\mathrm{SE}_1, \mathrm{SE}_2, \ldots, \mathrm{SE}_K$ known, we may also consider the following intervals whose joint coverage equals $1-\alpha$:

$$(\hat{\theta}_k - z_{\gamma/2}\mathrm{SE}_k, \hat{\theta}_k + z_{\gamma/2}\mathrm{SE}_k), \qquad \text{for } k = 1, 2, \ldots, K, \tag{13}$$

where $\gamma = 1 - (1-\alpha)^{1/K}$. We note that

Table 2 Joint confidence region for ranking based on joint confidence intervals (θ_ks): Bonferroni or independence (travel time to work data)†

$\hat{r}_k$	*State (k)*	$\hat{\theta}_k$	MOE_k	*Results for Bonferroni (10)*		*Results for independence (13)*	
				90% joint confidence intervals for θ_ks	*90% joint confidence region for ranking*	*90% joint confidence intervals for θ_ks*	*90% joint confidence region for ranking*
51	Maryland (MD)	32.2	0.2	(31.8, 32.6)	{50, 51}	(31.8, 32.6)	{50, 51}
50	New York (NY)	31.5	0.2	(31.1, 31.9)	{50, 51}	(31.1, 31.9)	{50, 51}
49	New Jersey (NJ)	30.5	0.2	(30.1, 30.9)	{48, 49}	(30.1, 30.9)	{48, 49}
48	District of Columbia (DC)	30.1	0.5	(29.2, 31.0)	{48, 49}	(29.2, 31.0)	{48, 49}
47	Illinois (IL)	28.2	0.2	(27.8, 28.6)	{45, 46, 47}	(27.8, 28.6)	{45, 46, 47}
46	Massachusetts (MA)	28.0	0.2	(27.6, 28.4)	{43, ..., 47}	(27.6, 28.4)	{43, ..., 47}
45	Virginia (VA)	27.7	0.2	(27.3, 28.1)	{43, ..., 47}	(27.3, 28.1)	{43, ..., 47}
44	California (CA)	27.1	0.1	(26.9, 27.3)	{42, 43, 44}	(26.9, 27.3)	{42, 43, 44}
44	Georgia (GA)	27.1	0.3	(26.5, 27.7)	{42, ..., 46}	(26.5, 27.7)	{42, ..., 46}
42	New Hampshire (NH)	26.9	0.5	(26.0, 27.8)	{37, ..., 46}	(26.0, 27.8)	{37, ..., 46}
41	Pennsylvania (PA)	25.9	0.1	(25.7, 26.1)	{36, ..., 42}	(25.7, 26.1)	{36, ..., 42}
40	Florida (FL)	25.8	0.2	(25.4, 26.2)	{35, ..., 42}	(25.4, 26.2)	{35, ..., 42}
39	Hawaii (HI)	25.7	0.4	(24.9, 26.5)	{32, ..., 42}	(25.0, 26.4)	{33, ..., 42}
38	West Virginia (WV)	25.6	0.5	(24.7, 26.5)	{30, ..., 42}	(24.7, 26.5)	{30, ..., 42}
37	Washington (WA)	25.5	0.2	(25.1, 25.9)	{34, ..., 41}	(25.1, 25.9)	{34, ..., 41}
36	Delaware (DE)	25.3	0.6	(24.2, 26.4)	{25, ..., 42}	(24.2, 26.4)	{25, ..., 42}
35	Connecticut (CT)	25.0	0.3	(24.4, 25.6)	{27, ..., 40}	(24.4, 25.6)	{27, ..., 40}
34	Arizona (AZ)	24.8	0.2	(24.4, 25.2)	{27, ..., 39}	(24.4, 25.2)	{27, ..., 39}
34	Texas (TX)	24.8	0.1	(24.6, 25.0)	{29, ..., 38}	(24.6, 25.0)	{30, ..., 37}
32	Colorado (CO)	24.5	0.3	(23.9, 25.1)	{23, ..., 38}	(23.9, 25.1)	{23, ..., 38}
32	Louisiana (LA)	24.5	0.2	(24.1, 24.9)	{23, ..., 37}	(24.1, 24.9)	{24, ..., 37}
