An Overview of Some Concepts for Potential Use in Ranking Populations Based on Sample Survey Data

Tommy Wright, Martin Klein, and Jerzy Wieczorek
U. S. Bureau of the Census; Washington, D. C., USA
Corresponding author: Tommy Wright, e-mail: tommy.wright@census.gov

Abstract

Our main objective is to call national statistical agencies' attention to the need to express uncertainty in rankings based on data from sample surveys. We share some simple and easy to use methods for presenting uncertainty in estimated rankings. We see promise with the bootstrap.

Keywords: Bootstrap, Nonparametrics, Official statistics, Uncertainty in rankings.

1. INTRODUCTION

We rank k populations with random variables Y_i from smallest to largest based on the values of some associated real-valued parameters θ_i, for $i = 1, ..., k$. When the value of θ_i is unknown, we use observed values $Y_{i1}, ..., Y_{in_i}$ in a random sample from population i to compute $\hat{\theta}_i$ as an estimate of θ_i, for all i. So assume k independent sample survey estimates $\hat{\theta}_1, ..., \hat{\theta}_k$ and associated standard errors $SE_1, ..., SE_k$ are used to produce estimated rankings. For example, the U. S. Census Bureau's American Community Survey (ACS) produced 85 different rankings of the $k = 51$ states (50 states and Washington, D.C.) based on observed sample estimates during 2011. One of those rankings ranks the states based on $\hat{\theta}_i$, the estimated mean travel time to work for workers 16 years and over who did not work at home (minutes) for state i.

Beginning with pair-wise comparisons, we consider and modify some known practices, assisted by visualizations. In Section 2, we illustrate four methods comparing a pair of populations using normal theory, and in Section 3, we define three uncertainty measures and their estimates for the estimated ranks using the bootstrap (nonparametric/parametric). For all seven methods illustrated, we only need the k sample estimates $\hat{\theta}_i$ and their associated standard errors SE_i. We end by introducing the concept of "a most probable ranking" inspired by classical nonparametric methods.

A nation's official statistics should be *widely understood* and *robust*, among many other properties. These two desired characteristics are shared by classical probability design-based and model-assisted sampling methods (e.g., Cochran, 1977; Fuller, 2009; Lohr, 2010; Särndal, Swensson, and Wretman, 2003) that are commonly used by national statistical agencies around the world, and they are also inherent in classical nonparametric methods using ranks (e.g., see Hollander and Wolfe, 1999).

2. VISUALLY COMPARING PAIRS OF STATES USING NORMAL THEORY

2.1. Comparing One Reference State Against Each of the Other States - Version I

For population i, we treat the SE_i estimates as though they were known constants. Let * be a specific reference population among the k populations with estimate $\hat{\theta}_*$ and standard error SE_*.

Assuming $\hat{\theta}_*$ and $\hat{\theta}_i$ are independent and each normally distributed for $i \neq *$, it is known that a $100(1-\alpha)\%$ confidence interval for $\theta_i - \theta_*$ is given by

$$\left((\hat{\theta}_i - \hat{\theta}_*) - z_{\frac{\alpha}{2}}\sqrt{(SE_i)^2 + (SE_*)^2} \ , \ (\hat{\theta}_i - \hat{\theta}_*) + z_{\frac{\alpha}{2}}\sqrt{(SE_i)^2 + (SE_*)^2}\right) \quad (1)$$

where $z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2})$. At significance level α, to test $H_0 : \theta_i = \theta_* \ \ vs \ \ H_A : \theta_i \neq \theta_*$, reject H_0 in favor of H_A if the interval (1) does not contain zero (0); otherwise, we do not reject H_0.

Figure 1 gives an estimated ranking of the 51 states based on estimates $\hat{\theta}_i$ for the i^{th} state's mean travel time to work. Each column presents results of fifty tests - each comparing the reference state noted at the bottom with one of the other fifty states. In each column, shaded states do (unshaded states do not) differ from reference state for mean travel time to work (minutes). The significance level for each test comparing $\hat{\theta}_*$ with each $\hat{\theta}_i$ is $\frac{\alpha}{50} = .002$ (Bonferroni corrected). The family-wide (or overall) significance level for all fifty tests simultaneously being compared in each column is 0.1. For the USA, $\hat{\theta} = 25.51$ and $SE = 0.02$.

2.2. Comparing One Reference State Against Each of the Other States - Version II

Relative to zero (0), Figure 2 gives 50 different confidence intervals for the differences $\theta_i - \theta_*$ for reference state * $\equiv$ Colorado (CO) for mean travel time to work and $i \neq *$. We use a Bonferroni correction for the tests as noted in Section 2.1. The bold intervals show the states that are statistically significantly different

from CO, while the non-bold intervals show the states that are not statistically significantly different from CO.

2.3. Comparing One Reference State Showing Its Confidence Interval with Each of the Other States Showing Their "Comparison Intervals"

Given a reference state * with a $100(1-\alpha)\%$ confidence interval for θ_* as given in (2), Almond, Lewis, Tukey, and Yan (2000) show that it is possible to construct an interval $(\hat{\theta}_i - w_i, \hat{\theta}_i + w_i)$ for state $i \neq *$ such that when the two intervals overlap, θ_i and θ_* are not statistically significantly different at level α, whereas if the two intervals do not overlap, then θ_i and θ_* are statistically significantly different.

$$\left(\hat{\theta}_* - z_{\frac{\alpha}{2}} SE_* \ , \ \hat{\theta}_* + z_{\frac{\alpha}{2}} SE_*\right) \tag{2}$$

The appropriate value of w_i is $w_i = z_{\frac{\alpha}{2}}\sqrt{(SE_*)^2 + (SE_i)^2} - z_{\frac{\alpha}{2}} SE_*$. Relative to $\hat{\theta}_*$, we refer to the interval $(\hat{\theta}_i - w_i, \hat{\theta}_i + w_i)$ as a "θ_* *comparison interval for* θ_i." The comparison interval for θ_i is not a confidence interval, while the interval for θ_* is a confidence interval.

Figure 3 shows results of fifty tests (each Bonferroni corrected at level $\frac{\alpha}{50} = .002$, and overall level $\alpha = 0.1$) comparing reference state CO with each of the other states. We see that CO's mean travel time to work is significantly different from all of the states except MS, AL, NV, MI, TN, LA, AZ, TX, CT, DE, and WV.

Remark 1: While visually different, Figure 3, Figure 2, and the column for CO in Figure 1, all provide the same comparison results for CO. In Figure 3, the usual 99.8% confidence interval for the reference state CO (θ_*) is shown explicitly; the (Bonferroni-corrected) "comparison intervals" are not usual confidence intervals. On the other hand, all of the intervals in Figure 2 are the usual 99.8% confidence intervals for $\theta_i - \theta_*$, but we do not see the 99.8% confidence interval for θ_*. No confidence intervals are shown in Figure 1.

2.4. Comparing A Pair of States by Presenting Meaningful Overlapping/Non-overlapping Confidence Intervals Appropriately for Each State in the Pair

Figure 4 gives 90% confidence intervals for each state. What can we say about θ_i vs θ_j if their confidence intervals overlap or if they do not overlap? If the desire is that the level of significance be α, Goldstein and Healy (1995) show how to adjust the confidence coefficient to a value, say $100(1-\alpha_A)\%$, such that if the $100(1-\alpha_A)\%$ confidence interval for θ_i does not overlap an independent $100(1-\alpha_A)\%$ confidence interval for θ_j, then we can declare θ_i and θ_j as statistically significantly different at significance level α. Also if the $100(1-\alpha_A)\%$ confidence intervals do overlap, we can not declare that θ_i and θ_j differ at level α.

Comparing One Pair of Populations i and j: Let $(SE_{ij})^2 \equiv Var(\hat{\theta}_i - \hat{\theta}_j) = (SE_i)^2 + (SE_j)^2$. One can show that $|\hat{\theta}_i - \hat{\theta}_j| > z_{\frac{\alpha_A}{2}}(SE_i + SE_j)$ if and only if the $100(1-\alpha_A)\%$ confidence intervals for θ_i and θ_j do not overlap. Thus the probability of a Type I error under the hypothesis $\theta_i = \theta_j$ is

$$\gamma_{ij} \quad = \quad P\left(|\hat{\theta}_i - \hat{\theta}_j| > z_{\frac{\alpha_A}{2}}(SE_i + SE_j)\right) \quad = \quad 2\left(1 - \Phi(z_{\frac{\alpha_A}{2}} \tfrac{(SE_i + SE_j)}{SE_{ij}})\right) \ . \tag{3}$$

Thus (3) relates γ_{ij} and $z_{\frac{\alpha_A}{2}}$ (hence α and $z_{\frac{\alpha_A}{2}}$) for given values of SE_i and SE_j. So if we want the probability of a Type I error γ_{ij} to be equal to a specific value, say α, then we are able to determine α_A such that when the two $100(1-\alpha_A)\%$ confidence intervals for θ_i and θ_j do not overlap we can correctly say that θ_i and θ_j are statistically significantly different at significance level α. Also when they do overlap, we would not be able to say they differ at level α.

Using estimates $\hat{\theta}_i$ and SE_i from Figure 1, we give two examples illustrating the method of Goldstein and Healy (1995). *Example 1: Comparing the Pair of States AZ and CO* - For AZ (θ_{AZ}) and CO (θ_{CO}), we can determine for $\alpha = 0.1$ that we have $100(1-\alpha_A)\% = 76\%$ confidence intervals for θ_{AZ} and θ_{CO} respectively as (24.62, 24.98) and (24.28, 24.72) which do overlap. Thus we would infer that θ_{AZ} and θ_{WY} are not different at $\alpha = 0.1$. Also, a 90% confidence interval for $\theta_{AZ} - \theta_{CO}$ is (-0.10, 0.70) which does contain 0. *Example 2: Comparing the Pair of States AZ and WY* - For AZ (θ_{AZ}) and WY (θ_{WY}), we can determine for $\alpha = 0.1$ that we have $100(1-\alpha_A)\% = 81\%$ confidence intervals for θ_{AZ} and θ_{WY} respectively as (24.56, 24.96) and (17.44, 18.76) which do not overlap. Thus we would infer that θ_{AZ} and θ_{WY} differ at $\alpha = 0.1$. Also, a 90% confidence interval for $\theta_{AZ} - \theta_{WY}$ is (6.57, 6.75) which does not contain 0.