30	Tennessee (TN)	24.2	0.2	(23.8, 24.6)	{22, ..., 35}	(23.8, 24.6)	{22, ..., 35}
29	Michigan (MI)	24.1	0.2	(23.7, 24.5)	{21, ..., 35}	(23.7, 24.5)	{21, ..., 35}
29	Nevada (NV)	24.1	0.4	(23.3, 24.9)	{19, ..., 37}	(23.4, 24.8)	{20, ..., 37}
27	Alabama (AL)	23.9	0.2	(23.5, 24.3)	{21, ..., 33}	(23.5, 24.3)	{21, ..., 33}
27	Mississippi (MS)	23.9	0.4	(23.1, 24.7)	{17, ..., 36}	(23.2, 24.6)	{17, ..., 35}
25	South Carolina (SC)	23.6	0.3	(23.0, 24.2)	{16, ..., 32}	(23.0, 24.2)	{16, ..., 32}
24	Indiana (IN)	23.5	0.2	(23.1, 23.9)	{17, ..., 30}	(23.1, 23.9)	{17, ..., 30}
23	Maine (ME)	23.4	0.4	(22.6, 24.2)	{15, ..., 32}	(22.7, 24.1)	{15, ..., 31}
23	North Carolina (NC)	23.4	0.2	(23.0, 23.8)	{16, ..., 29}	(23.0, 23.8)	{16, ..., 29}
23	Rhode Island (RI)	23.4	0.5	(22.5, 24.3)	{15, ..., 33}	(22.5, 24.3)	{15, ..., 33}
20	Missouri (MO)	23.1	0.2	(22.7, 23.5)	{15, ..., 27}	(22.7, 23.5)	{15, ..., 27}
20	Ohio (OH)	23.1	0.1	(22.9, 23.3)	{16, ..., 26}	(22.9, 23.3)	{16, ..., 26}
18	Minnesota (MN)	23.0	0.2	(22.6, 23.4)	{15, ..., 27}	(22.6, 23.4)	{15, ..., 26}
17	Kentucky (KY)	22.9	0.2	(22.5, 23.3)	{15, ..., 26}	(22.5, 23.3)	{15, ..., 26}
16	Oregon (OR)	22.5	0.3	(21.9, 23.1)	{11, ..., 24}	(21.9, 23.1)	{11, ..., 24}
15	Vermont (VI)	21.9	0.5	(21.0, 22.8)	{10, ..., 21}	(21.0, 22.8)	{10, ..., 21}
15	Wisconsin (WI)	21.9	0.2	(21.5, 22.3)	{11, ..., 16}	(21.5, 22.3)	{11, ..., 16}
13	Utah (UT)	21.6	0.3	(21.0, 22.2)	{10, ..., 16}	(21.0, 22.2)	{10, ..., 16}
12	New Mexico (NM)	21.4	0.4	(20.6, 22.2)	{10, ..., 16}	(20.7, 22.1)	{10, ..., 16}
11	Arkansas (AR)	21.3	0.4	(20.5, 22.1)	{10, ..., 16}	(20.6, 22.0)	{10, ..., 16}
10	Oklahoma (OK)	21.1	0.2	(20.7, 21.5)	{10, ..., 14}	(20.7, 21.5)	{10, ..., 14}
9	Idaho (ID)	19.7	0.4	(18.9, 20.5)	{4, ..., 9}	(19.0, 20.4)	{4, ..., 9}
8	Kansas (KS)	18.9	0.3	(18.3, 19.5)	{3, ..., 9}	(18.3, 19.5)	{3, ..., 9}
7	Iowa (IA)	18.8	0.2	(18.4, 19.2)	{3, ..., 9}	(18.4, 19.2)	{3, ..., 9}
6	Alaska (AK)	18.4	0.5	(17.5, 19.3)	{1, ..., 9}	(17.5, 19.3)	{1, ..., 9}
5	Montana (MT)	18.2	0.5	(17.3, 19.1)	{1, ..., 9}	(17.3, 19.1)	{1, ..., 9}
4	Nebraska (NE)	18.1	0.3	(17.5, 18.7)	{1, ..., 8}	(17.5, 18.7)	{1, ..., 8}
4	Wyoming (WY)	18.1	0.8	(16.6, 19.6)	{1, ..., 9}	(16.6, 19.6)	{1, ..., 9}
2	North Dakota (ND)	16.9	0.6	(15.8, 18.0)	{1, ..., 6}	(15.8, 18.0)	{1, ..., 6}
2	South Dakota (SD)	16.9	0.5	(16.0, 17.8)	{1, ..., 6}	(16.0, 17.8)	{1, ..., 6}

†Source: based on 2011 1-year ACS, ranking table R0801.

$$P\left[\bigcap_{k=1}^{K}\{\theta_k \in (\hat{\theta}_k - z_{\gamma/2}\mathrm{SE}_k, \hat{\theta}_k + z_{\gamma/2}\mathrm{SE}_k)\}\right]$$

$$= P\left(-z_{\gamma/2} < \frac{\hat{\theta}_1 - \theta_1}{\mathrm{SE}_1} < z_{\gamma/2}, -z_{\gamma/2} < \frac{\hat{\theta}_2 - \theta_2}{\mathrm{SE}_2} < z_{\gamma/2}, \ldots, -z_{\gamma/2} < \frac{\hat{\theta}_K - \theta_K}{\mathrm{SE}_K} < z_{\gamma/2}\right)$$

$$= \prod_{k=1}^{K} P\left(-z_{\gamma/2} < \frac{\hat{\theta}_k - \theta_k}{\mathrm{SE}_k} < z_{\gamma/2}\right)$$

$$= \prod_{k=1}^{K} (1-\gamma) = (1-\gamma)^K = [1 - \{1 - (1-\alpha)^{1/K}\}]^K = 1 - \alpha.$$

Thus the confidence intervals given by expression (13) have joint coverage probability equal to $1-\alpha$. As with the Bonferroni-corrected confidence intervals, we apply this proposed methodology to the ACS travel time to work data. The seventh column of Table 2 shows the joint confidence intervals for $\theta_1, \theta_2, \ldots, \theta_{51}$ as given by expression (13) with $\alpha = 0.10$. The last column of Table 2 shows the joint confidence region for the ranking $(r_1, r_2, \ldots, r_{51})$ that is obtained by using expression (9) as applied to the independent confidence intervals for $\theta_1, \theta_2, \ldots, \theta_{51}$.

Fig. 1 (independence) makes it easy to identify all overall rankings in the 90% joint confidence region as specified in expression (9). We are 90% confident that the true overall ranking of the 51 states lies within this joint confidence region. Ties are permitted. For example, states Vermont and Wisconsin are tied at estimated rank 15.

The joint confidence region in Fig. 1 contains many possible rankings of the 51 states based on the data. By result 1, one of those overall rankings is the observed estimated overall ranking of states in the first and second columns of Table 2 and shown in the bold rectangles in Fig. 1. One other example of an overall ranking in the joint confidence region is (ordered from rank 51 to rank 1; see Table 2 for definitions of the state abbreviations) (NY, MD, DC, NJ, IL, MA, VA, CA, GA, NH, PA, FL, HI, WV, WA, DE, CT, TX, AZ, LA, CO, TN, MI, NV, AL, MS, SC, IN, ME, NC, RI, OH, MO, MN, KY, OR, VT, WI, UT, NM, AR, OK, ID, KS, IA, AK, MT, NE, WY, SD, ND). To select other rankings from the joint confidence region, start with the first row at the top and select NY or MD for rank 51. From the second row from the top, select NY or MD for rank 50—the one state that is not selected for rank 51. (If NY and MD are both selected for rank 51 (tied), then we leave row 50 blank and proceed to row 49. We proceed similarly for each remaining row.) For rank 49, select NJ or DC a state from the third row of the region. Continue in similar fashion until selecting a state for rank 1. A state can be selected only once in any ranking, but a rank can be assigned to more than one state.

Moreover, the joint confidence region of Fig. 1 makes it easy to read off marginal confidence intervals. Each row of the joint confidence region (Fig. 1) shows which states could occupy each rank. Similarly, each column k of the joint confidence region (Fig. 1) shows the marginal confidence set for the rank r_k of state k, i.e. the possible values of r_k.

From Table 2, note that the confidence interval for θ_{Nevada} under independence is shorter than the corresponding confidence interval for θ_{Nevada} under Bonferroni correction. The same is true for some other states, e.g. Mississippi and Maine. In most cases, the given corresponding intervals under independence and under Bonferroni correction are equal in length. We observe in result 2 that the confidence interval under independence will always be no longer than the corresponding confidence interval under Bonferroni correction. As a consequence, the joint confidence region for $(r_1, \ldots, r_{51})$ based on independence is at least as tight as the corresponding joint confidence region based on the Bonferroni correction.