Comparing All Pairs of Populations i and j: Goldstein and Healy (1995) note, "Where there are more than two (populations), we propose that (α_A) should be selected so that the average value of γ_{ij} over all (i,j) is a predetermined value, say α, typically 0.05 or 0.01. For a given data set, this can be determined by a straightforward search procedure. A starting point for $z_{\frac{\alpha_A}{2}}$ is the average of $z_{\frac{\alpha_A}{2}} \frac{SE_i + SE_j}{SE_{ij}}$ taken over all the pairs (i,j). The confidence interval for the i^{th} (population) is then given by $\left(\hat{\theta}_i - z_{\frac{\alpha_A}{2}} SE_i \ , \ \hat{\theta}_i + z_{\frac{\alpha_A}{2}} SE_i\right)$."

Remark 2: Figure 5 permits a visual comparison of any pair of states. For example, the 77.49% confidence intervals for Iowa and Kansas overlap; hence we would not be able to say that Iowa and Kansas differ for an average significance level of $\alpha = 0.1$. On the other hand, the 77.49% confidence intervals for Iowa and Idaho do not overlap. Thus we would say that Iowa and Idaho differ for an average significance level of $\alpha = 0.1$.

3. BOOTSTRAPPING AND RANKING

3.1. Some Uncertainty Measures for Estimated Ranks

In Section 2, we presented uncertainty in the estimated ranking through confidence intervals and hypothesis tests for individual θ_i's and for the pairwise differences $\theta_i - \theta_j$. Alternatively, one may consider the individual ranks as the parameters of interest and make inferences on them directly. The unknown true ranks are denoted by $r_1, r_2, ..., r_k$, and formally, we define the rank for the i^{th} smallest population as

$$r_i = \sum_{j=1}^{k} I(\theta_j \leq \theta_i) = 1 + \sum_{j:j\neq i} I(\theta_j \leq \theta_i), \quad \text{for} \quad i = 1, 2, ..., k. \tag{4}$$

The estimated ranking, computed based on the estimates $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$, is denoted by $\hat{r}_1, \hat{r}_2, ..., \hat{r}_k$, where

$$\hat{r}_i = 1 + \sum_{j:j\neq i} I(\hat{\theta}_j \leq \hat{\theta}_i), \quad \text{for} \quad i = 1, 2, ..., k. \tag{5}$$

Uncertainty in $\hat{\theta}_i$ induces uncertainty in $\hat{r}_i$. Some uncertainty measures for $\hat{r}_i$ follow.

(a) A collection of k confidence intervals for the unknown ranks $r_1, r_2, ..., r_k$ as suggested by Barker, Smith, Gerzoff, Luman, McCauley, and Strine (2005) and Goldstein and Spiegelhalter (1996).

(b) A collection of k estimates of the probabilities $P(|\hat{r}_i - r_i| \leq c)$ for some chosen value of c, as suggested by Klein and Wright (2011).

(c) An estimate of the joint probability $P(|\hat{r}_1 - r_1| \leq c, |\hat{r}_2 - r_2| \leq c, ..., |\hat{r}_k - r_k| \leq c)$ as mentioned by Klein and Wright (2011).

How are estimates of the quantities (a) - (c) computed (by the statistical agency) and how are they interpreted (by the data user)? See Sections 3.2 and 3.3.

3.2. Bootstrap Estimation

The bootstrap (Efron, 1979) provides a clear-cut way to compute/estimate the uncertainty measures (a) - (c) above; it is a computer intensive statistical method with broad applications (Shao and Tu, 1995).

Nonparametric Bootstrap: In the nonparametric bootstrap, we estimate each of the k population cumulative distribution functions $F_1(y), F_2(y), ..., F_k(y)$ by the empirical distribution functions defined as

$$\hat{F}_i(y) = \frac{1}{n_i} \sum_{j=1}^{n_i} I(Y_{ij} \leq y), \quad \text{for} \quad i = 1, 2, ..., k. \tag{6}$$

Note that $\hat{F}_i(y)$ places equal probability on each of the observed data points $Y_{i1}, Y_{i2}, ..., Y_{in_i}$. An estimate of a quantity such as $P(|\hat{r}_i - r_i| \leq c)$ is then obtained by computing this probability for the case that $F_1(y), F_2(y), ..., F_k(y)$ are replaced by their estimates $\hat{F}_1(y), \hat{F}_2(y), ..., \hat{F}_k(y)$. Even when $F_1(y), F_2(y), ..., F_k(y)$ are replaced by the estimates, quantities such as (a) - (c) in Section 3.1 may still be difficult to calculate analytically, and therefore a Monte Carlo estimator is used. Thus to obtain nonparametric bootstrap estimates, we use the following algorithm.

Step 1. Draw $Y^*_{i1}, Y^*_{i2}, ..., Y^*_{in_i}$ as a simple random sample with replacement from $Y_{i1}, Y_{i2}, ..., Y_{in_i}$. Do this independently for each $i = 1, 2, ..., k$.