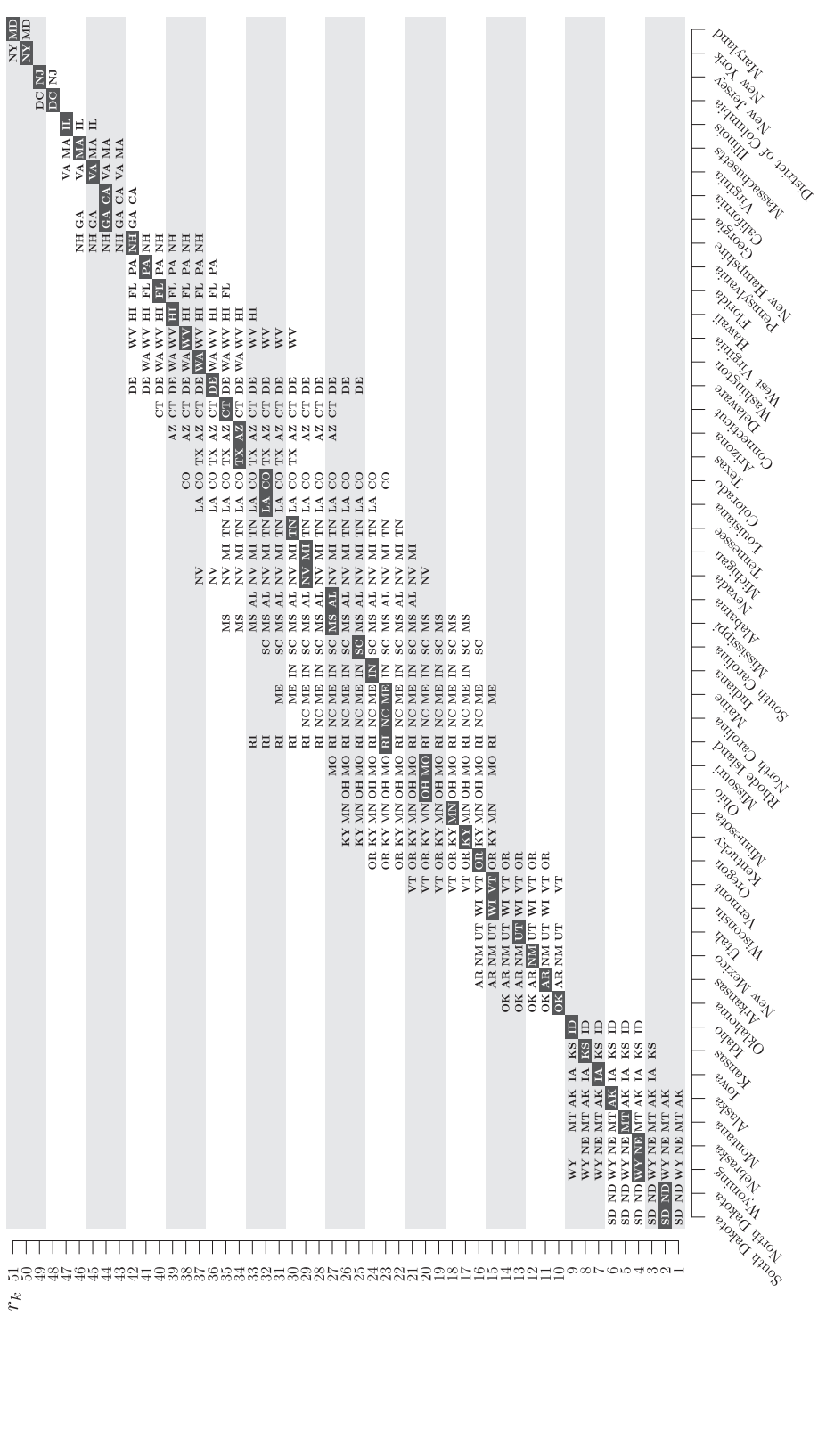

Fig. 1. (Independence) visualization of the 90% joint confidence region for the ranking by using travel time to work data: highlighted states in the joint confidence region show the observed estimated ranking of Table 2; each specific row of the joint confidence region shows which states could occupy the associated rank; each specific column of the joint confidence region shows which ranks the associated state could occupy; for example, rank (row) 10 could be occupied by states OK, AR, NM, UT or VT; state (column) MN could occupy ranks 15, 16, 17, 18, 19, . . . , 25 or 26 (source, based on data from 2011 1-year ACS, ranking table R0801; US Bureau of the Census, Washington DC)

Result 2. The intervals in expression (13) based on independence are shorter than the corresponding intervals in expression (10) based on Bonferroni correction.

Proof. Note that the intervals in expression (13) are shorter than the corresponding intervals in expression (10) if and only if $z_{\gamma/2} < z_{(\alpha/K)/2}$, which is equivalent to

$$1-\alpha < \left(1-\frac{\alpha}{K}\right)^K. \tag{14}$$

Thus it is sufficient to show that the inequality in expression (14) is true. A simple proof based on the binomial theorem is found in Klein *et al.* (2018).

Although the intervals that are based on independence are shorter than the corresponding intervals based on Bonferroni correction, the 90% joint confidence regions are similar in Table 2 (Bonferroni and independence) because the values of $(\alpha/K)/2 = 0.00098$ and $\gamma/2 = 0.00103$ are nearly equal, with corresponding z-values 3.096 and 3.081 respectively. Thus the corresponding confidence intervals in expressions (10) and (13) are close as shown in Table 2.

5. Discussion

Remark 1. A simple and useful $100(1-\alpha)\%$ joint confidence region is given for an overall ranking $(r_1, r_2, \ldots, r_K)$ of K populations that gives a measure of uncertainty for the estimated overall ranking $(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_K)$ based on sample survey data. When none of the confidence intervals for the θ_k overlap, the joint confidence region is as 'tight' as it can be and contains only the observed estimated overall ranking $(\hat{r}_1, \ldots, \hat{r}_K)$. National statistical agencies may increase the release of estimated overall rankings now that a measure of uncertainty for each exists that can be shared with users. A possible wording of the release is 'Our estimated overall ranking is abc ... xyz, and we are 90% confident that the true overall ranking lies within the joint confidence region shown in Fig. 1 (or equivalently Table 2)'.

Remark 2. A proposed visualization makes it easy to communicate this uncertainty in the estimated overall ranking while also revealing many other possible overall rankings (for example see Fig. 1).

Remark 3. The key concept underlying the joint confidence region presented is found in the following quote from a reviewer:

> 'Assume that, for each unit (i.e., state), we have identified a confidence interval (CI) that places an upper and lower bound on what its true mean can be. Then the rank of any unit can reliably be placed above that of all of the units whose CIs lie entirely below its own CI, and conversely (below all of the units whose CIs lie entirely above its own CI).'

Of course there is the possibility of other units' confidence intervals that overlap this unit's confidence interval. The full details form this paper's content. A simulation study which confirms the method and its properties for both Bonferroni and independence cases is given in Klein *et al.* (2018).

Remark 4. There are several differences between the current treatment of rankings (Table 1) and the proposed treatment of overall rankings (Fig. 1 or equivalently Table 2) of this paper. First, and perhaps most significantly, Table 1 does not provide a measure of uncertainty for the estimated overall ranking; Fig. 1 does, and it is the joint confidence region. As is the case with a confidence interval for a scalar parameter, the joint confidence region shows the many other

possible true overall rankings. Fig. 1 focuses on rankings and ranks, which we desire; Table 1 seems to place most of its attention directly on the parameters θ_k. Table 1 which compares Alabama separately with each of the other 50 states is just one of 51 displays that are needed to show how each state ranks against the others; Fig. 1 presents only one needed display which shows (at once) how all states stand relative to each and all of the others. Beyond the estimated rank for a given state, Table 1 does not show the other possible ranks for that state; Fig. 1 does show the other possible ranks for that state. For a given rank, Table 1 does not show which states can hold that rank; Fig. 1 does show the other states that can hold a given rank. Table 1 is a table which requires more effort to absorb its contents; Fig. 1 is a visualization which helps to simplify and to deliver its content correctly. Moreover, Fig. 1 shows which estimates of individual ranks are more reliable. Those estimates of states with shorter vertical columns are more reliable than those estimates of states with longer vertical columns.

To be specific, Table 1 shows that the estimated rank of AL is 27 and that AL is tied with MS. We also see that AL's estimated mean travel time to work is not statistically different from those for TN, MI, NV, MS, SC and RI. From Fig. 1, the estimated rank for AL is 27 and it is tied with MS. Fig. 1 shows that the rank of AL could range from 21 to 33. We also see the other states that could occupy these ranks. In particular, rank 27 could be occupied by states MO, RI, NC, ME, IN, SC, MS, AL, NV, MI, TN, LA, CO, AZ, CT and DE. From Fig. 1, we immediately see that the estimated rank for AL is more reliable than the estimated rank for MS.

Remark 5. As we briefly mentioned earlier, Bayesian methods have been used to address ranking problems. For example, Laird and Louis (1989) provided

> '... ranking methods based on the conditional (posterior) distribution of the ranks rather than the conditional (posterior) distribution of the θ_ks'.

They proposed a ranking based on the posterior mean ranks which has the effect of shrinking each of the estimated ranks towards the mean rank $(K+1)/2$ when there is high variability in the posterior distributions of the θ_ks. To provide uncertainty in the individual posterior mean ranks, Laird and Louis used the posterior standard deviations.

Also in a Bayesian setting, Shen and Louis (1998) sought to provide a collection of estimates of parameters (θ_ks) to satisfy three goals:

(a) good estimates of the parameters;
(b) a good estimate of the histogram of the parameters;
(c) good estimates of the individual ranks of the parameters.