Step 2. (a) Compute the bootstrap analog of $\hat{\theta}_i$ which is defined as $\hat{\theta}^*_i = \hat{\theta}_i(Y^*_{i1}, Y^*_{i2}, ..., Y^*_{in_i}) \; \forall \; i$.

(b) Compute the bootstrap analog of $\hat{r}_i$ which is defined as $\hat{r}^*_i = 1 + \sum_{j:j\neq i} I(\hat{\theta}^*_j \leq \hat{\theta}^*_i) \; \forall \; i$.

Step 3. Repeat Steps 1 and 2 a total of B times where B is sufficiently large to get $(\hat{r}^*_{1,1}, \hat{r}^*_{2,1}, ..., \hat{r}^*_{k,1})$, $(\hat{r}^*_{1,2}, \hat{r}^*_{2,2}, ..., \hat{r}^*_{k,2}), ..., (\hat{r}^*_{1,B}, \hat{r}^*_{2,B}, ..., \hat{r}^*_{k,B})$, a collection of bootstrap replications of the ranks.

Given the bootstrap replications of ranks, a bootstrap estimate of $P\{|\hat{r}_i - r_i| \leq c\}$ is obtained as

$$\hat{P}_{boot}\{|\hat{r}_i - r_i| \leq c\} = \frac{1}{B} \sum_{b=1}^{B} I\{|\hat{r}^*_{i,b} - \hat{r}_i| \leq c\}, \tag{7}$$

and a bootstrap estimate of $P(|\hat{r}_1 - r_1| \leq c, |\hat{r}_2 - r_2| \leq c, ..., |\hat{r}_k - r_k| \leq c)$ is obtained as

$$\hat{P}_{boot}(|\hat{r}_1 - r_1| \leq c, ..., |\hat{r}_k - r_k| \leq c) = \frac{1}{B} \sum_{b=1}^{B} I\{|\hat{r}^*_{1,b} - \hat{r}_1| \leq c, ..., |\hat{r}^*_{k,b} - \hat{r}_k| \leq c\}. \tag{8}$$

An approximate $100(1-\alpha)\%$ bootstrap confidence interval for r_i can be obtained as

$$[\hat{r}_i^{*(\frac{\alpha}{2})}\ ,\ \hat{r}_i^{*(1-\frac{\alpha}{2})}] \tag{9}$$

where $\hat{r}_i^{*(\frac{\alpha}{2})}$ and $\hat{r}_i^{*(1-\frac{\alpha}{2})}$ denote, respectively, the empirical $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the bootstrap replications $\hat{r}^*_{i,1}, \hat{r}^*_{i,2}, ..., \hat{r}^*_{i,B}$. The confidence interval (9) is called the *bootstrap percentile interval* (Efron, 1981).

Parametric Bootstrap: When the sampling distribution of each $\hat{\theta}_i$ is well approximated by a normal distribution, and the SE_i are provided for all i, it is natural to use a parametric bootstrap procedure in which we generate bootstrap replications of $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ directly from normal distributions as given in the following algorithm.

Step 1. Draw $\hat{\theta}_i^*$ from $N(\hat{\theta}_i, SE_i)$, independently for $i = 1, 2, ..., k$.

Step 2. Compute the bootstrap analog of $\hat{r}_i$ which is defined as $\hat{r}_i^* = 1 + \sum_{j:j\neq i} I(\hat{\theta}_j^* \leq \hat{\theta}_i^*)$ for $i = 1, 2, ..., k$.

Step 3. Repeat Steps 1 and 2 a total of B times where B is sufficiently large to get $(\hat{r}^*_{1,1}, \hat{r}^*_{2,1}, ..., \hat{r}^*_{k,1})$, $(\hat{r}^*_{1,2}, \hat{r}^*_{2,2}, ..., \hat{r}^*_{k,2}), ..., (\hat{r}^*_{1,B}, \hat{r}^*_{2,B}, ..., \hat{r}^*_{k,B})$, a collection of bootstrap replications of the ranks.

Given the bootstrap replications of the ranks using this procedure, estimates of the various uncertainty measures are obtained using the estimators (7) - (9). Notice that the parametric bootstrap algorithm generates $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_k^*$ directly, as opposed to the nonparametric bootstrap which first takes a random sample from the underlying data $Y_{i1}, Y_{i2}, ..., Y_{in_i}$. Thus the parametric bootstrap in this case has three potential advantages over the nonparametric bootstrap: (i) it requires less computation and hence will run more quickly, (ii) it has less code to debug, and (iii) it can be applied in situations where $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ and $SE_1, SE_2, ..., SE_k$ are available or otherwise known but the underlying data are not. The normality approximation is critical.

Remark 3: In the spirit of (7) - (9), we can also compute various other *"nonparametric or parametric bootstrap estimates of probabilities,"* such as: (1) P(estimated rank of state i is among 5 highest) $= P(\hat{r}_i \in \{47, 48, 49, 50, 51\})$; (2) P(estimated ranks of states i and j are among 4 highest) $= P(\hat{r}_i \in \{48, ..., 51\}$, $\hat{r}_j \in \{48, ..., 51\})$; and (3) P(estimated rank of state i is higher than estimated rank of state j) $= P(\hat{r}_i > \hat{r}_j)$.