In the context of the bootstrap, Hall and Miller (2010) observed that

> '... one feature of many rankings reported over time is that the ordering (of K objects) at the extreme top or bottom remains relatively invariant (over time) ... (their explanation). Those scores at the extreme of a range are more likely to be sufficiently "spaced out" to overcome the problems of data noise, whereas less extreme scores are likely to be bunched more closely together.'

However, each of these three approaches studies the estimates of individual ranks. None of them provides an explicit statement of uncertainty for the estimated overall ranking, as is the main contribution of our paper.

Remark 6. Fig. 1 shows varying levels of uncertainty among the estimated ranks. Specifically, there is more uncertainty around the mid-estimated ranks than at the extremes. Also the uncertainty is particularly small for the four states holding the estimated ranks 51, 50, 49 and 48. As can be seen immediately from the main result (7), the size of the set which gives the

possible ranks for state k is $|\Lambda_{Ok}| + 1$, i.e. 1 plus the number of states whose confidence interval for $\theta_{k'}$ overlaps the confidence interval for θ_k where $k' \neq k$ (Fig. 2). We comment on three groups of states, though it seems difficult to make generalizations about strong relationships between travel times to work, varying geography and levels of uncertainty among the estimated ranks in Fig. 1. Other groupings are possible.

(a) *Group 1—states with estimated ranks 51,..., 39*: these states are mainly along the east coast of the USA (MD, NY, NJ, DC, MA, VA, GA, NH, PA and FL), part of the west coast and pacific (CA and HI) and part of the central USA (IL). The states tend to be highly urban with large populations and large population densities. They tend to report the longest travel times to work. The relatively large individual estimates $\hat{\theta}_k$ among these states are more 'spaced out' and as a result the confidence intervals for these (θ_k) do not overlap much (Fig. 2). The average value of $|\Lambda_{Ok}|$ for states in this group is 3.9 states. These states tend to show the least uncertainties in estimated ranks.

(b) *Group 2—states with estimated ranks 15,..., 2*: these states are mainly among the mountain states (UT, NM, ID, MT and WY), parts of central USA (WI, AR, OK, KS, IA, NE, ND and SD) and AK. The states tend to have large unpopulated land areas with relatively few large population areas and large population density areas. They tend to report the shortest travel times to work. Although the individual estimates $\hat{\theta}_k$ among these states are 'somewhat spaced out', they are not as spaced out as those in group 1, and there is a little more overlap among their confidence intervals than in group 1 (Fig. 2). The average value of $|\Lambda_{Ok}|$ for states in this group is 6.1 states. These states tend to show the second least uncertainties in estimated ranks.

(c) *Group 3—states with estimated ranks 38,..., 15*: the remaining states are all over the USA, but mainly among the east central (north and south) states (MI, IN, OH, KY, TN, MS and AL), some other southern states (NC, SC, LA and TX) and some western states (WA, OR, NV, AZ and CO). The states tend to have mixes of urban and rural areas, a mix of large and small population areas and large and small population density areas. The individual estimates $\hat{\theta}_k$ among these states are more 'bunched' together and their confidence intervals overlap more than in groups 1 or 2 (Fig. 2). They tend to report the middle travel times to work, and each state tends to have confidence intervals for θ_k that overlap more states than is true for states in the other two groups. The average value of $|\Lambda_{Ok}|$ for states in this group is 13.1 states. These states tend to show the largest uncertainties in estimated ranks.

Remark 7. In the presence of sampling error, inferences about a population's rank are not necessarily equivalent to pairwise comparisons between populations. Therefore, attempts to make pairwise comparisons by using the joint confidence region proposed may lead to misleading inferences, as the joint confidence region is not designed for this.

For instance, imagine a set of three populations A, B and C with estimated ranks 1, 2 and 3 respectively, i.e. $\hat{r}_A = 1$, $\hat{r}_B = 2$ and $\hat{r}_C = 3$. If $\mathrm{SE}(\hat{\theta}_A)$ is sufficiently large, whereas B and C have sufficiently small SEs, it is possible to be highly confident that $r_C > r_B$, and so $r_C > 1$, but not to be confident that r_C is different from r_A. In other words, confidence that $r_C \neq 1$ does not imply confidence that $r_C \neq r_A$, even though $\hat{r}_A = 1$.

As a concrete example with reference to Table 2, the confidence region for Alabama's rank is $\{21, \ldots, 33\}$, and Delaware's estimated rank is 36. Although we are 90% confident that $r_{AL} \neq 36$, this is not necessarily equivalent to being 90% confident that Alabama and Delaware have different ranks or different mean travel times. The underlying confidence intervals (23.5, 24.3) and (24.2, 26.4) for mean travel time (Table 2, under independence) respectively for Alabama

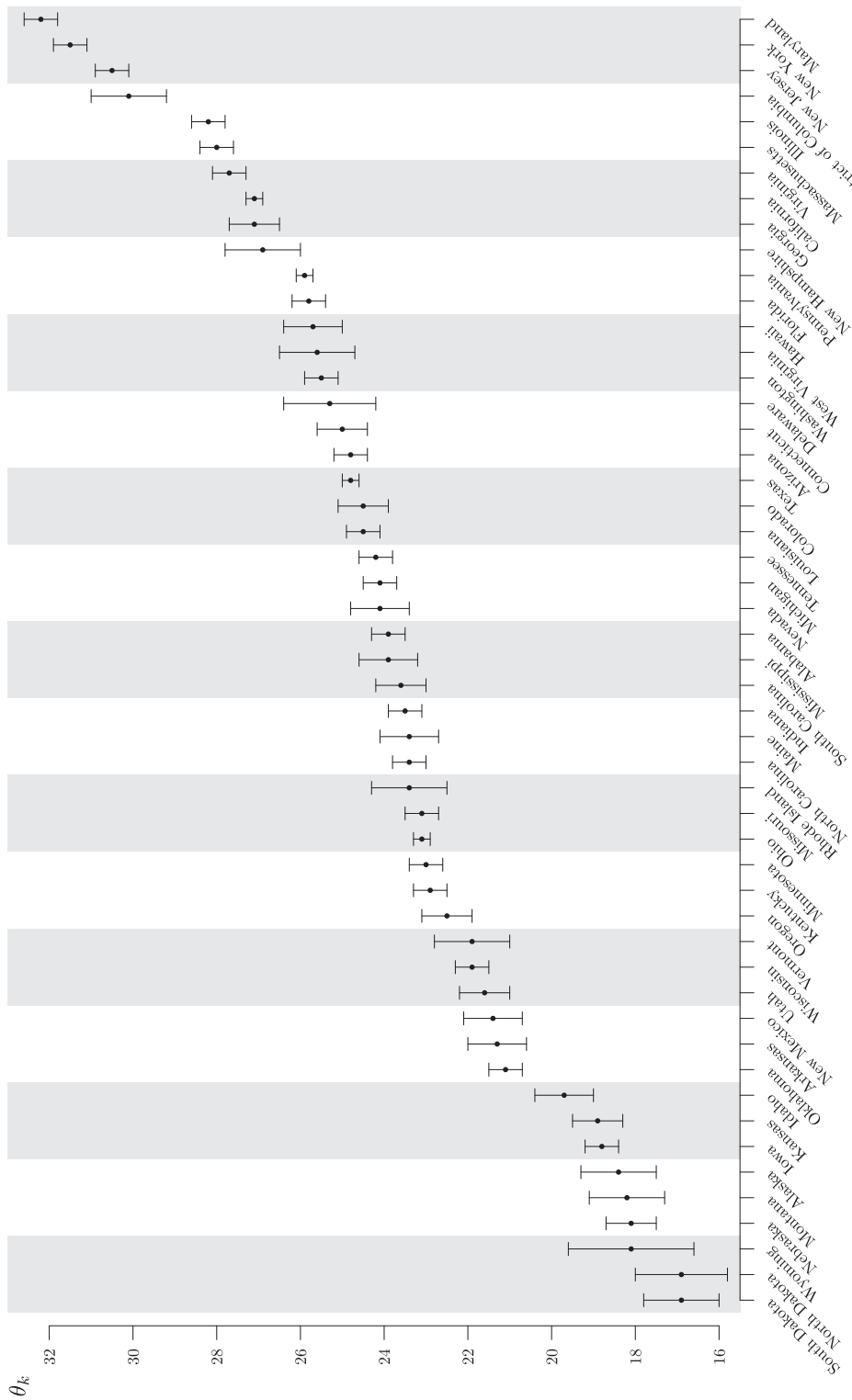

Fig. 2. (Independence) visualization of the 90% joint confidence intervals for θ_ks from Table 2 and their overlap for the ranking by using travel time to work data: the intervals provide the foundation for the joint confidence region in Fig. 1 (source: based on data from 2011 1-year ACS, ranking table R0801; US Bureau of the Census, Washington DC)

and for Delaware overlap, as do their underlying confidence intervals for rank. In either case, we cannot safely determine whether the mean travel times or the ranks of Alabama and Delaware are necessarily significantly different, because it is known that the method of examining overlap of pairs of confidence intervals should not be used for formal hypothesis testing (Schenker and Gentleman, 2001; Wright *et al.*, 2019). If the primary goal is to make formal pairwise comparisons, then we suggest using a method that is specifically designed for this such as discussed by Wright *et al.* (2019).