Remark 4: An alternative form of the parametric bootstrap (actually the more commonly used form in some applications) can be obtained if we assume a parametric model $F_i(y|\varphi_i)$ for $F_i(y)$, $i = 1, 2, ..., k$, where φ_i is an unknown parameter vector and $F_i(y|\varphi_i)$ is known when the value of φ_i is known. We draw samples of sizes $n_1, ..., n_k$ from the estimated populations $F_1(y|\hat{\varphi}_1), ..., F_k(y|\hat{\varphi}_k)$, respectively, where $\hat{\varphi}_i$ is an appropriate estimate of φ_i. To be clear, this alternative form of the bootstrap requires a model assumption on the data, and not an assumption about the estimates $\hat{\theta}_i$ where the central limit theorem is more likely to apply.

3.3. Application to American Community Survey Travel Time to Work Data

Using the parametric bootstrap and the $\hat{\theta}_i$ and SE_i from Figure 1 for the $k = 51$ states, we estimate the uncertainty measures (a) - (c) from Section 3.1; the results are reported in Tables 1 and 2 for $B = 100000$.

Remark 5: Note that the event $|\hat{r}_i - r_i| \leq c$ is equivalent to the event $\hat{r}_i - c \leq r_i \leq \hat{r}_i + c$, and hence $P\{|\hat{r}_i - r_i| \leq c\} = P\{\hat{r}_i - c \leq r_i \leq \hat{r}_i + c\}$. Therefore, noting that $1 \leq r_i \leq k$, one can think of

$$[\max\{\hat{r}_i - c, 1\}, \min\{\hat{r}_i + c, k\}] \tag{10}$$

as a confidence interval for the unknown rank r_i, where the bootstrap estimated probabilities in Table 1 give estimates of the confidence coefficient of the interval for $c = 0, 1, 2, 3$. As an illustration, suppose we want a 0.90 level confidence interval for Nebraska's rank (whose estimate is $\hat{r}_i = 3$). From Table 1, we find that the estimates of $P\{|\hat{r}_i - r_i| \leq c\}$ are 0.31, 0.71, 0.94, and 1.00 for c =0, 1, 2, and 3, respectively. Thus we would take $[3-2, 3+2] = [1, 5]$ as an approximate level 0.90 (approximate confidence coefficient is actually 0.94) confidence interval for Nebraska's rank.

Remark 6: Next consider the quantities $\hat{r}_i^{*(.05)}$ and $\hat{r}_i^{*(.95)}$, also displayed in the last two columns of Table 1 for each state $i = 1, ..., k$. Based on the bootstrap percentile method for obtaining a confidence interval, these quantities can be interpreted as the left and right endpoints, respectively, of an approximate level 0.90 confidence interval for the unknown rank r_i. Thus, based on this method, we find that a 0.90 level confidence interval for the rank of Nebraska is $[3, 6]$, which is different from the interval of $[1, 5]$ reported in the preceding paragraph as an approximate 0.94 level confidence interval for Nebraska's rank. It is worth noting that Nebraska's point estimate (18.06) with $\hat{r}_i = 3$ is much closer to those point estimates (18.10, 18.18, and 18.39) of states with $\hat{r}_j = 4, 5$, or 6 than those (16.86 and 16.91) for states with $\hat{r}_j = 1$ or 2. So the symmetric $\hat{r}_i \pm 2$ confidence interval is quite different from the equal tail bootstrap percentile interval.

Remark 7: The estimates of the joint probability $P(|\hat{r}_1 - r_1| \leq c, |\hat{r}_2 - r_2| \leq c, ..., |\hat{r}_k - r_k| \leq c)$ are presented in Table 2 for $c = 0, 1, ..., 8$. One can interpret the estimated probabilities as approximate confidence coefficients for a joint confidence set on the entire ranking $(r_1, r_2, ..., r_k)$ whose form is the rectangular region:

$$[\max\{\hat{r}_1 - c, 1\}, \min\{\hat{r}_1 + c, k\}] \times \cdots \times [\max\{\hat{r}_k - c, 1\}, \min\{\hat{r}_k + c, k\}]. \quad (11)$$

For example, we see from Table 2 that with $c = 5$ the estimated confidence coefficient of the above region (11) is approximately 0.93. Therefore we can claim that we are an estimated 90% confident that simultaneously the rank of each state is contained within the interval formed by adding and subtracting 5 from each estimated rank. This method provides a straightforward way to make an overall inference on the ranking, without the need for any further adjustment for multiple comparisons.

Remark 8: Notice that we have two reasonable methods for obtaining an approximate confidence interval on an individual rank r_i, namely, (i) take the interval as (10) and use the bootstrap to estimate the confidence coefficient, and (ii) the bootstrap percentile confidence interval given by (9). The question of which of these intervals is preferable requires further investigation and we will not pursue it here.

4. MOST PROBABLE RANKING

So far in this paper, we have focused on ranking by considering an ordering of θ_i. We will briefly explore consideration of comparing populations with probability statements such as $P(Y_1 < Y_2)$, where Y_1 and Y_2 are independent continuous random variables associated with populations 1 and 2, respectively. In the context of the travel time to work variable used throughout this paper, this probability can be viewed as the probability that a random person selected from state 1 has a shorter travel time to work than a random person from state 2.

Assuming $k = 3$, we motivate the concept of *a most probable ranking*. Assume random variables Y_1, Y_2, and Y_3 with independent continuous distributions. For the $3! = 6$ possible orderings, define the probabilities $P_{123} = P\{Y_1 < Y_2 < Y_3\}$, $P_{132} = P\{Y_1 < Y_3 < Y_2\}$, $P_{213} = P\{Y_2 < Y_1 < Y_3\}$, $P_{231} = P\{Y_2 < Y_3 < Y_1\}$, $P_{312} = P\{Y_3 < Y_1 < Y_2\}$, and $P_{321} = P\{Y_3 < Y_2 < Y_1\}$. We define *the most probable ranking* of the three populations as the permutation of $\{1, 2, 3\}$ that corresponds to $max\{P_{123}, P_{132}, P_{213}, P_{231}, P_{312}, P_{321}\}$. Because there could be more than one ordering with the largest probability, we will think of the expression, *a most probable ranking*. Given independent random samples of sizes n_1, n_2, and n_3 respectively from the three populations, it seems reasonable that one could estimate a most probable ranking by considering all tuplets of order $k = 3$ with one observation from each sample and counting the frequency of each of the 6 orderings based on the samples. An estimator of a most probable ranking is the ordering among the tuplets with the highest frequency. If more than one ordering has the highest frequency, we can report all such orderings as most probable rankings. The extension to k populations is conceptually straightforward. There seems to be a role for bootstrap in this.

5. CONCLUDING COMMENTS

See Remarks 1 - 8. Future research will evaluate the confidence regions in (9) - (11) and explore the concept of a most probable ranking.

Disclaimer and Acknowledgments: This paper is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U. S. Bureau of the Census. We are grateful to our colleagues who read various drafts of this paper: Derrick Simmons, Jun Shao, Carolina Franco, Josh Tokle, Pat Hunley, and Sarah Wilson.

REFERENCES

- Almond, R. G., Lewis, C., Tukey, J. W., and Yan, D. (2000). "Displays for Comparing a Given State to Many Others", *The American Statistician, Vol. 54, No. 2*, 89-93.
- Barker, L. E., Smith, P. J., Gerzoff, R. B., Luman, E. T., McCauley, M. M., and Strine, T. W. (2005). "Ranking States' Immunization Coverage: An Example from the National Immunization Survey", *Statistics in Medicine, 24*, 605 - 613.
- Cochran, W. G. (1977). *Sampling Techniques (3^{rd} Edition)*, New York, NY: John Wiley & Sons.
- Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife", *Annals of Statistics, 1*, 1-26.
- Efron, B. (1981). "Nonparametric Standard Errors and Confidence Intervals", *The Canadian Journal of Statistics, Vol. 9, No. 2*, 139-158.
- Fuller, W. A. (2009). *Sampling Statistics*, Hoboken, NJ: John Wiley & Son
- Goldstein, H. and Healy, M.J.R. (1995). "The Graphical Presentation of a Collection of Means", *Journal of the Royal Statistical Society, Series A, Vol. 158, No. 1*, 175-177.
- Goldstein, H. and Spiegelhalter, D. J. (1996). "League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance", *Journal of the Royal Statistical Society, Series A, Vol. 159 No. 3*, 385-443.
- Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods (2^{nd})*, New York, NY: John Wiley & Sons.
- Klein, M. and Wright, T. (2011). "Ranking Procedures for Several Normal Populations: An Empirical Investigation", *International Journal of Statistical Sciences, 11*, 37-58.
- Lohr, S. L. (2010). *Sampling: Design and Analysis (2^{nd} Edition)*, Boston, MA: Brooks/Cole.
- Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*, New York, NY: Springer.
- Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, New York, NY: Springer-Verlag.