Remark 8. Finally, the methodology in this paper has been developed from the frequentist perspective; and from this perspective we have developed a joint confidence region for the overall ranking. If one instead works from the Bayesian perspective, then this paper's main result can also be used to construct a credible region that contains the overall ranking $(r_1, \ldots, r_K)$ with posterior probability at least $1 - \alpha$. Such a region can be constructed by first using the posterior distribution of $(\theta_1, \ldots, \theta_K)$ to construct a set of joint credible intervals $(L_1, U_1), \ldots, (L_K, U_K)$ such that the posterior probability of the event $\cap_{k=1}^{K} \{\theta_k \in (L_k, U_k)\}$ is at least $1 - \alpha$. By applying this paper's main result, one can then obtain a credible region for the overall ranking $(r_1, \ldots, r_K)$, such that the posterior probability that the overall ranking is contained in this region is at least $1 - \alpha$.

6. R software for figures

All figures in this paper were made in R (R Core Team, 2019). Our data set, plotting functions and example code are in the `RankingProject` R package (Wieczorek, 2020), which is available on line from the CRAN repository, `https://CRAN.R-project.org/package=RankingProject`. The package also contains a vignette which reproduces both figures in the paper.

Acknowledgements

The authors are grateful for insightful comments from the referees that have helped to clarify and improve the paper's content. The views expressed are those of the authors and not those of the US Census Bureau or the US Food and Drug Administration.

References

Bechhofer, R. E. (1954) A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Statist.*, **25**, 16–29.

Brand, R., van Houwelingen, H., le Cessie, S. and Louis, T. A. (1996) Empirical Bayes analysis for the Dutch 'Mirroring' Project'. Unpublished. Department of Medical Statistics, Leiden University Medical Center, Leiden.

Goldstein, H. and Spiegelhalter, D. J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *J. R. Statist. Soc.* A, **159**, 385–443.

Govindarajulu, Z. and Harvey, C. (1974) Bayesian procedures for ranking and selection problems. *Ann. Inst. Statist. Math.*, **26**, 35–53.

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

Hall, P. and Miller, H. (2010) Modeling the variability of rankings. *Ann. Statist.*, **30**, 2652–2677.

Klein, M. and Wright, T. (2011) Ranking procedures for several normal populations: an empirical investigation. *Int. J. Statist. Sci.*, **11**, 37–58.

Klein, M., Wright, T. and Wieczorek, J. (2018) A simple joint confidence region for a ranking of K populations: application to American Community Survey's travel time to work data. *Research Report Statistics #2018-04*. Center for Statistical Research and Methodology, US Bureau of the Census, Washington DC.

Laird, N. M. and Louis, T. A. (1989) Empirical Bayes ranking methods. *J. Educ. Statist.*, **14**, 29–46.

Louis, T. A. (1984) Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Am. Statist. Ass.*, **79**, 393–398.

Morris, C. N. and Christiansen, C. L. (1996) Hierarchical models for ranking and for identifying extremes, with applications (with discussion). In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 277–296. Oxford: Oxford University Press.

Mukhopadhyay, N. (2000) *Probability and Statistical Inferences*. New York: Dekker.

R Core Team (2019) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Schenker, N. and Gentleman, J. E. (2001) On judging the significance of differences by examining the overlap between confidence intervals. *Am. Statistn*, **55**, 182–186.

Shen, W. and Louis, T. A. (1998) Triple-goal estimates in two-stage hierarchical models. *J. R. Statist. Soc.* B, **60**, 455–471.

Siddiqui, F. (2018) Washington-area residents spent an average of nearly 35 minutes commuting each way in 2017, the Census Bureau Reports. *Washington Post*, Sept. 17th.

Wieczorek, J. (2020) RankingProject: the Ranking Project: visualizations for comparing populations. *R Package Version 0.2.0*. (Available from `https://CRAN.R-project.org/package=RankingProject`.)

Wright, T., Klein, M. and Wieczorek, J. (2013) An overview of some concepts for potential use in ranking populations based on sample survey data. In *Proc. 59th Wrld Statistics Congr., Hong Kong*. The Hague: International Statistical Institute.

Wright, T., Klein, M. and Wieczorek, J. (2014) Ranking populations based on sample survey data. *Research Report Statistics # 2014-07*. Center for Statistical Research and Methodology, US Bureau of the Census, Washington DC.

Wright, T., Klein, M. and Wieczorek, J. (2019) A primer on visualizations for comparing populations, including the issue of overlapping confidence intervals. *Am. Statistn*, **73**, 165–178.