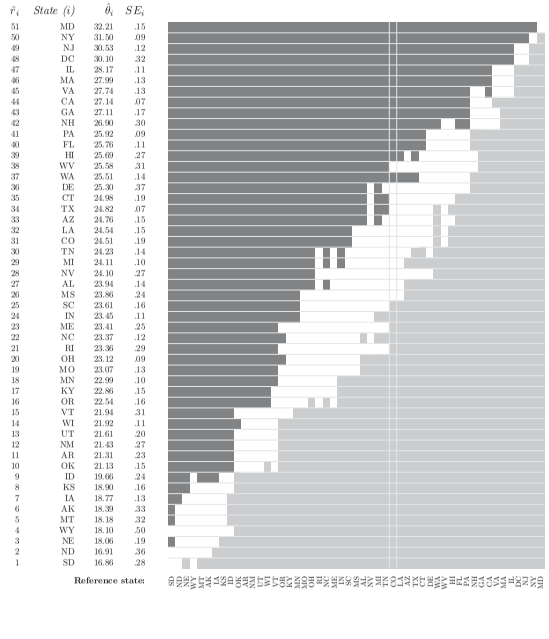

Figure 1: Results of Pairwise Hypothesis Tests

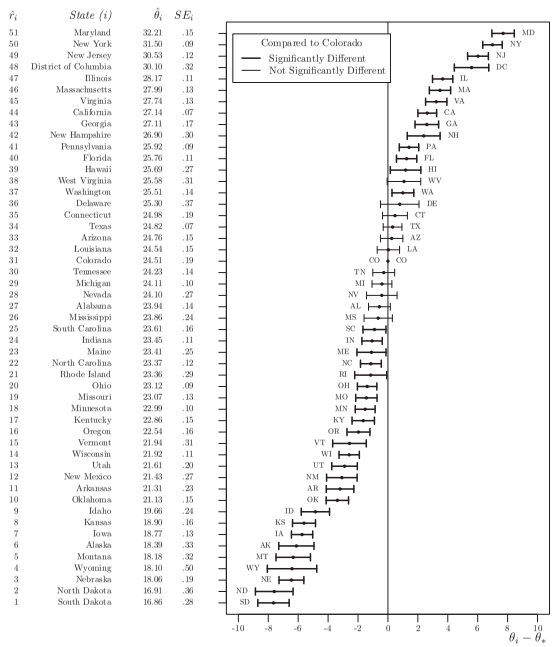

Figure 2: Confidence Intervals for Differences

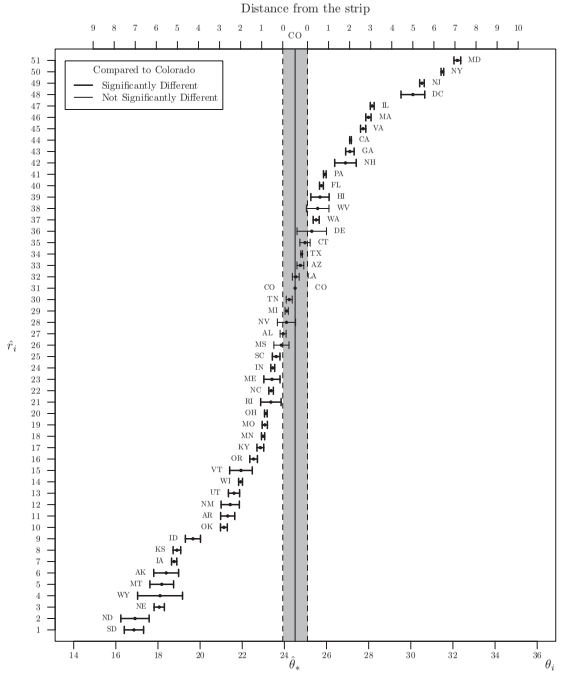

Figure 3: Comparison Intervals

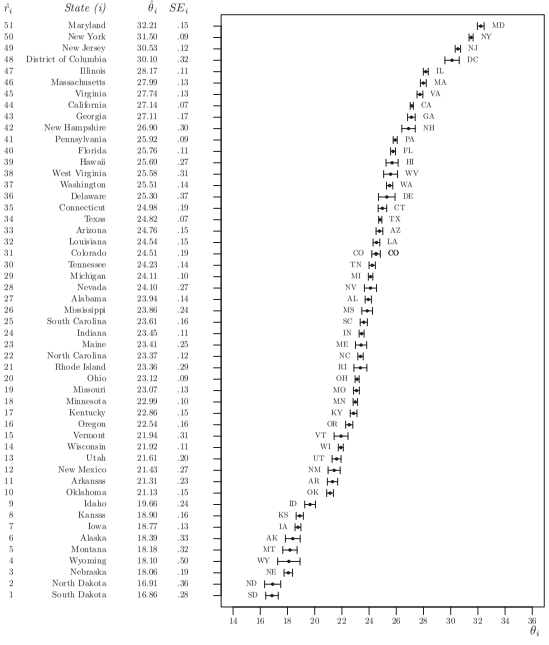

Figure 4: 90% Confidence Intervals

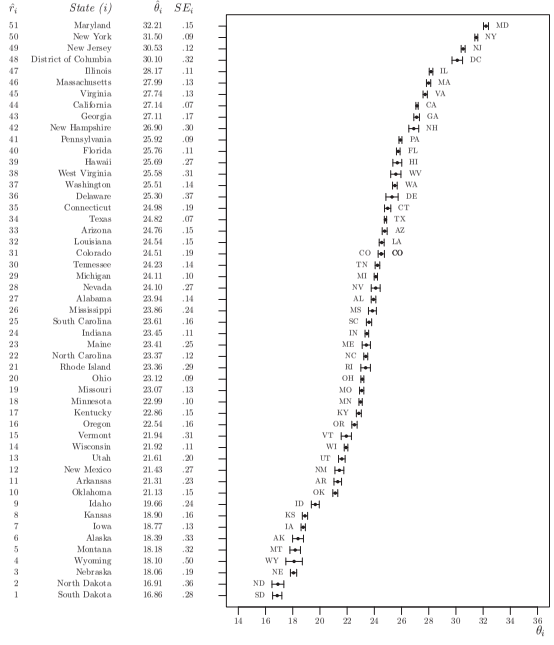

Figure 5: 77.49% Confidence Intervals

Table 1: Parametric Bootstrap Estimates of Some Uncertainty Measures for Estimated Ranks

Estimated Rank ($\hat{r}_i$)	State (i)	$\hat{\theta}_i$	SE_i	$\hat{P}_{boot}\{\lvert\hat{r}_i - r_i\rvert \le c\}$ $c=0$	$c=1$	$c=2$	$c=3$	90% Confidence Interval $\hat{r}_i^{*(.05)}$	$\hat{r}_i^{*(.95)}$
51	Maryland	32.21	0.15	1.00	1.00	1.00	1.00	51	51
50	New York	31.50	0.09	1.00	1.00	1.00	1.00	50	50
49	New Jersey	30.53	0.12	0.89	1.00	1.00	1.00	48	49
48	District of Columbia	30.10	0.32	0.89	1.00	1.00	1.00	48	49
47	Illinois	28.17	0.11	0.85	1.00	1.00	1.00	46	47
46	Massachusetts	27.99	0.13	0.77	1.00	1.00	1.00	45	47
45	Virginia	27.74	0.13	0.90	1.00	1.00	1.00	45	46
44	California	27.14	0.07	0.45	0.89	1.00	1.00	42	44
43	Georgia	27.11	0.17	0.41	1.00	1.00	1.00	42	44
42	New Hampshire	26.90	0.30	0.68	0.83	0.99	1.00	42	44
41	Pennsylvania	25.92	0.09	0.58	0.89	0.98	1.00	39	41
40	Florida	25.76	0.11	0.34	0.79	0.95	1.00	38	41
39	Hawaii	25.69	0.27	0.21	0.57	0.91	0.99	36	41
38	West Virginia	25.58	0.31	0.20	0.57	0.84	0.99	36	41
37	Washington	25.51	0.14	0.38	0.85	0.97	1.00	36	39
36	Delaware	25.30	0.37	0.35	0.64	0.79	0.89	33	40
35	Connecticut	24.98	0.19	0.50	0.84	0.96	0.99	33	36
34	Texas	24.82	0.07	0.48	0.93	1.00	1.00	33	35
33	Arizona	24.76	0.15	0.36	0.78	0.97	1.00	31	35
32	Louisiana	24.54	0.15	0.39	0.86	0.97	1.00	30	33
31	Colorado	24.51	0.19	0.37	0.78	0.92	0.98	29	34
30	Tennessee	24.23	0.14	0.41	0.79	0.94	0.99	27	31
29	Michigan	24.11	0.10	0.35	0.82	0.97	1.00	27	30
28	Nevada	24.10	0.27	0.18	0.51	0.81	0.93	25	31
27	Alabama	23.94	0.14	0.38	0.83	0.97	1.00	25	29
26	Mississippi	23.86	0.24	0.30	0.64	0.82	0.92	23	30
25	South Carolina	23.61	0.16	0.38	0.75	0.91	0.97	22	26
24	Indiana	23.45	0.11	0.27	0.68	0.90	0.99	21	25
23	Maine	23.41	0.25	0.16	0.48	0.77	0.89	19	26
22	North Carolina	23.37	0.12	0.31	0.77	0.94	0.99	20	24
21	Rhode Island	23.36	0.29	0.16	0.39	0.58	0.76	18	26
20	Ohio	23.12	0.09	0.39	0.84	0.98	1.00	18	22
19	Missouri	23.07	0.13	0.29	0.75	0.95	0.99	17	21
18	Minnesota	22.99	0.10	0.42	0.84	0.96	0.99	17	20
17	Kentucky	22.86	0.15	0.60	0.86	0.94	0.98	16	20
16	Oregon	22.54	0.16	0.88	0.99	1.00	1.00	16	17
15	Vermont	21.94	0.31	0.47	0.77	0.92	0.97	12	15
14	Wisconsin	21.92	0.11	0.49	0.99	1.00	1.00	13	15
13	Utah	21.61	0.20	0.46	0.87	0.99	1.00	11	14
12	New Mexico	21.43	0.27	0.34	0.78	0.98	1.00	10	14
11	Arkansas	21.31	0.23	0.38	0.87	0.98	1.00	10	13
10	Oklahoma	21.13	0.15	0.63	0.92	0.99	1.00	10	12
9	Idaho	19.66	0.24	0.99	1.00	1.00	1.00	9	9
8	Kansas	18.90	0.16	0.65	0.92	0.99	1.00	6	8
7	Iowa	18.77	0.13	0.56	0.97	1.00	1.00	6	8
6	Alaska	18.39	0.33	0.36	0.71	0.92	1.00	3	8
5	Montana	18.18	0.32	0.29	0.74	0.98	1.00	3	6
4	Wyoming	18.10	0.50	0.18	0.72	0.90	0.95	3	8
3	Nebraska	18.06	0.19	0.31	0.71	0.94	1.00	3	6
2	North Dakota	16.91	0.36	0.26	0.52	1.00	1.00	1	2
1	South Dakota	16.86	0.28	0.54	0.99	1.00	1.00	1	2

Table 2: Parametric Bootstrap Estimates of $P(\lvert\hat{r}_1 - r_1\rvert \le c, \lvert\hat{r}_2 - r_2\rvert \le c, \ldots, \lvert\hat{r}_k - r_k\rvert \le c)$

c	0	1	2	3	4	5	6	7	8
Estimated Probability	0.00	0.00	0.06	0.41	0.76	0.93	0.98	0.99	1.00